Yan Liu
Junichi Hirukawa
Yoshihide Kakizawa *Editors*

Research Papers in Statistical Inference for Time Series and Related Models

Essays in Honor of Masanobu Taniguchi

Springer

Research Papers in Statistical Inference for Time Series and Related Models

Yan Liu · Junichi Hirukawa · Yoshihide Kakizawa
Editors

Research Papers in Statistical Inference for Time Series and Related Models

Essays in Honor of Masanobu Taniguchi

Springer

Editors
Yan Liu
Faculty of Science and Engineering
Waseda University
Shinjuku-ku, Tokyo, Japan

Junichi Hirukawa
Faculty of Science
Niigata University
Nishi-ku, Niigata, Japan

Yoshihide Kakizawa
Faculty of Economics and Business
Hokkaido University
Kita-ku, Sapporo, Hokkaido, Japan

ISBN 978-981-99-0802-8 ISBN 978-981-99-0803-5 (eBook)
https://doi.org/10.1007/978-981-99-0803-5

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Foreword

The Long and Curved Road

It was about 50 years ago when I was a master's student at Osaka University, I started time series analysis by Hannan's book: Multiple Time Series (1970), which deals with statistical analysis for multivariate stochastic processes. It is a very difficult and high-level book for beginners. At that time, I felt hectic and discouraged to study time series analysis.

Entering into the Ph.D. course, I had an occasion to hear a junior student's seminar talk on Beran (1977, Ann. Stat.), which discusses minimum Hellinger distance estimation in i.i.d. case. These days, because there have been no approaches using Hellinger distance for time series, I was much influenced by this paper. Motivated by Beran's paper, I could introduce two divergences between the periodogram and parametric spectral densities and developed their asymptotic theory in view of efficiency and robustness.

In 1978, I got a research position at Hiroshima University, where there was an active group of researchers in multivariate analysis and experimental design. There, multivariate analyzers were applying the asymptotic expansion to a variety of statistics. I got much influenced by them and could proceed to higher-order asymptotic theory for time series by use of the asymptotic expansion method.

Around that time, Ted Hannan visited Hiroshima. I was very pleased to meet the VIP person in time series analysis and responded to this opportunity with best Japanese hospitality, e.g., sightseeing, gourmet, etc. However, Ted did not appreciate such Japanese hospitality, but, he invited me to Australian National University (ANU) in 1983. In this era, ANU was in a golden period. At the seminar, beside Ted, Heyde C., Gani, J., Hall, P., Speed, T., Daley, D., etc. were there. It was a high-standard one. My first foreign stay was a very impressive one.

Coming back to Hiroshima, I was interested in higher-order asymptotic theory in time series analysis and econometrics. Especially, I tried to establish the validity of Edgeworth expansion for the maximum likelihood estimators and related statistics.

In 1986, I was invited to the Center for Multivariate Analysis at the University of Pittsburgh by Krishnaiah. Because he was the editor of Journal of Multivariate Analysis, he provided me with the latest achievement in multivariate analysis. Then I could write papers on the problem of eigenvalue for multivariate time series efficiently. Also, the third-order asymptotic efficiency was discussed in time series observations. As a related topic, I could write a paper on statistical inference of dyadic processes with Zhidong Bai, etc.

After this, I was interested in the differential geometry of curved exponential families, which was introduced by Amari. He defined an ancillary space corresponding to the estimator and its curvature, which measures the goodness of the estimator. Then I tried to generalize this approach to time series models.

Around that time, Dahlhaus and Hardle organized workshops in Heidelberg and at the Oberwolfach Research Institute for Mathematics, respectively. I was invited there and got acquainted with many people, e.g., Durbin, J., Robinson, P.M., Deistler, M., Tjøstheim, D., Stoffer, D., Rao, Subba, Giraitis, L., Bhansali, R., etc.

After I got transferred to Osaka University, I completed a paper on "Curved Probability Models" which includes multivariate stochastic processes, and the inference is very general. I felt that curved models include a lot of stochastic models, and the applications are vast.

In 1993, by Madan Puri, I was invited to Indiana University, USA, and developed the higher-order asymptotic theory for normalizing transformations of MLE. Based on the integral functional of the spectral density matrix, a new nonparametric approach was proposed for non-Gaussian vector stationary processes, which has a variety of applications in statistics and econometrics. In Indiana, I recognized that Madan was a good cook for Indian cuisine. I enjoyed his cuisines.

In 1996, Marc Hallin invited me to Universite libre de Brussels as a visiting professor. The Grand-Place is the central square of the City of Brussels. All over the world, it is known for its decorative and aesthetic wealth. I felt it was a very impressive place. Near my staying place, there was a very nice Thai–French restaurant, whose taste was a connoisseur. At the university, Marc and I developed the optimal asymptotic inference for time series regression models with long memory disturbances based on the local asymptotic normality (LAN). The results led to our Annals paper.

At Osaka University, Kakizawa and I were proceeding the problem of discrimination and clustering for multivariate time series by using the integral functional of spectral density matrix. Bob Shumway kindly joined our work providing us seismic data, leading to our JASA paper, and together with the above LAN and higher-order asymptotic theory, yielding Taniguchi–Kakizawa (2000, Springer book).

In 2000, Garderen K.J. invited me to the University of Bristol as a visiting professor. Although Bristol was a nice city, K.J. kindly took me to the neighbourhood places, e.g., Salisbury Cathedral, Stonehenge, etc., which were very impressive ones. In our research, I could publish a paper on the higher-order asymptotic estimation for semiparametric models in time series, inviting Madan as a co-author.

Around that time, Sangeol Lee invited me to Seoul National University, which accelerate my research exchange with Asian countries, e.g., Hong Kong, Taiwan, China, etc. Sangeol and I developed the asymptotic theory for the empirical residual process for the stochastic regression model by use of LAN. Transferred to Waseda University (2003), I organized an international symposium, inviting Marc, Peter Robinson, Linton, O., Howell Tong, Dahlhaus, R., Chan, N.H., Sangeol, Koul, H., Stoffer, D., etc. I was much influenced by these high-standard people.

After this, I was much interested in the empirical likelihood approach for time series. Anna Clara Monti was a leader of this field. For 2007–2008, Marc and I started Japan–Belgium Research Cooperative Program (JSPS and FNRS), and organized Belgian side workshops and Japanese side workshops. We invited Anna Clara to the Izu workshop. There she called disciples "kids". From this occasion, we have been using her way of calling disciple students.

Simultaneously I got a large JSPS funding Kiban (A) for 2007–2010 and organized 10 workshops in nice local cities. Many foreign and domestic researchers joined, mixed, and were much influenced each other. In 2006, Akaike, H. received Kyoto Prize. At that time, I was the editor of Journal of Japanese Statistical Society, and arranged the Celebration Volume for Akaike in 2008, inviting Philiips, P.C.B., Atkinson, A.C., Robinson, P.M., etc. Successively, I got Kiban (A) for 2011–2014, and Kiban (A) for 2015–2018, and organized 16 workshops for each. I invited many foreign researchers, e.g., Hall, P., Dette, H., Hallin, M., Koenker, R., Prakasa Rao, Mikosch, T., Bosq, D., Kedem, B., Kutoyants, Y.A., Khmaladze, E., Anna Clara Monti, Proietti, T., Luati, A., Tsay, R., Giraitis, L., Taqqu, M., Yao, Q., Chan, N.H., Ing, C.K., Meihui, G., Lee, S., Chen, Cathy, Dunsmuir, W., Negri, I., Bose, A., Chen, Y., Delaigle, A., Tjøstheim, D., Davis, R., Ombao, H., Yoon-Jae Whang, Francq, C., Zakoian, M., Patilea, V., Hansen, P., Pourahmadi, M., Fokianos, K., Sachs, R., Lahiri, S.N., and Chen, K.

A lot of collaborative researches and exchanges were proceeded by some of the above mentioned people, in many ways, especially, for young people. Academically, we published four English research books in Springer, Chapman Hall. Also, we organized a lot of local workshops in nice resort places, e.g., Mt. Fuji area, Hakone, Izu, Kyoto, Ise-Shima, Kanazawa, Niigata, Kagawa, Kōchi, Kumamoto, Kagoshima, Sapporo, etc. We could deepen our friendship, and enjoy gourmet there.

After this, I got a bigger JSPS grant, Kiban (S), for the period 2018–2022. The title is "Introduction of General Causality to Various Data and Its Applications". Hereafter, I exported our workshops to foreign countries. In 2019, Anna Clara organized Italian Workshop "Statistical Methods and Models for Complex Data" in Benevento. Below is a photo of the workshop.

After this, Solvang Kato organized the Norwegian workshop "Workshop on causal inference in complex marine ecosystems" in Bergen. We enjoyed the fjord tour and seafood. In September, Meihui arranged NSYSU–Waseda International Symposium "Time Series, Machine Learning and Causality Analysis" in Kaohsiung. We felt a very nice tropical atmosphere by the seaside. Moving to Tainan, Ray-Bing Chen arranged NCKU–Waseda International Symposium "Time Series, Machine Learning and Causality Analysis". Liang-Ching Lin showed us an interesting big salt pan and oyster field. Moving to Taichung, Cathy arranged FCU–Waseda International Symposium at Feng Chia University. We enjoyed the old town and sea coast. The university was planning to construct new buildings designed by Kengo Kuma (famous Japanese architect).

After this, Japan was infected by COVID-19. For about three years, we could not arrange any workshop. But, fortunately, in 2022, the epidemic waned. So I planned Italian workshops. Tommaso organized Rome–Waseda Time Series Symposium in Villa Mondragone. The conference site, dinners and lunches were very nice. Also, there were many high-standard talks and I enjoyed them. Moving to Bologna, Alessandra organized Bologna–Waseda Time Series Workshop at Accademia delle Scienze in Bologna. There were many interesting talks. The sanctuary of San Luca is one of the most important symbols of Bologna, so, we went there. The scenery from the top of church was fantastic. In Japan, I usually order Limoncello "with soda". But I recognized that the way was not normal in Bologna.

My statistical life is a sort of "The long and curved road". In the course of it, I had a lot of collaborations, friendships, and hospitalities. I thank all of the above people, all of the contributors of this volume, and all of the "kids" who always stimulated my research. Thanks are extended to my teachers: Hannan, E.J., Okamoto, M., Nagai, T. and Huzii, M., and leaders of my research: Takeuchi, Kei, Hosoya, Y., Fujikoshi, Y., Akahira, M. and Kitagawa, G., etc.

Tokyo, Japan
November 2022

Masanobu Taniguchi

Preface

The papers in this volume were contributed by the friends and students of Masanobu Taniguchi on the occasion of his 70th birthday. We wish him a belated happy birthday. In addition, we celebrate his very extensive contributions to asymptotic theory for time series and related topics.

This book of 26 contributed papers compiles theoretical developments on statistical inference for time series and related models in honor of Masanobu Taniguchi. Chapter 1 studies self-weighted GEL estimator and test statistics for hypotheses testing of parameters in vector autoregressive models with possibly infinite variance. Chapter 2 proposes a new estimator, called the excess mean of power estimator, for estimating the extreme value index, and the asymptotic normality of this new estimator is established for dependent observations. Chapter 3 proposes an exclusive topic model consisting of a pairwise Kullback–Leibler divergence and a weighted LASSO penalty for interpretation and topic detection. Chapter 4 proposes a new family of isotropic correlation functions whose members display long-range dependence and can also model different degrees of smoothness. Chapter 5 considers three portmanteau tests for general nonlinear conditionally heteroscedastic time series, and compares them in terms of Bahadur slope. Chapter 6 discusses the estimation of the AR(1) and MA(1) models driven by non-i.i.d. noise, with an emphasis on the impact of non-i.i.d. noise for the standard errors and confidence interval. Chapter 7 considers three tests for non-negative integer-valued time series, and the asymptotic properties of the tests are elucidated. Chapter 8 considers M-estimation in GARCH models under milder moment assumptions, together with weighted bootstrap approximations of the distributions of these M-estimators. Chapter 9 proposes the linear serial rank statistics for the problem of testing randomness against time-varying MA alternative, and then establishes its asymptotic normality under both null and alternative hypotheses. Chapter 10 discusses asymptotic expansions for the distributions of χ^2 type test statistics from GEL framework in the possibly over-identified moment restrictions, including the bootstrap-based Bartlett-type correction. Chapter 11 addresses the problem of testing the goodness-of-fit hypothesis in multiple linear regression models with non-random and random predictors. Chapter 12 proposes a new contrast function for multivariate time series and establishes the asymptotic normality of the proposed minimum contrast estimators. Chapter 13 considers the models based on the Box–Cox transform of the spectral density function of locally stationary processes. Chapter 14 describes a model of elements that drive the dance, and allow the attunement and synchronization between a couple of dancers. Chapter 15 overviews the Z-process method for detecting changes in parameters of stationary time series models. Chapter 16 extends the Fréchet–Hoeffding copula bounds for circular data to the modified Fréchet and Mardia families of copulas for modelling the dependency of two circular random variables. Chapter 17 considers the topological data analysis approaches to decompose the weighted connectivity network into its symmetric and anti-symmetric components for comparing the anti-symmetric components between prior-to-seizure and post-seizure. Chapter 18 develops an orthogonal impulse response analysis for time-varying covariance functions. Chapter 19 overviews many attractive aspects of optimal transportation and leaves important open problems to achieve robust optimal transportation. Chapter 20 derives the asymptotic distribution of the estimated ruin probability from an insurance surplus model with long memory. Chapter 21 proposes complex-valued time series models, related to two famous circular distributions in direction statistics. Chapter 22 considers the semiparametric estimation problem of an optimal dividend barrier for spectrally negative Levy Processes. Chapter 23 presents a method of fitting a local spectral envelope to heterogeneous sequences for choosing the best-fitting model. Chapter 24 reviews recent studies on the statistical inference for persistent homology of EEG signals to address clinical questions in brain disorders. Chapter 25 deals with the uniformly minimum variance unbiased estimation problems in time series analysis. Chapter 26 proposes a new estimator, without bias or median adjustment, for the coefficient of the Gaussian AR(1) model, which is shown to be better than the MLE in the sense of the third-order MSE.

In addition to the 26 contributed papers, we have included a foreword from Prof. Masanobu Taniguchi, a short vita/a list of publications of Taniguchi, together with a list of Taniguchi's Ph.D. students. We would like to express our thanks to all contributors of this Festschrift. Our deepest thanks go to Sivananth Siva Chandran and Yutaka Hirachi at Springer Nature for their continued support and warm encouragement.

Tokyo, Japan — Yan Liu
Niigata, Japan — Junichi Hirukawa
Hokkaido, Japan — Yoshihide Kakizawa
October 2022

Photos of Masanobu Taniguchi

Benevento, Italy, 2019

Masanobu Taniguchi and his “kids” at his celebration ceremony for receiving Commendation for Science and Technology by the Minister of Education, Culture, Sports, Science and Technology (MEXT), 2022. Left to right front row: Masao Urata, Yoshihide Kakizawa, Masanobu Taniguchi, Junichi Hirukawa, Hiroshi Shiraishi. Back row: Yuichi Goto, Yoichi Nishiyama, Yan Liu, Yujie Xue, Xiaoling Dou, Fumiya Akashi, Hiroaki Ogata, and Takayuki Shiohama

Biography of Masanobu Taniguchi

Education

1974	Bachelor of Science, Osaka University
1976	Master of Science, Osaka University
1981	Doctor of Engineering, Osaka University

Positions Held

1977–1983	Research Associate, Hiroshima University
1983–1990	Associate Professor, Hiroshima University
1990–2003	Associate Professor, Osaka University
2000	Visiting Professor, University of Bristol
2003–2022	Professor, Waseda University
2022	Professor Emeritus, Waseda University

Professional Activities

1987	Fellow of the Institute of Mathematical Statistics, USA
2006–2009	Editor, Journal of the Japan Statistical Society

Publications of Masanobu Taniguchi

List of Books

1. Taniguchi, M. (1991). *Higher Order Asymptotic Theory for Time Series Analysis*. Lecture Notes in Statistics, Vol. 68, Springer-Verlag, Heidelberg.
2. Taniguchi, M. and Kakizawa, Y. (2000). *Asymptotic Theory of Statistical Inference for Time Series*. Springer in Statistics, Springer-Verlag, New York.
3. Taniguchi, M., Hirukawa, J. and Tamaki, K. (2008). *Optimal Statistical Inference in Financial Engineering*. Financial Mathematics Series, Chapman and Hall/CRC, New York.
4. Taniguchi, M., Amano, T., Ogata, H. and Taniai, H. (2014). *Statistical Inference for Financial Engineering*. Springer Briefs in Statistics, Springer-Verlag, Heidelberg.
5. Taniguchi, M., Shiraishi, H., Hirukawa, J., Kato, H. S. and Yamashita, T. (2018). *Statistical Portfolio Estimation*. Financial Mathematics Series, Chapman and Hall/CRC, New York.
6. Liu, Y., Akashi, F. and Taniguchi, M. (2018). *Empirical Likelihood and Quantile Methods for Time Series*. Springer Briefs in Statistics, Springer-Verlag.
7. Akashi, F., Taniguchi, M., Monti, A.C. and Amano, T. (2021). *Diagnostic Methods in Time Series*. Springer Briefs in Statistics, Springer-Verlag.

List of Papers

1. Taniguchi, M. (1978). On a generalization of a statistical spectrum analysis. *Math. Japon.* 23–44.
2. Taniguchi, M. (1979). On estimation of parameters of Gaussian stationary processes. *J. Appl. Prob.* **16** 575–591.
3. Taniguchi, M. (1980). On estimation of the integrals of certain functions of spectral density. *J. Appl. Prob.* **17** 73–83.

4. Taniguchi, M. (1980). On selection of the order of the spectral density model for a stationary process. *Ann. Inst. Stat. Math.* **32** 401–419.
5. Taniguchi, M. (1980). Regression and interpolation for time series. *Recent Developments in Statistical Inference and Data Analysis.* Ed: K. Matusita. 311–321. North-Holland Publishing Company.
6. Taniguchi, M. (1981). An estimation procedure of parameters of a certain spectral density model. *J. Roy. Statist. Soc. B* **43** 34–40.
7. Taniguchi, M. (1981). Robust regression and interpolation for time series. *J. Time Ser. Anal.* **2** 53–62.
8. Hosoya, Y. and Taniguchi, M. (1982). A central limit theorem for stationary processes and the parameter estimation of linear processes. *Ann. Statist.* **10** 132–153. Correction: *Ann. Statist.* **21** 1115–17.
9. Taniguchi, M. (1982). On estimation of the integrals of the fourth order cumulant spectral density. *Biometrika* **69** 117–122.
10. Taniguchi, M. (1982). Note on Clevenson's result in estimating the parameters of a moving average time series. *Math. Japon.* **27** 599–601.
11. Fujikoshi, Y., Morimune, K., Kunitomo, N. and Taniguchi, M. (1982). Asymptotic expansions of the distributions of the estimates of coefficients in a simultaneous equation system. *J. Econometrics.* **18** 191–205.
12. Taniguchi, M. (1983). On the second order asymptotic efficiency of estimators of Gaussian ARMA processes. *Ann. Statist.* **11** 157–169.
13. Taniguchi, M. (1984). Validity of Edgeworth expansions for statistics of time series. *J. Time Ser. Anal.* **5** 37–51.
14. Taniguchi, M. (1985). An asymptotic expansion for the distribution of the likelihood ratio criterion for a Gaussian autoregressive moving average process under a local alternative. *Econometric Theory* **1** 73–84.
15. Taniguchi, M. (1985). Third order efficiency of the maximum likelihood estimator in Gaussian autoregressive moving average processes. *Statistical Theory and Data Analysis.* Ed: K. Matusita. 725–743. North-Holland Publishing Company.
16. Taniguchi, M. (1986). Third order asymptotic properties of maximum likelihood estimators for Gaussian ARMA processes. *J. Multivariate Anal.* **18** 1–31.
17. Taniguchi, M. (1986). Berry–Esseen theorems for quadratic forms of Gaussian stationary processes. *Prob. Th. Rel. Fields* **72** 185–194.
18. Taniguchi, M. (1987). Validity of Edgeworth expansions of minimum contrast estimators for Gaussian ARMA processes. *J. Multivariate Anal.* **21** 1–28.
19. Taniguchi, M. and Krishnaiah, P.R. (1987). Asymptotic distributions of functions of the eigenvalues of sample covariance matrix and canonical correlation matrix in multivariate time series. *J. Multivariate Anal.* **22** 156–176.
20. Nagai, T. and Taniguchi, M. (1987). Walsh spectral analysis of multiple dyadic stationary processes and its applications. *Stochastic Processes and their Applications* **24** 19–30.
21. Taniguchi, M. (1987). Third order asymptotic properties of BLUE and LSE for a regression model with ARMA residual. *J. Time Ser. Anal.* **8** 111–114.

22. Taniguchi, M. (1987). Minimum contrast estimation for spectral densities of stationary processes. *J. Roy. Statist. Soc. B* **49** 315–325.
23. Taniguchi, M. (1988). Asymptotic expansions of the distributions of some test statistics for Gaussian ARMA processes. *J. Multivariate Anal.* **27** 495–511.
24. Taniguchi, M. and Taniguchi, R. (1988). Asymptotic ancillarity in time series analysis. *J. Japan Statist. Soc.* **18** 107–121.
25. Taniguchi, M. (1988). A Berry–Esseen theorem for the maximum likelihood estimator in Gaussian ARMA processes. *Statistical Theory and Data Analysis II*. Ed: K. Matusita. 535–549. North-Holland Publishing Company.
26. Taniguchi, M., Krishnaiah, P.R. and Chao, R. (1989). Normalizing transformations of some statistics of Gaussian ARMA processes. *Ann. Inst. Stat. Math.* **41** 187–197.
27. Taniguchi, M., Zhao, L.C., Krishnaiah, P.R. and Bai, Z.D. (1989). Statistical analysis of dyadic stationary processes. *Ann. Inst. Stat. Math.* **41** 205–225.
28. Taniguchi, M., Swe, M. and Taniguchi, R. (1989). An identification problem for moving average models. *J. Japan Statist. Soc.* **19** 145–149.
29. Taniguchi, M. and Maekawa, K. (1990). Asymptotic expansions of the distributions of statistics related to the spectral density matrix in multivariate time series and their applications. *Econometric Theory* **6** 75–96.
30. Swe, M. and Taniguchi, M. (1990). Higher order asymptotic properties of a weighted estimator for Gaussian ARMA process. *J. Time Ser. Anal.* **12** 83–93.
31. Hosoya, Y. and Taniguchi, M. (1990). Higher order asymptotic theory for time series analysis. *Sugaku* **42** 48–67. (in Japanese).
32. Taniguchi, M. (1991). Third-order asymptotic properties of a class of test statistics under a local alternatives. *J. Multivariate Anal.* **37** 223–238.
33. Taniguchi, M. (1992). An automatic formula for the second order approximation of the distributions of test statistics under contiguous alternatives. *International Statistical Review* **60** 211–225.
34. Taniguchi, M. and Kondo, M. (1993). Non-parametric approach in time series analysis. *J. Time Ser. Anal.* **14** 397–408.
35. Kondo, M. and Taniguchi, M. (1993). Two sample problem in time series analysis. *Statistical Theory and Data Analysis.* Eds: K. Matusita et al. 165–174.
36. Taniguchi, M. andWatanabe, Y. (1994). Statistical analysis of curved probability densities. *J. Multivariate Anal.* **48** 228–248.
37. Taniguchi, M. (1994). Higher order asymptotic theory for discriminant analysis in exponential families of distributions. *J. Multivariate Anal.* **48** 169–187.
38. Zhang, G. and Taniguchi, M. (1994). Discriminant analysis for stationary vector time series. *J. Time Ser. Anal.* **15** 117–126.
39. Kakizawa, Y. and Taniguchi, M. (1994). Asymptotic efficiency of the sample covariances in a Gaussian stationary process. *J. Time Ser. Anal.* **15** 303–311.
40. Kakizawa, Y. and Taniguchi, M. (1994). Higher order asymptotic relation between Edgeworth approximation and saddlepoint approximation. *J. Japan Statist. Soc.* **24** 109–119.

41. Zhang, G. and Taniguchi, M. (1995). Nonparametric approach for discriminant analysis in time series. *J. Nonparametric Statistics* **5** 91–101.
42. Taniguchi, M. and Puri, M.L. (1995). Higher order asymptotic theory for normalizing transformations of maximum likelihood estimators. *Ann. Inst. Stat. Math.* **47** 581–600.
43. Taniguchi, M. (1995). Higher order asymptotic theory for some statistics in time series analysis. *Bull. Inter. Stat. Inst. Proc.* 50th Session Book **2** 481–496.
44. Taniguchi, M., Puri, M.L. and Kondo, M. (1996). Nonparametric approach for non-Gaussian vector stationary processes. *J. Multivariate Anal.* **56** 259–283.
45. Taniguchi, M. and Puri, M.L. (1996). Valid Edgeworth expansions of Mestimators in regression models with weakly dependent residuals. *Econometric Theory* **12** 331–346.
46. Taniguchi, M. (1996). Higher order asymptotic theory for tests and studentized statistics in time series. *Athens Conference on Applied Probability and Time Series, Vol.II: Time Series Analysis.* In Memory of E. J. Hannan. *Lecture Notes in Statistics* **115** 406–419. Springer-Verlag.
47. Taniguchi, M. (1996). On the stochastic difference between the Newton–Raphson and scoring iterative methods. *Festschrift in Honor of Madan L. Puri on the Occasion of his 65th Birthday*. Eds: E. Brunner and M. Denker. 239–255. VSP Utrecht.
48. Sato, T., Kakizawa, Y. and Taniguchi, M. (1998). On large deviation principle of several statistics for short- and long-memory processes. *Austral. New Zealand J. Statist.* **40** 17–29.
49. Kakizawa, Y., Shumway, R.H. and Taniguchi, M. (1998). Discrimination and clustering for multivariate time series. *J. Amer. Statist. Assoc.* **93** 328–340.
50. Taniguchi, M. (1999). Statistical analysis based on functionals of nonparametric spectral density estimators. *Asymptotics, Nonparametrics and Time Series*. Ed: S. Ghosh. 351–394.
51. Nakamura, S. and Taniguchi, M. (1999). Asymptotic theory for the Durbin–Watson statistic under long-memory dependence. *Econometric Theory* **15** 847–866.
52. Hallin, M., Taniguchi, M., Serroukh, A. and Choy, K. (1999). Local asymptoticnormality for regression models with long-memory disturbance. *Ann. Statist.* **27** 2054–2080.
53. Tsuga, N., Taniguchi, M. and Puri, M.L. (2000). An estimation method in time series errors-in-variables models. *J. Japan Statist. Soc.* **30** 75–87.
54. Choy, K. and Taniguchi, M. (2001). Stochastic regression model with dependent disturbances. *J. Time Ser. Anal.* **22** 175–196.
55. Taniguchi, M. (2001). On large deviation asymptotics of some tests in time series analysis. *J. Statist. Plan. Inf*. **97** Special Issue on Rao's Score Test. 191–200.
56. Chandra, A. and Taniguchi, M. (2001). Estimating functions for nonlinear time series models. *Ann. Inst. Stat. Math.* **53** 125–141.

57. Shiohama, T. and Taniguchi, M. (2001). Sequential estimation for a functional of the spectral density of a Gaussian stationary process. *Ann. Inst. Stat. Math.* **53** 142–158.
58. Choi, I-B. and Taniguchi, M. (2001). Misspecified prediction for time series. *J. Forecasting* **20** 543–564.
59. Sakiyama, K., Taniguchi, M. and Puri, M.L. (2002). Asymptotics of tests for a unit root in autoregression. The special issue of *J. Statist. Plan. Inf*. in honor of Professor C.R. Rao. **108** 351–364.
60. Taniguchi, M. and Puri, M.L. (2002). LAN based asymptotic theory for time series. Proceedings of the 4th International Triennial Calcutta Symposium. *Calcutta Statistical Association Bulletin* **52** 143–170.
61. Chandra, A. and Taniguchi, M. (2003). Asymptotics of rank order statistics for ARCH residual empirical processes. *Stochastic Processes and Their Applications* **104** 301–324.
62. Sakiyama, K. and Taniguchi, M. (2003). Testing composite hypotheses for locally stationary processes. *J. Time Ser. Anal.* **24** 483–504.
63. Choi, I-B. and Taniguchi, M. (2003). Prediction problems for square-transformed stationary processes. *Statistical Inference for Stochastic Processes* **6** 43–64.
64. Vellaisamy, P., Sankar, S. and Taniguchi, M. (2003). Estimation and design of sampling plans for monitoring dependent production processes. *Methodology and Computing in Applied Probability* **5** 85–108.
65. Shiohama, T., Taniguchi, M. and Puri, M.L. (2003). Asymptotic estimation theory of change-point problems for time series regression models and its applications. *Probability, Statistics and their Applications.* Papers in Honors of Rabi Bhattacharya, Eds: K. Athreya, M. Majumdar, M.L. Puri and E. Waymire, *IMS Lecture Notes* **41** 257–284.
66. Taniguchi, M., Kees Jan van Garderen and Puri, M.L. (2003). Higher order asymptotic theory for minimum contrast estimators of spectral parameters of stationary processes. *Econometric Theory* **19** 984–1007.
67. Shiohama, T. and Taniguchi, M. (2004). Sequential estimation for time series regression models. *J. Statist. Plan. Inf*. **123** 295–312.
68. Sakiyama, K. and Taniguchi, M. (2004). Discriminant analysis for locally stationary processes. *J. Multivariate Anal.* **90** 282–300.
69. Lee, S. and Taniguchi, M. (2004). Asymptotic theory for ARCH-SM models: LAN and residual empirical processes. *Statistica Sinica* **15** 215–234.
70. Taniguchi, M. (2004). Recent developments in statistical asymptotic theory for time series analysis. *Ouyo-Suri* **14** 13–23. (in Japanese).
71. Taniguchi, M. and Hirukawa, J. (2005). The Stein–James estimator for short and long-memory Gaussian processes. *Biometrika* **92** 737–746.
72. Chandra, A. S. and Taniguchi, M. (2006). Minimum -divergence estimation for ARCH models. *J. Time Ser. Anal.* **27** 19–39.
73. Hirukawa, J. and Taniguchi, M. (2006). LAN theorem for non-Gaussian locally stationary processes and its applications. *J. Statist. Plan. Inf*. **136** 640–688.

74. Taniguchi, M., Maeda, K. and Puri, M.L. (2006). Statistical analysis of a class of factor time series models. The special issue of *J. Statist. Plan. Inf*. **136** 2367–2380, in honor of Professor Shanti Gupta.
75. Kato, H., Taniguchi, M. and Honda, M. (2006). Statistical analysis for multiplicatively modulated nonlinear autoregressive model and its applications to electrophysiological signal analysis in humans. *IEEE Trans. Signal Processing* **54** 3414–3425.
76. Senda, M. and Taniguchi, M. (2006). James-Stein estimators for time series regression models. *J. Multivariate Anal*. **97** 1984–1996 Fujikoshi Volume.
77. Tamaki, K. and Taniguchi, M. (2007). Higher order asymptotic option valuation for non-Gaussian dependent return. *J. Statist. Plan. Inf*. Special Issue in honor of Professor Madan L. Puri. **137** 1043–1058.
78. Shiraishi, H. and Taniguchi, M. (2007). Statistical estimation of optimal portfolios for locally stationary returns of Assets. *International Journal of Theoretical and Applied Finance* **10** 129–154.
79. Taniguchi, M., Shiraishi, H. and Ogata, H. (2007). Improved estimation for the autocovariances of a Gaussian stationary process. *Statistics* **41** 269–277.
80. Amano, T. and Taniguchi, M. (2008). Asymptotic efficiency of conditional least squares estimators for ARCH models. *Statist. Prob. Letters* **78** 179–185.
81. Taniguchi, M. (2008). Non-regular estimation theory for piecewise continuous spectral densities. *Stochastic Processes and Their Applications* **118** 153–170.
82. Taniai, H. and Taniguchi, M. (2008). Statistical estimation errors of VaR under ARCH returns. *J. Statist. Plan. Inf*. Special Issue in honor of Professor J. Ogawa. **138** 3568–3577.
83. Kato, H., Taniguchi, M., Nakatani, T. and Amano, S. (2008). Classification and similarity analysis of fundamental frequency patterns in infant spoken language acquisition. *Statistical Methodology* **5** 187–208.
84. Shiraishi, H. and Taniguchi, M. (2008). Statistical estimation of optimal portfolios for non-Gaussian dependent returns of assets. *J. Forecasting* **27** 193–215.
85. Hirukawa, J., Kato, H., Tamaki, K. and Taniguchi, M. (2008). Generalized information criteria in model selection for locally stationary processes. *J. Jap. Statist. Soc*. **38** 157–171.
86. Ogata, H. and Taniguchi, M. (2009). Cressie–Read power divergence statistics for non Gaussian stationary processes. *Scandinavian J. Statistics* **36** 141–156.
87. Nishikawa, M. and Taniguchi, M. (2009). Discriminant analysis for dynamics of stable processes. *Statistical Methodology* **6** 82–96.
88. Taniguchi, M., Ogata, H. and Shiraishi, H. (2009). Preliminary test estimation for regression models with long-memory disturbance. *Commun. Stat*. **38** 1–12.
89. Shiraishi, H. and Taniguchi, M. (2009). Statistical estimation of optimal portfolios depending on higher order cumulants. *Annales de I'I.S.U.P*. 3–18.
90. Taniguchi, M. and Amano, T. (2009). Systematic approach for portmanteau tests in view of Whittle likelihood ratio. *J. Jap. Stat. Soc*. **39** 177–192.

91. Ishioka, T., Kawamura, S., Amano, T. and Taniguchi, M. (2009). Spectral analysis for intrinsic time processes. *Statist. Probab. Letters* **79** 2389–2396.
92. Watanabe, T., Shiraishi, H. and Taniguchi, M. (2010). Cluster analysis for stable processes. *Commun. Stat*. **39**1630–1642.
93. Ogata, H. and Taniguchi, M. (2010). Empirical likelihood approach for non-Gaussian vector stationary processes and its application to minimum contrast estimation. *Austral. New Zeal. J. Statist*. **52** 451–468.
94. Kanai, H., Ogata, H. and Taniguchi, M. (2010). Estimating function approach for CHARN models. *Metron* **68** 1–21.
95. Hatai, I., Shiraishi, H. and Taniguchi, M. (2010). Statistical testing for asymptotic no-arbitrage in financial markets. *J. Statistics: Advance in Theory and Applications* **3** 27–42.
96. Naito,T., Asai, K., Amano,T. andTaniguchi, M. (2010). Local Whittle likelihood estimators and tests for non-Gaussian stationary processes. *Statistical Inference for Stochastic Processes* **13** 163–174.
97. Shiohama, T., Hallin, M., Veredas, D. and Taniguchi, M. (2010). Dynamic portfolio optimization using generalized dynamic conditional heteroskedastic factor models. *J. Japan Statist. Soc*. **40** 145–166.
98. Hirukawa, J., Taniai, H., Hallin, M. and Taniguchi, M. (2010). Rank-based inference for multivariate nonlinear and long-memory time series models. *J. Japan Statist. Soc*. **40** 167–187.
99. Taniguchi, M. (2010). Asymptotic theory for time series analysis. *Sugaku* **62** 50–74 Iwanami. (in Japanese).
100. Taniguchi, M. (2011). Optimal statistical inference in financial engineering. *International Encyclopedia of Statistical Science*. Springer.
101. Amano, T. and Taniguchi, M. (2011). Control variate method for stationary processes. *J. Econometrics* **165** 20–29.
102. Maeyama, Y., Tamaki, K. and Taniguchi, M. (2011). Preliminary test estimation for spectra. *Statist. and Probab. Letters* **81** 1580–1587.
103. Amano, T., Kato, T. and Taniguchi, M. (2012). Statistical estimation for CAPM with long-memory dependence. *Advances in Decision Sciences: Special Issue on "Statistical Estimation of Portfolios for Dependent Financial Returns*. Lead Guest Editor: Taniguchi, M. Article ID 571034, 12 pages.
104. Taniguchi, M. and Hirukawa, J. (2012). Generalized information criterion. *J. Time Ser. Anal*. **33** 287–297.
105. Taniguchi, M., Tamaki, K., DiCiccio, T.J. and Monti, A.C. (2012). Jackknifed Whittle estimators. *Statistica Sinica* **22** 1287–1304.
106. Taniguchi, M., Petkovic, A., Kase, T., DiCiccio, T.J. and Monti, A.C. (2012). Robust portfolio estimation under skew-normal return processes. *The European Journal of Finance* **21** 1–22.
107. Hamada, K., Dong, W-Y. and Taniguchi, M. (2012). Statistical portfolio estimation under the utility function depending on exogenous variables. *Advances in Decision Sciences: Special Issue on "Statistical Estimation of Portfolios for Dependent Financial Returns"*. Lead Guest Editor: M. Taniguchi. Article ID 127571, 16 pages.

108. Shiraishi, H., Ogata, H., Amano, T., Patilea, V., Veredas, D. and Taniguchi, M. (2012). Optimal portfolios with end-of-period target. *Advances in Decision Sciences: Special Issue on "Statistical Estimation of Portfolios for Dependent Financial Returns"*. Lead Guest Editor: M. Taniguchi. Article ID 703465, 13 pages.
109. Nagahata, H., Suzuki, T., Usami, Y., Yokoyama, A., Ito, J., Hasegawa, H. and Taniguchi, M. (2012). Various problems in time series analysis. *ASTE*, *Research Institute for Science and Engineering, Waseda University Special Issue "Financial and Pension Mathematical Science"*. Editor: M. Taniguchi. **8** 1–30.
110. Hamada, K. and Taniguchi, M. (2012). Multi-step ahead portfolio estimation for dependent return processes. *ASTE*, *Research Institute for Science and Engineering, Waseda University Special Issue "Financial and Pension Mathematical Science"*. Editor: M. Taniguchi. **8** 47–56.
111. Taniai, H., Usami, T., Suto, N. and Taniguchi, M. (2012). Asymptotics of realized volatility with non-Gaussian ARCH(1) microstructure noise. *J. Financial Econometrics* **10** 617–636.
112. Yokozuka, J. and Taniguchi, M. (2013). Asymptotic theory for near unit root autoregression with infinite innovation variance. *ASTE, Research Institute for Science and Engineering, Waseda University Special Issue "Financial and Pension Mathematical Science"*. Editor: M. Taniguchi. **9** 1–24.
113. Shoyama, Y., Shirataki, T., Nagayasu, S. and Taniguchi, M. (2013). Statistical asymptotic theory for financial time series. *ASTE*, *Research Institute for Science and Engineering, Waseda University Special Issue "Financial and Pension Mathematical Science"*. Editor: M. Taniguchi. **9** 25–44.
114. Hamada, K. and Taniguchi, M. (2013). Constrained Whittle estimators and shrunk Whittle estimators. *ASTE*, *Research Institute for Science and Engineering, Waseda University Special Issue "Financial and Pension Mathematical Science"*. Editor: M. Taniguchi. **9** 55–66.
115. Ando, H., Yamashita, T. and Taniguchi, M. (2013). Portfolios influenced by causal variables under long-range dependent returns. *ASTE*, *Research Institute for Science and Engineering, Waseda University Special Issue "Financial and Pension Mathematical Science"*. Editor: M. Taniguchi. **9** 67–84.
116. Kobayashi, I., Yamashita, T. and Taniguchi, M. (2013). Portfolio estimation under causal variables. *ASTE*, *Research Institute for Science and Engineering, Waseda University Special Issue "Financial and Pension Mathematical Science"*. Editor: M. Taniguchi. **9** 85–100.
117. Hamada, K. and Taniguchi, M. (2014). Shrinkage estimation and prediction for time series. *ASTE*, *Research Institute for Science and Engineering, Waseda University Special Issue "Financial and Pension Mathematical Science"*. Editor, M. Taniguchi. **10** 3–18.
118. Hamada, K. and Taniguchi, M. (2014). Statistical portfolio estimation for non-Gaussian return process. *ASTE*, *Research Institute for Science and Engineering, Waseda University Special Issue "Financial and Pension Mathematical Science"*. Editor: M. Taniguchi. **10** 69–86.

119. Iizuka, K., Matsuura, K. and Taniguchi, M. (2015). Asymptotic properties of the sample variance matrix for high dimensional dependent data. *ASTE, Research Institute for Science and Engineering, Waseda University Special Issue "Financial and Pension Mathematical Science"*. Editor: M. Taniguchi, **12** 1–10.
120. Haga, K., Koji, M. and Taniguchi, M. (2015). High dimensional statistical analysis for time series. *ASTE*, *Research Institute for Science and Engineering, Waseda University Special Issue "Financial and Pension Mathematical Science"*. Editor: M. Taniguchi. **12** 11–30.
121. Suto, Y., Liu, Y. and Taniguchi, M. (2015). Asymptotic theory of parameter estimation by a contrast function based on interpolation error. *Statistical Inference for Stochastic Processes* **19** 93–110.
122. Akashi, F., Liu, Y. and Taniguchi, M. (2015). An empirical likelihood approach for symmetric α-stable processes. *Bernoulli* **21** 2093–2119.
123. Koike, R., Dou, X., Taniguchi, M. and Xue, Y. (2016). Granger causality test via Box-Cox transformations. *ASTE*, *Research Institute for Science and Engineering, Waseda University Special Issue "Financial and Pension Mathematical Science"*. Editor: M. Taniguchi. **13** 3–18.
124. Suto, Y. and Taniguchi, M. (2016). Shrinkage interpolation for stationary processes. *ASTE*, *Research Institute for Science and Engineering, Waseda University Special Issue "Financial and Pension Mathematical Science"*. Editor: M. Taniguchi. **13** 35–42.
125. Giraitis, L., Taniguchi, M. and Taqqu, M.S. (2017). Asymptotic normality of quadratic forms of martingale differences. *Statistical Inference for Stochastic Processes* **20** 315–327.
126. Kato, Solvang H. and Taniguchi, M. (2017). Portfolio estimation for spectral density of categorical time series data. *Far East J. theoretical statistics* **53** 19–33.
127. Kato, Solvang H. and Taniguchi, M. (2017). Microarray analysis using rank order statistics for ARCH residual empirical process. *Open Journal of Statistics* **7** 54–71.
128. Liu,Y., Nagahata, H., Uchiyama, H. and Taniguchi, M. (2017). Discriminant and cluster analysis of possibly high-dimensional time series data by a class of disparities. *Com. Stat.-Simul. Computation* **46** 8014–8027.
129. Chen, C.W.S., Hsu, Y.C. and Taniguchi, M. (2017). Discriminant analysis by quantile regression with application on the climate change problem. *J. Stat. Plan. Inf*. **187** 17–27.
130. Akashi, F., Odashima, H., Taniguchi, M. and Monti, A.C. (2018). A new look at portmanteau test. *Sankhya A* **80** 121–137.
131. Monti, A.C. and Taniguchi, M. (2018). Adjustments for a class of tests under nonstandard conditions. *Statistica Sinica* **28** 1437–1458.
132. Liu, Y., Tamura, Y. and Taniguchi, M. (2018). Asymptotic theory of test statistic for sphericity of high-dimensional time series. *J. Time Ser. Anal.* **39** 402–416.

133. Shiraishi, H., Taniguchi, M. and Yamashita, T. (2018). Higher-order asymptotic theory of shrinkage estimation for general statistical models. *J. Multivariate Anal.* **166** 198–211.
134. Nagahata, H. and Taniguchi, M. (2018). Analysis of variance for multivariate time series. *Metron* **76** 69–82.
135. Nagahata, H. andTaniguchi, M. (2018). Analysis of variance for high-dimensional time series. *Statistical Inference for Stochastic Processes* **21** 455–468.
136. Giraitis, L., Taniguchi, M. and Taqqu, M.S. (2018). Estimation pitfalls when the noise is not i.i.d. *Jpn. J. Stat. Data Sci.* **1** 59–80.
137. Goto, Y. and Taniguchi, M. (2019). Robustness of zero crossings estimator. *J. Time Series Anal.* **40** 815–830.
138. Liu, Y., Xue, Y. and Taniguchi, M. (2020). Robust linear interpolation and extrapolation of stationary time series in L_P. *J. Time Ser. Anal.* **41** 229–248.
139. Goto, Y. and Taniguchi, M. (2020). Discriminant analysis based on binary time series. *Metrika* **83** 569–595.
140. Xue, Y. and Taniguchi, M. (2020). Local Whittle likelihood approach for generalized divergence. *Scand. J. Statist*. **4** 182–195.
141. Akashi, F., Taniguchi, M. and Monti, A.C. (2020). Robust causality test of infinite variance processes. *J. Econometrics* **216** 235–245.
142. Xue, Y. and Taniguchi, M. (2020). Modified LASSO estimators for time series regression models with dependent disturbances. *Statistical Methods and Applications* **29** 845–869.
143. Taniguchi, M., Kato, S., Ogata, H. and Pewsey, A. (2020). Models for circular data from time series spectra. *J. Time Series Anal.* **41** 808–829.
144. Chen, C.W.S., Lee, S., Dong, M.C. and Taniguchi, M. (2021). What factors drive the satisfaction of citizens on governments' responses to COVID-19? *International J. of Infectious Diseases* **102** 327–331.
145. Akashi, F., Taniguchi, M. and Tanida, Y. (2021). Estimation of linear functional of large spectral density matrix and application to Whittle's approach, *Jpn. J. Stat. Data Sci.* **4** 449–474.
146. Liu, Y. and Taniguchi, M. (2021). Minimax estimation for time series models. *Metron* **79** 353–359.
147. Liu,Y., Tanida, Y. and Taniguchi, M. (2021). Shrinkage estimation for multivariate time series. *Statistical Inference for Stochastic Processes* **24** 733–751.
148. Goto, Y., Kaneko, T., Kojima, S. and Taniguchi, M. (2022). Likelihood ratio processes under nonstandard settings. *Theory Probab. Appl.* **67** 246–260.
149. Xu, X., Liu, Y. and Taniguchi, M. (2023). Higher order asymptotic theory for minimax estimation of time series. *J. Time Series Anal.* **44** 247–257. https://doi.org/10.1111/jtsa.12661.
150. Xu, X., Li, Z. and Taniguchi, M. (2022). Comparison between the exact likelihood and Whittle likelihood for moving average processes. *STATISTICA* **82** 3–13. https://doi.org/10.6092/issn.1973-2201/13609.

151. Goto, Y., Arakaki, K., Liu, Y. and Taniguchi, M. (2023). Homogeneity tests for one-way models with dependent errors under correlated groups. *TEST* **32** 163–183. https://doi.org/10.1007/s11749-022-00828-9.
152. Goto, Y., Suzuki, K., Xu, X. and Taniguchi, M. (2023). Two-way random ANOVA model with interaction disturbed by dependent errors. *Ann. Inst. Statist. Math.* **75** 511–532.
153. Taniguchi, M. and Xue Y. (2022). Hellinger distance estimation for non-regular spectra. To appear in *Theory Probab. Appl.*
154. Fujimori, K., Goto, Y., Liu, Y. and Taniguchi, M. (2023). Sparse principal component analysis for high-dimensional stationary time series. To appear in *Scand. J. Stat.*

Students of Masanobu Taniguchi

MYINT SWE

Higher order asymptotic investigations of weighted estimators for Gaussian ARMA processes 1990

GUOQIANG ZHANG (KOKYO NAGA)

Discriminant analysis for time series by spectral distribution 1995

YOSHIHIDE KAKIZAWA

Statistical inference and asymptotic theory for stationary time series 1996

KENJI SAKIYAMA

Statistical analysis for nonstationary time series 2002

IN-BONG CHOI

Misspecified prediction of time series and its applications 2002

SUBHASH AJAY CHANDRA

Statistical inference for nonlinear time series models 2003

TAKAYUKI SHIOHAMA

Sequential analysis and change-point problems in time series 2003

JUNICHI HIRUKAWA

LAN theorem for non-Gaussian locally stationary processes and their discriminant analysis 2005

KENICHIRO TAMAKI

Statistical higher order asymptotic theory and its applications to analysis of financial time series 2006

HIROSHI SHIRAISHI

Statistical estimation of optimal portfolios for dependent returns of assets 2007

HIROAKI OGATA

Empirical likelihood method for time series analysis 2007

TOMOYUKI AMANO

Various statistical methods in time series analysis 2008

HIROYUKI TANIAI

Inference for the quantiles of ARCH processes (Supervisor: Marc Hallin) 2009

FUMIYA AKASHI

Frequency domain empirical likelihood method for stochastic processes with infinite variance and its application to discriminant analysis 2015

YAN LIU

Asymptotic theory for non-standard estimating function and self-normalized method in time series analysis 2015

YOSHIHIRO SUTO

Interpolation and shrinkage estimation for time series 2017

HIDEAKI NAGAHATA

Theory of multivariate time series analysis and its application 2019

YUJIE XUE

Robust prediction and sparse estimation for long memory time series 2021

YUICHI GOTO

Asymptotic theory of statistical inference for binary time series and count time series and its applications 2021

Contents

Contributors

Fumiya Akashi University of Tokyo, Hongo Bunkyo-ku, Tokyo, Japan

Narayana Balakrishna Cochin University of Science and Technology, Kochi, KL, India

Ngai Hang Chan City University of Hong Kong, Kowloon Tong, Hong Kong SAR, China

Ying Chen National University of Singapore, Singapore, Singapore

Fuxia Cheng Illinois State University, Normal, IL, USA

Enzo D'Innocenzo Vrije Universiteit Amsterdam, Amsterdam, Netherlands

Violetta Dalla National and Kapodistrian University of Athens, Athens, Greece

Victor De Oliveira The University of Texas at San Antonio, San Antonio, TX, USA

Anass El Yaagoubi King Abdullah University of Science and Technology, Thuwal, Thuwal, Kingdom of Saudi Arabia

Christian Francq CREST-ENSAE and University of Lille, Villeneuve d'Ascq cedex, France

Liudas Giraitis Queen Mary University of London, London, UK

Yuichi Goto Kyushu University, Nishi-ku, Fukuoka, Japan

Marc Hallin ECARES and Département de Mathématique, Université Libre de Bruxelles CP 114/4, Bruxelles, Belgium

Junichi Hirukawa Faculty of Science, Niigata University, Nishi-ku, Niigata, Japan

Yoshihide Kakizawa Hokkaido University, Kita-ku, Sapporo, Hokkaido, Japan

Hira L. Koul Michigan State University, East Lansing, MI, USA

Hao Lei National University of Singapore, Singapore, Singapore

Yuxin Li The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China

Hang Liu International Institute of Finance, School of Management, University of Science and Technology of China, Anhui, China

Kailiang Liu National University of Singapore, Singapore, Singapore

Yan Liu Waseda University, Shinjuku-ku, Tokyo, Japan

Alessandra Luati Imperial College London, London, UK

Nao Mimoto The University of Akron, Akron, OH, USA

Anna Clara Monti University of Sannio, Benevento, Italy

Kanchan Mukherjee Department of Mathematics and Statistics, Lancaster University, Lancaster, United Kingdom

Naoya Murata Waseda University, Shinjuku-ku, Tokyo, Japan

Shota Nakamura Waseda University, Shinjuku-ku, Tokyo, Japan

Ilia Negri University of Calabria, Rende, CS, Italy

Hiroaki Ogata Tokyo Metropolitan University, Hachioji, Tokyo, Japan

Hernando Ombao King Abdullah University of Science and Technology, Thuwal, Thuwal, Kingdom of Saudi Arabia

Valentin Patilea CREST Ensai, Campus de Ker-Lann, BRUZ Cedex, France

Tommaso Proietti Università di Roma Tor Vergata, Roma, Italy

Hamdi Raïssi Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile

Elvezio Ronchetti University of Geneva, Geneva, Switzerland

Shunsuke Sakai Graduate School of Science and Technology, Niigata University, Niigata, Japan

Pietro Scalera Conservatorio di Musica di San Pietro a Majella, Naples, Italy

Yasutaka Shimizu Waseda University, Shinjuku-ku, Tokyo, Japan

Takayuki Shiohama Nanzan University, Nagoya, Japan

Hiroshi Shiraishi Keio University, Kohoku-ku, Yokohama, Japan

Tony Sit The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China

David S. Stoffer University of Pittsburgh, Pittsburgh, Pennsylvania, USA

Masanobu Taniguchi Waseda University, Shinjuku-ku, Tokyo, Japan

Murad S. Taqqu Boston University, Boston, MA, USA

Thomas Verdebout ECARES and Département de Mathématique, Université libre de Bruxelles (ULB), Boulevard du Triomphe, Brussels, Belgium

Yuan Wang University of South Carolina, Columbia, South Carolina, USA

Xiaofei Xu Wuhan University, Wuhan, Hubei, P.R. China

Yujie Xue Waseda University, Shinjuku-ku, Tokyo, Japan

Jean-Michel Zakoian University of Lille and CREST-ENSAE, Palaiseau, France

Chapter 1
Spatial Median-Based Smoothed and Self-Weighted GEL Method for Vector Autoregressive Models

Fumiya Akashi

Abstract This paper considers the estimation and testing problems for the coefficient matrices of vector autoregressive models, including infinite variance processes. The self-weighted generalized empirical likelihood (GEL) estimator and test statistic for the hypotheses of the nonlinear restriction of the parameters are proposed. The limiting distributions of the GEL estimator and test statistic are derived under mild distributional conditions for the innovation processes. The proposed testing procedure does not require any prior information for the nuisance parameters of the process, such as the behavior of the distributional tail of the innovation processes; hence, the results in this paper provide a feasible testing procedure for the hypothesis. Simulation experiments illustrate the finite sample performance of the proposed methods.

1.1 Introduction

In various fields including economics and financial engineering, we often encounter data that exhibit heavy tails. Such data are poorly grasped by Gaussian models, and a variety of infinite variance time series models have been proposed by several authors. For example, [5–7] studied limit theorems for stochastic processes generated by random variables with regularly varying tails. Especially [7] investigated the limit behavior of the sample covariance and autocorrelation functions for the infinite variance processes. On the other hand, one of the famous classes of infinite variance processes is the stable processes. The book [32] summarized the fundamental properties of stable processes. In the frequency domain approach, [17] considered parameter estimation for autoregressive moving average (ARMA) processes with stable innovations, and investigated the asymptotics of the Whittle estimator. The papers [11, 18] studied the properties of the periodograms and their integral functionals for the stable processes. These results showed the sharp contrast between finite and infinite variance cases in the sense of the different limit distributions and

F. Akashi (✉)
University of Tokyo, Hongo Bunkyo-ku, Tokyo 113-0033, Japan
e-mail: akashi@e.u-tokyo.ac.jp

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_1

rates of convergence of fundamental statistics. However, the limit distributions of the estimators and test statistics in these papers were represented by sums of stable random variables; hence, it was difficult to develop a feasible inference procedure based on the results in practice. For example, [35] derived the limit distribution of the least square estimator for autoregressive (AR) models with generalized autoregressive conditional heteroscedastic (GARCH)-type errors, but the rate of convergence contained an unknown tail index of the error term, and the limit distribution was non-Gaussian. Moreover, the probability density functions of the stable distributions cannot be represented in a closed form except for a few special cases, and it is also difficult to apply the likelihood ratio statistics directly. To overcome such difficulties, a useful weighting method has been proposed by [14], which is referred to as a *self-weighting method*. The self-weighting method for scalar AR models proposed by [14] enables us to control the heavy tails of stochastic processes and recovers the asymptotic normality of estimators. The paper [28] extended the self-weighting approach toward infinite variance ARMA models. Furthermore, the self-weighting approach is widely applicable to various complex models; for example, see [15, 36, 37] for ARMA-GARCH-type processes.

Another approach to circumvent this problem is the empirical likelihood (EL) method of [27]. The EL method can be used to build a nonparametric likelihood ratio function and provides a feasible statistical framework to carry out hypothesis testing under mild conditions for the population distributions of data-generating processes. Notably, the EL and its generalized version (GEL) have been studied in time and frequency domains. The paper [30] considered the EL based on estimating equations, and [22] gave a unified view of the higher order properties of the GEL estimators. In the context of time series analysis, [9] extended the EL approach to weakly dependent processes via a data blocking technique. In the frequency domain, [20] provided a version of the EL test statistic based on the Whittle estimator. The frequency domain approach was extended by [23, 25], respectively, to long-memory processes and non-Gaussian multivariate processes. A review of the developments of the EL approach in time series analysis was provided by [24].

Although the self-weighting and EL approaches are useful in the analysis of heavy-tailed time series models, most of the researches considered scalar processes because the self-weighting approach is always used with the least absolute deviation (LAD) regression; for the self-weighted EL and GEL approach in univariate time series models, see, for example, [1, 12, 13]. To extend the self-weighting and EL approaches to multivariate stochastic processes, we utilize the spatial median concept, which is a natural multivariate extension of univariate medians. There is a substantial body of literature on the spatial median concept and its related topics. For example, [4] discussed an estimation problem of multivariate median, and [3] showed the consistency and asymptotic normality of the least distance estimator, which is defined as the minimizer of the sum of the Euclidean norm of residuals. For some reviews on spatial median approaches, see [26, 33].

Motivated by the aforementioned studies, we extend the self-weighted GEL approach toward vector AR (VAR) models based on the spatial median. Although the self-weighted LAD method provides a robust statistical inference procedure, a lack of smoothness of objective functions causes problems in stochastic expansions and computations. To avoid this inconvenience, we also incorporate the smoothed EL approach of [34] into the GEL. The extension in this paper is highly nontrivial and contains a lot of novel aspects.

The rest of the paper is organized as follows. Section 1.2 defines the model in this paper and introduces the concept of the spatial median. Some fundamental results of the spatial median are presented. This section also introduces the GEL function. Section 1.3 provides the main results of this paper. Section 1.4 illustrates the finite sample performance of the proposed methods via simulation experiments. Auxiliary lemmas and technical proofs are relegated to Sect. 1.5.

In this paper, we use the following notations. The symbols $\mathbb{R}$, $\mathbb{Z}$, $\mathbb{N}$, and $\mathbb{C}$ denote the set of all real numbers, integers, natural numbers, and complex numbers, respectively, and $\mathbb{R}^+ := \{x : x \in \mathbb{R}, x \geq 0\}$. For any matrix A, we define $\|A\| := \sqrt{\mathrm{tr}(A^\top A)}$, where the symbol $\top$ is the transpose. For a differentiable function $f : \mathbb{R}^d \to \mathbb{R}$, we denote $\nabla f(x) := \partial f(x)/\partial x$. The symbol $A \otimes B$ denotes the Kronecker product of matrices A and B. Also, I_k, 0_k, and $O_{k\times l}$ denote, respectively, the $k \times k$ identity matrix, the $k \times 1$ zero vector, and the $k \times l$ zero matrix.

1.2 Settings

1.2.1 Model and Spatial Median

Let $\{X(1), ..., X(n)\}$ be an observed stretch from the d-dimensional VAR process of order p, i.e.,

$$X(t) = A_1 X(t-1) + \cdots + A_p X(t-p) + U(t), \tag{1.1}$$

where A_i is a $d \times d$ constant matrix ($i = 1, ..., p$) and $\{U(t) : t \in \mathbb{Z}\}$ is a sequence of i.i.d. random vectors with the density function f_U. Throughout this paper, we assume $d \geq 2$. By denoting

$$\mathbb{X}_{t-1} := [X(t-1)^\top, ..., X(t-p)^\top]^\top \in \mathbb{R}^{dp} \text{ and } \mathbb{A} := [A_1, ..., A_p] \in \mathbb{R}^{d\times dp},$$

we define the residuals

$$U(t;\theta) := X(t) - \mathbb{A}\mathbb{X}_{t-1} \quad (t = p+1, ..., n),$$

where $\theta = \mathrm{vec}(\mathbb{A})$. For any vector $\phi \in \mathbb{R}^m$ ($m := d^2 p$), define an operator $\Psi : \mathbb{R}^m \to \mathbb{R}^{d\times dp}$ as

$$\Psi(\phi) := \left[\psi_1^{(1)}(\phi) \cdots \psi_1^{(d)}(\phi) \cdots \psi_p^{(1)}(\phi) \cdots \psi_p^{(d)}(\phi)\right],$$

where $\psi_i^{(j)}(\phi)$ is a $d \times 1$ vector ($i = 1, ..., p$, $j = 1, ..., d$), which satisfies

$$\phi = (\psi_1^{(1)}(\phi)^\top, ..., \psi_1^{(d)}(\phi)^\top, ..., \psi_p^{(1)}(\phi)^\top, ..., \psi_p^{(d)}(\phi)^\top)^\top.$$

Evidently, we have $\mathrm{vec}(\Psi(\phi)) = \phi$ for any $\phi \in \mathbb{R}^m$ and $\Psi(\mathrm{vec}(M)) = M$ for any $M \in \mathbb{R}^{d\times dp}$. That is, Ψ reconstructs a $d \times dp$ matrix from an m-dimensional vector. We denote the true value of $\mathbb{A}$ as $\mathbb{A}_0$, and denote $\theta_0 := \mathrm{vec}(\mathbb{A}_0)$. Now, define the *spatial sign function* $\mathrm{Sign}(\cdot)$ as

$$\mathrm{Sign}(u) = \begin{cases} u/\|u\| & (u \neq 0_d), \\ 0_d & (u = 0_d). \end{cases}$$

In this paper, we assume the following conditions for $U(t)$ and the coefficient matrices $A_1, ..., A_p$.

Assumption 1.1

(i) $E(\mathrm{Sign}(U(t))) = 0_d$;
(ii) $f_U(x)$ is continuous on $\mathbb{R}^d$ and bounded by a constant uniformly in $x \in \mathbb{R}^d$;
(iii) $P(\alpha^\top U(t) = 0) < 1$ for any non-zero vector $\alpha \in \mathbb{R}^d$;
(iv) $E(\|U(t)\|^\delta) < \infty$ for some $\delta > 0$;
(v) for any $\theta \in \mathbb{R}^m$, let $A_i(\theta)$ be a $d \times d$ matrix ($i = 1, ..., p$) satisfying $[A_1(\theta), ..., A_p(\theta)] = \Psi(\theta)$, we have $\det(I_d - \sum_{j=1}^p z^j A_j(\theta)) \neq 0$ for all $z \in \mathbb{C}$ such that $|z| \leq 1$ and for all $\theta \in \Theta$, where Θ is a compact subset of $\mathbb{R}^m$.

Denote the ith component of $U(t)$ by $U_i(t)$ ($i = 1, ..., d$). If $U_1(t), ..., U_d(t)$ are mutually independent and each $U_i(t)$ has a zero-marginal median, then Assumption 1.1(i) is satisfied. More generally, if a matrix B satisfies $B^\top B = cI_d$ for some constant $c > 0$ and a random vector Z satisfies Assumption 1.1(i), then $U(t) =_d BZ$, where $=_d$ stands for equal in distribution, also satisfies Assumption 1.1(i). Assumption 1.1(ii) is required to bound the moment of $\|U(t)\|^{-r}$ for $r \in (0, d)$. Assumption 1.1(iii) guarantees the positive definiteness of the matrices

$$A_U := \int \frac{1}{\|u\|}\left(I_d - \frac{uu^\top}{\|u\|^2}\right) f_U(u)du \text{ and } B_U := \int \frac{uu^\top}{\|u\|^2} f_U(u)du, \qquad (1.2)$$

which appear in the stochastic expansions of the GEL statistics. Assumption 1.1(iv) and (v) are standard for the stationarity of the model (1.1). Assumption 1.1 allows the infinite variance of the error term so that the process (1.1) contains both finite and infinite variance models.

It is worth mentioning the relationship between Assumption 1.1 and the spatial median. Let us define an objective function

$$Q(v) := E(\|U(t) - v\| - \|U(t)\|). \qquad (1.3)$$

Since an inequality $|\|U(t) - v\| - \|U(t)\|| \leq \|v\|$ holds, the expectation (1.3) always exists. The minimizer of Q is called the spatial median of $U(t)$. It is also easy to show that Q is 1-Lipschtz continuous and convex. In addition, under Assumption 1.1, the dominated convergence theorem yields $\nabla Q(v) = 0_d$ if $v = 0_d$. From the elementary calculation and the properties of convex functions, we conclude that one of the minimizers of $Q(v)$ is 0_d. When $d \geq 2$, the uniqueness of the spatial median is shown by [19] under a milder assumption than Assumption 1.1(iii).[1]

Now, we state (without proof) a key lemma in the literature of the spatial median approach. A similar result as Lemma 1.1 below is found in [4].

Lemma 1.1 *Suppose that U is a d-dimensional random vector with a density function f satisfying Assumption 1.1 (ii). Then, $E(\|U + v\|^{-r}) \leq C_d(r, f)$ for any constant vector $v \in \mathbb{R}^d$ and $r \in (0, d)$, where*

$$C_d(r, f) := 1 + \frac{2\pi^{d-1} \sup_{x \in \mathbb{R}^d} f(x)}{d - r}.$$

The following result is a corollary of Lemma 1.1.

Corollary 1.1 *Suppose that U and X are independent d-dimensional random vectors, U has a density function f satisfying Assumption 1.1(ii), and $q : \mathbb{R}^d \to \mathbb{R}^+$ is a nonnegative and nonrandom function satisfying $E(q(X)) < \infty$. Then, we have $E(\|U + X\|^{-r} q(X)) \leq C_d(r, f) E(q(X))$ for any $r \in (0, d)$.*

Corollary 1.1 is an easy consequence of Lemma 1.1 and Fubini's theorem for nonnegative random variables; thus, the detail of the proof is omitted. Lemma 1.1 and Corollary 1.1 are frequently used for $(U, X) = (U(t), \Psi(\theta)\mathbb{X}_{t-1})$ in the proofs to bound the remainder terms in the stochastic expansions. It should be noted that $d \geq 2$ is essential in this paper, since Corollary 1.1 is used for $r > 1$ in the proofs of the main results. In addition, the matrices A_U and B_U, defined as (1.2), must be positive definite, since they appear in the asymptotic covariance matrix of the proposed estimator.

1.2.2 Self-weighted and Smoothed GEL Function

Motivated by the EL approach for quantile regression methods in the univariate case, one may consider the moment restriction of the form "$E(\text{Sign}(U(t; \theta)) \otimes \mathbb{X}_{t-1}) = 0_m$" and construct EL ratio statistics. However, the spatial sign function is not smooth at 0_d; besides, $X(t)$ is possibly an infinite variance process in our framework. Thus, we consider the *smoothed and self-weighted moment function*

$$g^*_{t,h}(\theta) := -\frac{U(t; \theta)}{\sqrt{\|U(t; \theta)\|^2 + h^2}} \otimes J(\mathbb{X}_{t-1}) \quad (t = p + 1, ..., n), \tag{1.4}$$

[1] The paper [19] assumed that the support of f_U is not concentrated on a line in $\mathbb{R}^d$.

where $h > 0$ is a smoothing parameter tending to zero as $n \to \infty$, $J : \mathbb{R}^{dp} \to \mathbb{R}^{dp+q}$ is defined as

$$J(x) := \begin{pmatrix} w(x)x \\ \varphi(x) \end{pmatrix},$$

with $w : \mathbb{R}^{dp} \to \mathbb{R}^{+}$ and $\varphi : \mathbb{R}^{dp} \to \mathbb{R}^{q}$ being user-specified functions. In particular, w is called the self-weight. The additional function φ expresses an over-identified moment condition, and if $q = 0$, then $g^{*}_{t,h}(\theta)$ corresponds to the just-identified (and smoothed) moment function. We set the following condition for w and φ.

Assumption 1.2 There exist some $\beta > 2$ and $r \in (0, 1)$ such that

$$E\left(\|J(\mathbb{X}_{t-1})\|^{\beta} + \left\{\|\mathbb{X}_{t-1}\| + \|\mathbb{X}_{t-1}\|^{1+r}\right\}\left\{\|J(\mathbb{X}_{t-1})\| + \|J(\mathbb{X}_{t-1})\|^{2}\right\}\right) < \infty.$$

Example 1.1 Assume that $J(x) = w(x)x$. If w has the bounded support (for example, $w(x) := \mathbb{I}(\|x\| \le c)$ with some $c \in \mathbb{R}^{+}$), then Assumption 1.2 is automatically satisfied. Moreover, it is easy to check that the following weight function fulfills Assumption 1.2;

$$w(\mathbb{X}_{t-1}) := \frac{1}{(1 + \sum_{s=1}^{p} s^{-a}\|X(t-s)\|)^{2}}, \tag{1.5}$$

where $a > 2$ is a constant. This weight function can be regarded as a truncated and multivariate version of [28].

By definition, we have $U(t;\theta_0) = U(t)$ and $U(t) \neq 0_d$ a.s. Then, for any $r \in (0, 1)$, a Taylor expansion yields

$$\begin{aligned}\left\| E\left(\frac{U(t;\theta_0)}{\sqrt{\|U(t;\theta_0)\|^{2} + h^{2}}}\right)\right\| &= \left\| E\left(\frac{U(t)}{\sqrt{\|U(t)\|^{2} + h^{2}}} - \frac{U(t)}{\|U(t)\|}\right)\right\| \\ &\le E\left(\left|\frac{\|U(t)\|}{\sqrt{\|U(t)\|^{2} + h^{2}}} - 1\right|^{r}\right) \\ &\le E\left(\|U(t)\|^{-2r}\right)h^{2r} \\ &\le C_d(2r, f_U)h^{2r}, \end{aligned} \tag{1.6}$$

where the last inequality in (1.6) follows from Lemma 1.1. Thus, under Assumption 1.2, the moment condition

$$E(g^{*}_{t,h}(\theta_0)) = o(1) \tag{1.7}$$

holds when $h \to 0$.

Remark 1.1 The form of our moment function (1.4) is reminiscent of the smoothed EL approach presented by [34]. In his paper, a linear regression model $Y_t = X_t^\top \beta_0 + U_t$ was considered, where Y_t and X_t are, respectively, a scalar response and a vector of regressors, and U_t is an unobserved error whose τth quantile ($0 < \tau < 1$) is zero. In this model, the moment restriction $E((\tau - \mathbb{I}(Y_t - X_t^\top \beta_0))) = 0$ is satisfied. For the estimation problem of the true coefficient vector β_0, [34] considered the EL estimator based on the smoothed moment function $Z_t(\beta) := -(\tau - I_h(X_t^\top \beta - Y_t))X_t$, where

$$I_h(u) := \int_{-\infty}^{u/h} K(v)dv$$

and K is a kernel function on $\mathbb{R}$. This function $I_h(u)$, when h is sufficiently small, is regarded as a smooth approximation of the indicator function $\mathbb{I}(u \geq 0)$. Now, consider a special case of $\tau = 1/2$. If we choose $K(v) := 1/(2(v^2 + 1)^{3/2})$, then

$$I_h(u) = \frac{1}{2}\left(\frac{u}{\sqrt{u^2 + h^2}} + 1\right).$$

Therefore, the moment function $Z_t(\beta)$ satisfies

$$Z_t(\beta) \propto -\frac{Y_t - X_t^\top \beta}{\sqrt{(Y_t - X_t^\top \beta)^2 + h^2}} X_t,$$

and this corresponds to an unweighted and scalar version of our moment function (1.4). In our framework, however, we restrict ourselves within the multivariate case, and K may be unbounded. Therefore, the extension is not straightforward and highlights some sharp contrasts between univariate and multivariate cases. Nonetheless, thanks to the smoothing, we can use the gradient of the objective function in numerical optimizations. This aspect is another advantage of the smoothed approach.

Remark 1.2 In this remark, we consider the just-identified case $J(x) = w(x)x$ to make the motivation of the moment function (1.4) clear. Define a function $l_h(u) := u/\sqrt{\|u\|^2 + h^2}$. Then, the moment function (1.4) is represented as $g^*_{t,h}(\theta) = -l_h(U(t;\theta)) \otimes w(\mathbb{X}_{t-1})\mathbb{X}_{t-1}$. By some calculation, it is shown that $l_h(u) = \nabla D_h(u)$, where $D_h(u) := \sqrt{\|u\|^2 + h^2} - h$. Therefore, the sum of the moment function (1.4) corresponds to the gradient vector of the objective function $W_h(\theta) := \sum_{t=p+1}^{n} w(\mathbb{X}_{t-1}) D_h(X(t) - \Psi(\theta)\mathbb{X}_{t-1})$. If $h = 0$ and $w(x) \equiv 1$, then the minimizer of $W_0(\theta)$ is a sort of the least distance estimator by [3] for VAR models. It should be noted that, for each $h > 0$, the function $D_h(u)$ is approximated by $\|u\|$ when $\|u\| \to \infty$. Therefore, the GEL estimator and test statistic based on (1.4) are expected to be robust to outliers.

The self-weighted GEL function for the moment restriction (1.7) is defined as

$$r^*_n(\theta) := \sup_{\lambda \in \hat{\Lambda}_n(\theta)} \hat{P}_n(\theta, \lambda),$$

where

$$\hat{P}_n(\theta, \lambda) := \sum_{t=1}^{n} \rho(\lambda^\top g^*_{t,h}(\theta)),$$
$$\Lambda_n(\theta) := \left\{\lambda \in \mathbb{R}^m : \lambda^\top g^*_{t,h}(\theta) \in \mathcal{V}_\rho \text{ for all } t = 1, ..., n\right\}$$

and $\rho : \mathcal{V}_\rho \to \mathbb{R}$ is a user specified function with domain $\mathcal{V}_\rho$ ($\subset \mathbb{R}$). By choosing ρ appropriately, the GEL statistic can represent famous statistics in econometrics. For example, the EL (see [27]), exponential tilting (ET) (see [10]), and continuous updating (CU) (see [8]) are special cases of the GEL with

(EL) $\rho(x) = \log(1 - x)$ and $\mathcal{V}_\rho = (-\infty, 1)$;
(ET) $\rho(x) = 1 - \exp(x)$ and $\mathcal{V}_\rho = \mathbb{R}$;
(CU) $\rho(x) = -(x + 2)x/2$ and $\mathcal{V}_\rho = \mathbb{R}$.

In general, we assume the following condition for ρ.

Assumption 1.3 A function ρ is concave and twice continuously differentiable on its domain $\mathcal{V}_\rho$, where $\mathcal{V}_\rho$ is an open interval of $\mathbb{R}$ and contains zero. Further, $\rho(0) = 0$ and $\dot{\rho}(0) = \ddot{\rho}(0) = -1$, where $\dot{\rho}$ and $\ddot{\rho}$ are the first and second derivatives of ρ, respectively.

1.3 Main Results

This section derives the asymptotic distributions of the GEL estimator and test statistic. We focus on the null hypothesis

$$H : R(\theta_0) = 0_s, \tag{1.8}$$

where $R : \mathbb{R}^m \to \mathbb{R}^s$ ($s \le m$) is a restriction function. For example, $R(\theta) := \Gamma\theta - \gamma$, where $\Gamma \in \mathbb{R}^{s\times m}$ and $\gamma \in \mathbb{R}^s$ are, respectively, the user-specified matrix and vector. By choosing an appropriate R, our framework represents various problems, such as model diagnostics and tests of causality. Define the GEL estimator $\hat{\theta}_n := \arg\min_{\theta\in\Theta} r^*_n(\theta)$ and the GEL test statistic

$$T_n := 2\left[\inf_{\theta\in\Theta_R} r^*_n(\theta) - \inf_{\theta\in\Theta} r^*_n(\theta)\right],$$

where $\Theta_R := \{\theta : \theta \in \Theta, R(\theta) = 0_s\}$.

Before presenting the main theorem of this paper, we introduce a supporting theorem. Let us define

$$\hat{g}_{n,h}(\theta) := \frac{1}{n} \sum_{t=p+1}^{n} g^*_{t,h}(\theta)$$

and $\bar{g}_h(\theta) := E(g^*_{t,h}(\theta))$.

Theorem 1.1 *Define $B_0(\delta) := \{\theta : \theta \in \Theta, \|\theta - \theta_0\| \leq \delta\}$. If $h \to 0$ as $n \to \infty$, and Assumptions 1.1 (ii) and 1.2 hold, then, for any sequence $\delta_n \to 0$ ($n \to \infty$),*

$$\sup_{\theta \in B_0(\delta_n)} \frac{\sqrt{n} \left\| \left\{ \hat{g}_{n,h}(\theta) - \bar{g}_h(\theta) \right\} - \left\{ \hat{g}_{n,h}(\theta_0) - \bar{g}_h(\theta_0) \right\} \right\|}{1 + \sqrt{n}\|\theta - \theta_0\|} \to 0 \quad (n \to \infty).$$

Here, we discuss the distinction between our approach and [22]. As the moment function (1.4) contains parameter θ in a complicated way, the uniform convergence results in [21] can not be applied directly. Therefore, the techniques in [22], based on the Taylor expansions and uniform results of [21], are not applicable to our framework. However, Theorem 1.1 corresponds to Assumption 2.2(d) of [29]. Thus, in the proof of our theorem below, we mainly follow a similar argument as in [29] with Theorem 1.1. To establish Theorem 1.1, we make use of the approximation

$$\|l_h(u - x) - l_h(u) + M_h(u)x\| \leq \sqrt{d+4} \frac{\|x\|^{1+\omega}}{\|u\|^{1+\omega}}$$

for any $u, x \in \mathbb{R}^d$, $x \neq u$, $u \neq 0_d$, and $\omega \in [0, 1]$, where

$$M_h(u) := \frac{1}{\sqrt{\|u\|^2 + h^2}} \left(I_d - \frac{uu^\top}{\|u\|^2 + h^2} \right). \tag{1.9}$$

The special case of this approximation ($h = 0$) was introduced in [3] and [26, Lemma 6.2]. This approximation plays a crucial role in the proofs of theorems in this paper.

To describe the limit distributions of the GEL estimators and test statistics, some important quantities need to be introduced. Let us define

$$\hat{G}_{n,h}(\theta) := \frac{1}{n} \sum_{t=p+1}^{n} \frac{\partial g^*_{t,h}(\theta)}{\partial \theta^\top}.$$

Through a lengthy but elementary calculation, we obtain a representation

$$\hat{G}_{n,h}(\theta) = \frac{1}{n} \sum_{t=p+1}^{n} \left[\mathbb{X}_{t-1}^\top \otimes M_h(U(t;\theta)) \otimes J(\mathbb{X}_{t-1}) \right].$$

Also, define $\bar{G}_h(\theta) := \partial \bar{g}_h(\theta)/\partial \theta^\top$. Note that, for each h,

$$\left\| \mathbb{X}_{t-1}^{\top} \otimes M_h(U(t;\theta)) \otimes J(\mathbb{X}_{t-1}) \right\| \le \|\mathbb{X}_{t-1}\| \|J(\mathbb{X}_{t-1})\| \frac{\sqrt{d+3}}{h} \tag{1.10}$$

and the right-hand side of (1.10) is integrable under Assumption 1.2. Then, Corollary 1.1 and the dominated convergence theorem yield

$$\bar{G}_h(\theta) = E\left(\mathbb{X}_{t-1}^{\top} \otimes M_h(U(t;\theta)) \otimes J(\mathbb{X}_{t-1})\right)$$

and

$$\begin{aligned}\bar{G}_h(\theta_0) &= E\left(\frac{1}{\sqrt{\|U(t)\|^2 + h^2}} \mathbb{X}_{t-1}^{\top} \otimes \left(I_d - \frac{U(t)U(t)^{\top}}{\|U(t)\|^2 + h^2}\right) \otimes J(\mathbb{X}_{t-1})\right) \\ &\to E\left(\mathbb{X}_{t-1}^{\top} \otimes A_U \otimes J(\mathbb{X}_{t-1})\right) \quad (h \to 0).\end{aligned}$$

Hereafter, denote $E(\mathbb{X}_{t-1}^{\top} \otimes A_U \otimes J(\mathbb{X}_{t-1}))$ by G_0.

We make the following assumptions.

Assumption 1.4 Let $h = h_n$. There exists some $\delta \in (0, 1)$ such that $nh^{4\delta} \to 0$ as $n \to \infty$.

Assumption 1.5

(i) The matrix $\Sigma_J := E(J(\mathbb{X}_{t-1})J(\mathbb{X}_{t-1})^{\top})$ exists and is nonsingular;
(ii) G_0 and $\partial R(\theta)/\partial\theta^{\top}|_{\theta=\theta_0}$ are full-rank;
(iii) $\bar{G}_0(\theta)$ is a continuous function of θ;
(iv) $\theta_0 \in \mathrm{Int}(\Theta)$.

Assumption 1.4 is required for a remainder term in the expansion of $\sqrt{n}\hat{g}_{n,h}(\theta_0)$ to be asymptotically negligible. Assumption 1.5 guarantees some regularity conditions for the asymptotic normality of fundamental quantities.

The following theorem gives the limit distributions of the proposed GEL estimator and test statistic.

Theorem 1.2 *Suppose that Assumptions 1.1–1.5 hold. Then, for $d \ge 2$,*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{d} N(0_{d^2 p}, (G_0^{\top}\Omega_0^{-1}G_0)^{-1}),$$

where $\Omega_0 := B_U \otimes \Sigma_J$. Further, under (1.8), $T_n \xrightarrow{d} \chi_s^2$ as $n \to \infty$, where χ_s^2 is the chi-squared random variable with s degrees of freedom.

By Theorem 1.2, the self-weighted GEL estimator has asymptotic normality, and the GEL test statistic has a pivotal limit distribution under fairly mild conditions for the moments and distribution of the error term. Because the rate of convergence and the limit distribution of T_n do not contain any nuisance parameters, such as the tail index of the error process, Theorem 1.2 provides a feasible testing procedure for the null hypothesis (1.8). That is, we reject the null hypothesis (1.8) whenever $T_n > q_{1-\alpha}$, where $q_{1-\alpha}$ is the $(1-\alpha)$th quantile of the χ_s^2 distribution. Throughout this paper, the test based on the test statistic T_n is referred to as the T_n-test.

Remark 1.3 It is naturally shown that the asymptotic behavior of $\hat{\theta}_n$ and T_n depends on the asymptotics of the quantity $\hat{g}_{n,h}(\theta_0)$. In the proof of the asymptotic normality of $\hat{g}_{n,h}(\theta_0)$, it is shown that $\sqrt{n}(\hat{g}_{n,h}(\theta_0) - \hat{g}_{n,0}(\theta_0)) = o_p(1)$. Note that

$$\sqrt{n}\hat{g}_{n,0}(\theta_0)) = \frac{1}{\sqrt{n}} \sum_{t=p+1}^{n} \mathrm{Sign}(U(t)) \otimes J(\mathbb{X}_{t-1}),$$

and then $\sqrt{n}\hat{g}_{n,0}(\theta_0)$ is a sum of a martingale difference array due to the spatial sign approach and self-weighting. Therefore, the central limit theorem in [16] can be used directly without the data-blocking technique by [9].

1.4 Finite Sample Performance

This section shows some simulation results to illustrate the finite sample performance of the proposed estimator and test statistic. Suppose that $X(1), ..., X(n)$ are an observed stretch from the bivariate VAR(2)-model

$$X(t) = \begin{bmatrix} 0.5 & a_1 \\ 0.3 & -0.5 \end{bmatrix} X(t-1) + \begin{bmatrix} 0.1 & a_2 \\ 0.5 & -0.2 \end{bmatrix} X(t-2) + U(t), \tag{1.11}$$

where a_1 and a_2 are some scalars. In this section, $\{U(t) : t \in \mathbb{Z}\}$ is assumed to be a sequence of i.i.d. random vectors, and $U(t) =_d BZ^{(j_1,j_2)}$, where B is a (2×2)-constant matrix and $Z^{(j_1,j_2)} = (Z_1^{(j_1)}, Z_2^{(j_2)})^\top$ is a vector of two independent random variables $Z_1^{(j_1)}$ and $Z_2^{(j_2)}$. We assume one of the following distributions for $Z_i^{(j_i)}$, according to a parameter j_i;

$$Z_i^{(j_i)} \sim \begin{cases} N(0,1) & (j_i = 1) \\ t_2 & (j_i = 2) \\ t_1 & (j_i = 3) \end{cases} \quad (i = 1, 2),$$

where t_k is the t-distribution with k degrees of freedom. In particular, $Z_i^{(3)}$ follows the Cauchy distribution. We also assume that B is one of the following matrices;

(B1) $B = I_2$;

(B2) $B = B_2 := \begin{bmatrix} 1/2 & -\sqrt{3}/2 \\ \sqrt{3}/2 & 1/2 \end{bmatrix}$.

For the case (B2), the components of $U(t)$ are not independent. Since the marginal medians of $Z^{(j_1,j_2)}$ are zero and B is a rotation matrix, $U(t)$ has a zero-spatial median in all cases. Note that $U(t)$ has infinite variance except for the case of $(j_1, j_2) = (1, 1)$. The function ρ is chosen as $\rho(v) = -(2 + v)v/2$ (CU), and the self-weight (1.5) is used with $a = 3$.

Table 1.1 The estimated MSEs (1.12) for $\hat{A}_1$ and $\hat{A}_2$ in the case of (B1)

	$n = 100$		$n = 500$		$n = 1000$	
(j_1, j_2)	MSE_1	MSE_2	MSE_1	MSE_2	MSE_1	MSE_2
(1, 1)	0.2407	0.2483	0.1033	0.1105	0.0740	0.0784
(1, 2)	0.2478	0.2519	0.1065	0.1067	0.0736	0.0745
(1, 3)	0.2885	0.2993	0.1191	0.1203	0.0843	0.0840
(2, 1)	0.2245	0.2239	0.0946	0.0947	0.0678	0.0674
(2, 2)	0.2152	0.2145	0.0896	0.0888	0.0616	0.0621
(2, 3)	0.2350	0.2247	0.0897	0.0859	0.0625	0.0595
(3, 1)	0.2437	0.2199	0.0969	0.0852	0.0659	0.0581
(3, 2)	0.2309	0.2148	0.0856	0.0786	0.0572	0.0548
(3, 3)	0.2225	0.1993	0.0798	0.0710	0.0541	0.0477

First, we check the accuracy of the CU estimator. We generate an observed stretch with length n from the model (1.11), and calculate the CU estimator $\hat{\theta}_n$. We estimate the true coefficient matrices A_1, A_2 by $\left[\hat{A}_1, \hat{A}_2\right] := \Psi(\hat{\theta}_n)$. By repeating the above-mentioned procedure 1000 times, we estimate the mean squared error (MSE) of the estimators by the quantity

$$\text{MSE}_j := \frac{1}{1000} \sum_{l=1}^{1000} \left\| \hat{A}_j^{(l)} - A_j \right\| \quad (j = 1, 2), \tag{1.12}$$

where $\hat{A}_j^{(l)}$ is the estimator for A_j in the lth iteration. The parameters are set as $a_1 = a_2 = 0$, $n = 100, 500, 1000$, and $h = n^{-1/2}$. The results are summarized in Tables 1.1 and 1.2.

Table 1.2 The estimated MSEs (1.12) for $\hat{A}_1$ and $\hat{A}_2$ in the case of (B2)

	$n = 100$		$n = 500$		$n = 1000$	
(j_1, j_2)	MSE_1	MSE_2	MSE_1	MSE_2	MSE_1	MSE_2
(1, 1)	0.2418	0.2539	0.1043	0.1096	0.0735	0.0786
(1, 2)	0.2368	0.2279	0.1005	0.0990	0.0702	0.0695
(1, 3)	0.2603	0.2458	0.1026	0.0944	0.0722	0.0653
(2, 1)	0.2482	0.2435	0.1025	0.1035	0.0730	0.0710
(2, 2)	0.2169	0.2159	0.0894	0.0909	0.0621	0.0628
(2, 3)	0.2276	0.2238	0.0893	0.0841	0.0601	0.0575
(3, 1)	0.2908	0.2482	0.1218	0.0984	0.0821	0.0659
(3, 2)	0.2415	0.2194	0.0923	0.0835	0.0639	0.0580
(3, 3)	0.2337	0.2151	0.0816	0.0769	0.0550	0.0521

When the components of $U(t)$ are independent (case (B1)), the estimated errors tend to be large when the first component of $U(t)$ follows a normal distribution (cases $(j_1, j_2) = (1, 1), (1, 2), (1, 3)$). Especially the estimator gives the smallest errors when $U(t)$ is a vector of independent Cauchy random variables (case $(j_1, j_2) = (3, 3)$) except for the case of $n = 100$, and this tendency is also observed in the case of $B = B_2$. When the components of $U(t)$ are not independent (case (B2)), the estimator tends to perform better when both $Z_1^{(j_1)}$ and $Z_2^{(j_2)}$ have infinite variance (cases $(j_1, j_2) = (2, 2), (2, 3), (3, 2), (3, 3)$), while the errors are large when $Z_1^{(j_1)}$ or/and $Z_2^{(j_2)}$ follows a normal distribution.

We also investigate the empirical type-I error rates and powers of the T_n-test. Let us set $s = 2$ and $R(\theta) = \Gamma\theta$, where

$$\Gamma = \begin{bmatrix} 0\,0\,1\,0\,0\,0\,0\,0 \\ 0\,0\,0\,0\,0\,0\,1\,0 \end{bmatrix}.$$

Then, the null hypothesis (1.8) is equivalent to $H : a_1 = 0$ and $a_2 = 0$. The parameters are chosen as $(a_1, a_2) = (0.0, 0.0), (0.1, 0.0), (0.0, 0.1)$, and $(0.1, 0.1)$. The nominal significance level of test is 5%, and the empirical rejection rates of the T_n-test are summarized in Tables 1.3 and 1.4. The upper left panel of each table shows the simulated type-I error rate, while the upper right, and lower left and right panels of each table show the empirical rejection rates under alternative hypotheses $(a_1, a_2) = (0.1, 0.0), (0.0, 0.1), (0.1, 0.1)$, respectively.

Regarding the type-I error, the proposed T_n-test is slightly oversized in a few cases, but the nominal significance level is well-approximated in almost all cases. For the case (B1), the powers of the test rapidly increase when $Z_2^{(j_2)}$ follows the Cauchy distribution. Even for the case (B2), a similar tendency is observed. Overall, the proposed method shows a better performance when the error term follows heavy-tailed distributions, and this implies the robustness of the proposed estimator and test statistic.

1.5 Proofs

1.5.1 Some Approximations

We introduce the following Lemma 1.2, where (i) was summarized in [26, Sect. 6.1.1], and originally based on [2, 3].

Lemma 1.2 *(i) If $u, x \in \mathbb{R}^d$, $u \neq 0_d$ and $x \neq u$, then*

$$\left\| \frac{u - x}{\|u - x\|} - \frac{u}{\|u\|} \right\| \leq \frac{2\|x\|}{\|u\|}. \tag{1.13}$$

Table 1.3 The empirical type-I error rates and powers of the T_n-test in the case of (B1)

$(a_1, a_2) = (0.0, 0.0)$				$(a_1, a_2) = (0.1, 0.0)$			
(j_1, j_2)	$n = 100$	$n = 500$	$n = 1000$	(j_1, j_2)	$n = 100$	$n = 500$	$n = 1000$
(1, 1)	0.040	0.056	0.051	(1, 1)	0.102	0.401	0.710
(1, 2)	0.037	0.050	0.043	(1, 2)	0.131	0.767	0.971
(1, 3)	0.032	0.040	0.046	(1, 3)	0.246	0.957	1.000
(2, 1)	0.031	0.039	0.052	(2, 1)	0.097	0.305	0.537
(2, 2)	0.047	0.049	0.050	(2, 2)	0.124	0.535	0.877
(2, 3)	0.031	0.040	0.055	(2, 3)	0.205	0.881	0.995
(3, 1)	0.050	0.062	0.047	(3, 1)	0.083	0.209	0.404
(3, 2)	0.051	0.031	0.048	(3, 2)	0.083	0.433	0.720
(3, 3)	0.066	0.041	0.049	(3, 3)	0.155	0.681	0.953
$(a_1, a_2) = (0.0, 0.1)$				$(a_1, a_2) = (0.1, 0.1)$			
(j_1, j_2)	$n = 100$	$n = 500$	$n = 1000$	(j_1, j_2)	$n = 100$	$n = 500$	$n = 1000$
(1, 1)	0.100	0.424	0.747	(1, 1)	0.105	0.522	0.854
(1, 2)	0.197	0.825	0.993	(1, 2)	0.206	0.886	0.997
(1, 3)	0.364	0.996	1.000	(1, 3)	0.363	0.996	1.000
(2, 1)	0.098	0.376	0.658	(2, 1)	0.098	0.481	0.786
(2, 2)	0.140	0.685	0.939	(2, 2)	0.174	0.719	0.969
(2, 3)	0.291	0.968	1.000	(2, 3)	0.294	0.951	1.000
(3, 1)	0.101	0.370	0.697	(3, 1)	0.109	0.473	0.779
(3, 2)	0.130	0.537	0.874	(3, 2)	0.134	0.622	0.925
(3, 3)	0.211	0.850	0.998	(3, 3)	0.204	0.872	0.994

In addition, if $d \geq 2$, then

$$\left\| \frac{u - x}{\|u - x\|} - \frac{u}{\|u\|} + M_0(u)x \right\| \leq \sqrt{d+3}\frac{\|x\|^{1+\omega}}{\|u\|^{1+\omega}} \tag{1.14}$$

for any $\omega \in [0, 1]$, where M_h is defined as (1.9).

(ii) If $u, x \in \mathbb{R}^d$ and $h > 0$, then $\|l_h(u - x) - l_h(u)\| \leq 2\|x\|/\|u\|$, where l_h is defined in Remark 1.2. In addition, if $d \geq 2$, then

$$\|l_h(u - x) - l_h(u) + M_h(u)x\| \leq \sqrt{d+4}\frac{\|x\|^{1+\omega}}{\|u\|^{1+\omega}}$$

for any $\omega \in [0, 1]$.

Proof We give a sketch of the proof. The inequality (1.13) can be obtained through some geometric considerations. To prove (1.14), define $S(u, x) := l_0(u - x) - l_0(u) + M_0(u)x$. By (1.13), it is easily shown that $\|S(u, x)\| \leq \sqrt{d+3}\|x\|/\|u\|$ for any $d \geq 2$, and then, the assertion is true when $\|x\| \geq \|u\|$. When $\|x\| < \|u\|$, we first

Table 1.4 The empirical type-I error rates and powers of the T_n-test in the case of (B2)

$(a_1, a_2) = (0.0, 0.0)$				$(a_1, a_2) = (0.1, 0.0)$			
(j_1, j_2)	$n = 100$	$n = 500$	$n = 1000$	(j_1, j_2)	$n = 100$	$n = 500$	$n = 1000$
(1, 1)	0.049	0.048	0.050	(1, 1)	0.089	0.432	0.695
(1, 2)	0.034	0.042	0.043	(1, 2)	0.073	0.342	0.599
(1, 3)	0.044	0.045	0.046	(1, 3)	0.069	0.251	0.507
(2, 1)	0.035	0.052	0.044	(2, 1)	0.119	0.619	0.907
(2, 2)	0.041	0.043	0.044	(2, 2)	0.113	0.568	0.909
(2, 3)	0.042	0.051	0.048	(2, 3)	0.103	0.453	0.798
(3, 1)	0.041	0.051	0.047	(3, 1)	0.150	0.752	0.973
(3, 2)	0.038	0.047	0.055	(3, 2)	0.140	0.775	0.981
(3, 3)	0.053	0.041	0.053	(3, 3)	0.121	0.669	0.961
$(a_1, a_2) = (0.0, 0.1)$				$(a_1, a_2) = (0.1, 0.1)$			
(j_1, j_2)	$n = 100$	$n = 500$	$n = 1000$	(j_1, j_2)	$n = 100$	$n = 500$	$n = 1000$
(1, 1)	0.104	0.435	0.746	(1, 1)	0.115	0.528	0.853
(1, 2)	0.106	0.494	0.847	(1, 2)	0.157	0.651	0.927
(1, 3)	0.145	0.665	0.945	(1, 3)	0.150	0.728	0.972
(2, 1)	0.106	0.587	0.893	(2, 1)	0.131	0.628	0.915
(2, 2)	0.142	0.659	0.936	(2, 2)	0.149	0.711	0.963
(2, 3)	0.140	0.691	0.958	(2, 3)	0.177	0.748	0.976
(3, 1)	0.111	0.694	0.969	(3, 1)	0.128	0.617	0.919
(3, 2)	0.145	0.757	0.969	(3, 2)	0.134	0.714	0.957
(3, 3)	0.162	0.803	0.987	(3, 3)	0.137	0.776	0.980

show that $\|S(u, x)\| \le \sqrt{d+3}\|x\|^2/\|u\|^2$ for $u = e_1$ by considering a polar coordinate system, where $e_1 = (1, 0_{d-1}^\top)^\top$. Second, for general u, note the relationships $\|u\|Re_1 = u$, where

$$R := \frac{(v + e_1)(v + e_1)^\top}{1 + v^\top e_1} - I_d \text{ and } v = \frac{u}{\|u\|}.$$

By a simple calculation, it is shown that $S(u, x) = RS(e_1, R^\top x/\|u\|)$ for general $u \in \mathbb{R}^d$. Therefore, the result in the case of $u = e_1$ leads to (1.14). Finally, (ii) is an easy consequence of (i).

1.5.2 Proofs of Theorems 1.1 and 1.2

This section provides some auxiliary lemmas and the proofs of Theorems 1.1 and 1.2. First, we prepare the following Lemmas 1.3 and 1.4, which show the convergence of the fundamental quantities appearing in the stochastic expansions of the GEL estimator and test statistic.

Lemma 1.3 *Under Assumptions 1.1, 1.2, 1.4, and 1.5,* $\Omega_0 = B_U \otimes \Sigma_J$ *is positive definite, and* $\sqrt{n}\hat{g}_{n,h}(\theta_0) \xrightarrow{d} N(0_m, \Omega_0)$ *as* $n \to \infty$.

Proof Under Assumptions 1.1(iii) and 1.5, B_U, as defined in (1.2), and Σ_J are positive definite. Thus, the first assertion follows. Define $m_{t,h} := -n^{-1/2} l_h(U(t))$ and decompose $\sqrt{n}\hat{g}_{n,h}(\theta_0)$ as

$$\sqrt{n}\hat{g}_{n,h}(\theta_0) = \sum_{t=p+1}^{n} \left(m_{t,h} - m_{t,0}\right) \otimes J(\mathbb{X}_{t-1}) + \sum_{t=p+1}^{n} m_{t,0} \otimes J(\mathbb{X}_{t-1}). \quad (1.15)$$

By noting that $m_{t,h}$ and $\mathbb{X}_{t-1}$ are independent, the first term in (1.15) is bounded as

$$E\left(\left\| \sum_{t=p+1}^{n} \left(m_{t,h} - m_{t,0}\right) \otimes J(\mathbb{X}_{t-1}) \right\|\right) \leq (n-p) E\left(\|J(\mathbb{X}_{t-1})\|\right) E\left(\|m_{t,h} - m_{t,0}\|\right). \quad (1.16)$$

By a similar argument as in (1.6), we have

$$E\left(\|m_{t,h} - m_{t,0}\|\right) = O(n^{-1/2}) E\left(\left\| \frac{U(t)}{\sqrt{\|U(t)\|^2 + h^2}} - \frac{U(t)}{\|U(t)\|} \right\|\right) = O\left(n^{-1/2} h^{2\delta}\right),$$

where δ is the same constant as in Assumption 1.4. Therefore, (1.16) is $O(\sqrt{nh^{4\delta}})$, which converges to zero under Assumption 1.4. Next, we show the asymptotic normality of the second term in (1.15). For any non-zero vector $c \in \mathbb{R}^m$, write $y_{t,n} := c^\top \{m_{t,0} \otimes J(\mathbb{X}_{t-1})\}$. Then, we have $\sqrt{n} c^\top \hat{g}_{n,h}(\theta_0) = \sum_{t=p+1}^{n} y_{t,n} + o_p(1)$. By Assumption 1.1(i), we have

$$E(y_{t,n} | \mathscr{F}_{t-1}) = -\frac{1}{\sqrt{n}} c^\top \left\{E(\mathrm{Sign}(U(t)) | \mathscr{F}_{t-1}) \otimes J(\mathbb{X}_{t-1})\right\} = 0 \text{ a.s.},$$

where $\mathscr{F}_{t-1}$ is the sigma-field generated by $\{U(s) : s \leq t-1\}$. That is, $\{y_{t,n} : 1 \leq t \leq n\}$ is a martingale difference array with respect to $\{\mathscr{F}_t : 1 \leq t \leq n\}$. Therefore, by checking the condition for the central limit theorem for a martingale difference array (c.f. [16]), we obtain the desired result.

Lemma 1.4 *Suppose that a sequence of random vectors* $\{\bar{\theta}_n : n \in \mathbb{N}\}$ *satisfies* $\|\bar{\theta}_n - \theta_0\| = o_p(1)$. *Then, under Assumptions 1.1 (ii), 1.2 and 1.4,*

$$\frac{1}{n} \sum_{t=1}^{n} g^*_{t,h}(\bar{\theta}_n) g^*_{t,h}(\bar{\theta}_n)^\top \xrightarrow{p} \Omega_0 \quad (n \to \infty).$$

Proof Denote $W_{n,h}(\theta) := n^{-1}\sum_{t=p+1}^{n} g_{t,h}^{*}(\theta)g_{t,h}^{*}(\theta)^{\top}$ and

$$S_h(\theta) := E(W_{n,h}(\theta)) = (1+o(1))E\left(L_h(U(t;\theta)) \otimes (J(\mathbb{X}_{t-1})J(\mathbb{X}_{t-1})^{\top})\right),$$

where $L_h(u) = uu^{\top}/(\|u\|^2 + h^2)$. Note that

$$S_0(\theta_0) = (1+o(1))E\left(\frac{U(t)U(t)^{\top}}{\|U(t)\|^2} \otimes (J(\mathbb{X}_{t-1})J(\mathbb{X}_{t-1})^{\top})\right) = (1+o(1))B_U \otimes \Sigma_J$$

and

$$\begin{aligned}&\|W_{n,h}(\bar{\theta}_n) - S_0(\theta_0)\| \\ &\leq \|W_{n,h}(\bar{\theta}_n) - W_{n,h}(\theta_0)\| + \|W_{n,h}(\theta_0) - W_{n,0}(\theta_0)\| + \|W_{n,0}(\theta_0) - S_0(\theta_0)\|.\end{aligned} \tag{1.17}$$

We evaluate the terms in (1.17), as follows.

For the first term of (1.17), we can rewrite $W_{n,h}(\theta)$ as

$$W_{n,h}(\theta) = \frac{1}{n}\sum_{t=p+1}^{n} L_h(U(t) - \Psi(\theta-\theta_0)\mathbb{X}_{t-1}) \otimes (J(\mathbb{X}_{t-1})J(\mathbb{X}_{t-1})^{\top}).$$

By Lemma 1.2, we have

$$\|L_h(u-x) - L_h(u)\| \leq 2\left\|\frac{u-x}{\sqrt{\|u-x\|^2+h^2}} - \frac{u}{\sqrt{\|u\|^2+h^2}}\right\| \leq 4\frac{\|x\|}{\|u\|}. \tag{1.18}$$

By (1.18), Corollary 1.1, and Assumption 1.2, we have

$$\begin{aligned}&\left\|W_{n,h}(\bar{\theta}_n) - W_{n,h}(\theta_0)\right\| \\ &\leq \frac{1}{n}\sum_{t=p+1}^{n}\left\|L_h(U(t) - \Psi(\bar{\theta}_n-\theta_0)\mathbb{X}_{t-1}) - L_h(U(t))\right\| \|J(\mathbb{X}_{t-1})\|^2 \\ &\leq \|\bar{\theta}_n - \theta_0\|\frac{4}{n}\sum_{t=p+1}^{n}\frac{\|\mathbb{X}_{t-1}\|\|J(\mathbb{X}_{t-1})\|^2}{\|U(t)\|} \xrightarrow{p} 0 \quad (n\to\infty).\end{aligned}$$

For the second term of (1.17), note that, for any $u \neq 0_d$, $\|L_h(u) - L_0(u)\| \leq 1$. In addition, a Taylor expansion with respect to h yields

$$L_h(u) = L_0(u) - \frac{2h^2uu^{\top}}{(\|u\|^2+\varepsilon h^2)^2}$$

for some $\varepsilon \in (0, 1)$. Hence,

$$\|L_h(u) - L_0(u)\| \leq \|L_h(u) - L_0(u)\|^{1/2} = \frac{2h\|u\|}{\|u\|^2 + \varepsilon h^2} \leq 2h\|u\|^{-1}.$$

Then, we have

$$\begin{aligned} E(\|W_{n,h}(\theta_0) - W_{n,h}(\theta_0)\|) &\leq (1 + o(1))E(\|L_h(U(t)) - L_0(U(t))\|\|J(\mathbb{X}_{t-1})\|^2) \\ &\leq 2hC_d(1, f_U)E(\|J(\mathbb{X}_{t-1})\|^2)(1 + o(1)) \to 0. \end{aligned}$$

This implies the convergence in probability of the second term of (1.17).

The convergence of the third term of (1.17) is guaranteed by the law of large numbers for ergodic processes.

***Proof of Theorem* 1.1** Recall that $\hat{g}_{n,h}(\theta)$ is represented as

$$\hat{g}_{n,h}(\theta) = -\frac{1}{n} \sum_{t=p+1}^{n} l_h(U(t) - \Psi(\theta - \theta_0)\mathbb{X}_{t-1}) \otimes J(\mathbb{X}_{t-1}).$$

Also define

$$Q_{t,h}(\theta) := M_h(U(t))\Psi(\theta - \theta_0)\mathbb{X}_{t-1} \text{ and } \hat{Q}_{n,h}(\theta) := \frac{1}{n} \sum_{t=p+1}^{n} Q_{t,h}(\theta) \otimes J(\mathbb{X}_{t-1}),$$

where M_h is defined in (1.9). Now, consider a decomposition

$$\begin{aligned} \sqrt{n}\left\|\hat{g}_{n,h}(\theta) - \hat{g}_{n,h}(\theta_0) - \bar{g}_h(\theta) + \bar{g}_h(\theta_0)\right\| &\leq \sqrt{n}\left\|\hat{g}_{n,h}(\theta) - \hat{g}_{n,h}(\theta_0) - \hat{Q}_{n,h}(\theta)\right\| \\ &\quad + \sqrt{n}\left\|\bar{g}_h(\theta) - \bar{g}_h(\theta_0) - E(\hat{Q}_{n,h}(\theta))\right\| \\ &\quad + \left\|\hat{Q}_{n,h}(\theta) - E(\hat{Q}_{n,h}(\theta))\right\| \\ &= A_1 + A_2 + A_3 \quad \text{(say)}. \end{aligned}$$

For the term A_1, Lemma 1.2 yields

$$\left\|l_h(U(t) - \Psi(\theta - \theta_0)\mathbb{X}_{t-1}) - l_h(U(t)) + Q_{t,h}(\theta)\right\| \leq \sqrt{d+4}\frac{\tau_0(\theta)^{1+r}\|\mathbb{X}_{t-1}\|^{1+r}}{\|U(t)\|^{1+r}}$$

for any $r \in [0, 1]$, where $\tau_0(\theta) := \|\theta - \theta_0\|$. Therefore,

$$\begin{aligned} A_1 &\leq \sqrt{n}\tau_0(\theta)^{1+r}\frac{\sqrt{d+4}}{n} \sum_{t=p+1}^{n} \frac{\|\mathbb{X}_{t-1}\|^{1+r}\|J(\mathbb{X}_{t-1})\|}{\|U(t)\|^{1+r}} \\ &= D_n\sqrt{n}\tau_0(\theta)^{1+r} \quad \text{(say)}, \end{aligned} \tag{1.19}$$

where $\{D_n : n \in \mathbb{N}\}$ is $O_p(1)$ by Corollary 1.1 and Assumption 1.2, and it is independent of θ.

For the term A_2, a similar argument yields

$$A_2 \leq (1 + o(1))\sqrt{n}\tau_0(\theta)^{1+r}\sqrt{d+4}E\left(\frac{\|\mathbb{X}_{t-1}\|^{1+r}\|J(\mathbb{X}_{t-1})\|}{\|U(t)\|^{1+r}}\right)$$
$$\leq C_0\sqrt{n}\tau_0(\theta)^{1+r}, \tag{1.20}$$

where C_0 is a constant that is independent of θ and n.

We further decompose A_3 as

$$A_3 \leq \left\|\hat{Q}_{n,h}(\theta) - \hat{Q}_{n,0}(\theta)\right\| + \left\|\hat{Q}_{n,0}(\theta) - E(\hat{Q}_{n,0}(\theta))\right\| + \left\|E(\hat{Q}_{n,h}(\theta)) - E(\hat{Q}_{n,0}(\theta))\right\|$$
$$= A_{3,1} + A_{3,2} + A_{3,3} \quad \text{(say)}.$$

For the term $A_{3,1}$,

$$E(A_{3,1}) \leq (1 + o(1))\tau_0(\theta)E\left(\|\mathbb{X}_{t-1}\|\|J(\mathbb{X}_{t-1})\|\ \|M_h(U(t)) - M_0(U(t))\|\right).$$

Note that $M_h(U(t)) - M_0(U(t)) \xrightarrow{p} O_{d\times d}$ as $h \to 0$ and $\|M_h(U(t)) - M_0(U(t))\|$ is bounded by $4/\|U(t)\|$, an integrable random variable. Therefore, the dominated convergence theorem yields $E(\|M_h(U(t)) - M_0(U(t))\|) \to 0$, hence, we have

$$A_{3,1} = o_p(1)\tau_0(\theta),$$

where the $o_p(1)$-term is uniform in θ. On the other hand, by lengthy but elementary argument,

$$Q_{t,h}(\theta) \otimes J(\mathbb{X}_{t-1}) = \left[\mathbb{X}_{t-1}^\top \otimes M_h(U(t)) \otimes J(\mathbb{X}_{t-1})\right](\theta - \theta_0).$$

Therefore,

$$A_{3,2}$$
$$\leq \left\|\frac{1}{n}\sum_{t=p+1}^{n}\left(\mathbb{X}_{t-1}^\top \otimes M_h(U(t)) \otimes J(\mathbb{X}_{t-1}) - E(\mathbb{X}_{t-1}^\top \otimes M_h(U(t)) \otimes J(\mathbb{X}_{t-1}))\right)\right\|\tau_0(\theta), \tag{1.21}$$

and the summation in (1.21) is $o_p(1)$ uniformly in θ by the ergodicity of the summands. The term $A_{3,3}$ is also evaluated as $o(1)\tau_0(\theta)$, where $o(1)$-term is uniform in θ, by the same reason for the term $A_{3,1}$. Thus,

$$A_3 = o_p(1)\tau_0(\theta), \tag{1.22}$$

where the $o_p(1)$-term is independent of θ.

By (1.19), (1.20) and (1.22), we have

$$\frac{\sqrt{n}\left\|\left\{\hat{g}_{n,h}(\theta)-\bar{g}_h(\theta)\right\}-\left\{\hat{g}_{n,h}(\theta_0)-\bar{g}_h(\theta_0)\right\}\right\|}{1+\sqrt{n}\|\theta-\theta_0\|}$$
$$\leq \frac{\sqrt{n}\tau_0(\theta)}{1+\sqrt{n}\tau_0(\theta)}\left[(D_n+C_0)\tau_0(\theta)^r+o_p(1)\right].$$

As $x/(1+x)<1$ for any $x \geq 0$, we have

$$\sup_{\theta\in B_0(\delta_n)} \frac{\sqrt{n}\tau_0(\theta)}{1+\sqrt{n}\tau_0(\theta)}\left[(D_n+C_0)\tau_0(\theta)^r+o_p(1)\right] \leq (D_n+C_0)\delta_n^r+o_p(1)\to 0$$

for any $\delta_n \to 0$. Thus, we obtain the desired result. □

***Proof of Theorem* 1.2** Because the proof of Theorem 1.2 is rather complicated, we give a sketch of the proof. The proof is mainly based on [22, 29, 30].

Step 1. By a similar argument as in [30] and by Theorem 1.1, we show the consistency of $\hat{\theta}_n$ as follows. We derive the stochastic expansion

$$\hat{P}_n(\theta,\lambda) = -n\lambda^\top \hat{g}_{n,h}(\theta) - \frac{n}{2}\lambda^\top \bar{\Omega}_{n,h}(\theta,\varepsilon_n\lambda)\lambda \tag{1.23}$$

for any θ and λ, where

$$\bar{\Omega}_{n,h}(\theta,\varepsilon_n\lambda) := -\frac{1}{n}\sum_{t=p+1}^{n}\ddot{\rho}(\varepsilon_n\lambda^\top g^*_{t,h}(\theta))g^*_{t,h}(\theta)g^*_{t,h}(\theta)^\top,$$

and $\{\varepsilon_n : n\in\mathbb{N}\}$ is a sequence of real numbers in $[0,1]$, which may depend on λ and n. By Lemmas 1.3, 1.4, and a similar argument as in [22, Proof of Lemma A2], we obtain $r_n^*(\theta_0)=O_p(1)$. However, by noting the stochastic expansion (1.23), we have a lower bound $r_n^*(\theta_n(u)) \geq D_n' n^{1-2\xi}\|u\|$, for $\theta_n(u) := \theta_0 + n^{-\xi}u$ with any $u\in\mathbb{R}^m$, $\|u\|\neq 0$, where $\xi\in(1/\beta,1/2)$ is a constant and $\{D_n' : n\in\mathbb{N}\}$ is a sequence of random variables that converges to a positive constant in probability. Now, fix any $\varepsilon>0$. If we define $\hat{u}_n := n^\xi(\hat{\theta}_n-\theta_0)$, then $\hat{\theta}_n=\theta_n(\hat{u}_n)$ and $r_n^*(\hat{\theta}_n)\geq D_n' n^{1-2\xi}\|\hat{u}_n\|$. By definition, $r_n^*(\hat{\theta}_n)\leq r_n^*(\theta_0)$. Then, we have $r_n^*(\theta_0)\geq D_n' n^{1-2\xi}\|\hat{u}_n\|$. Noting that $\|\hat{\theta}_n-\theta_0\| > n^{-\xi}\varepsilon$ implies $\|\hat{u}_n\|>\varepsilon$, we have

$$P(n^\xi\|\hat{\theta}_n-\theta_0\|>\varepsilon) \leq P(r_n^*(\theta_0)\geq D_n' n^{1-2\xi}\varepsilon)\to 0$$

for each $\varepsilon>0$. This implies $\|\hat{\theta}_n-\theta_0\| = o_p(n^{-\xi})$.

Step 2. Define

$$\hat{L}_n(\theta,\lambda) := -n\lambda^\top G_0(\theta-\theta_0) - n\lambda^\top \hat{g}_{n,h}(\theta_0) - \frac{n}{2}\lambda^\top \Omega_0 \lambda,$$
$$\tilde{\theta}_n := \arg\min_{\theta\in\Theta} \hat{L}_n(\theta, \tilde{\lambda}_n(\theta)),$$
$$\tilde{\lambda}_n(\theta) := \arg\max_{\lambda\in\mathbb{R}^{d^2 p}} \hat{L}_n(\theta,\lambda) \text{ and } \tilde{\lambda}_n := \tilde{\lambda}_n(\tilde{\theta}).$$

Then, it can be shown that $\hat{P}_n(\hat{\theta}_n, \hat{\lambda}_n) = \hat{L}_n(\tilde{\theta}_n, \tilde{\lambda}_n) + o_p(1)$, $\sqrt{n}(\hat{\theta}_n - \tilde{\theta}_n) = o_p(1)$ and $\sqrt{n}(\hat{\lambda}_n - \tilde{\lambda}_n) = o_p(1)$.

Step 3. We consider the asymptotics of $\tilde{\theta}_n$ and $\tilde{\lambda}_n$. The first-order conditions for $\tilde{\theta}_n$ and $\tilde{\lambda}_n$ are given by

$$-G_0^\top \tilde{\lambda}_n = 0_{d^2 p} \quad \text{and} \quad -G_0(\tilde{\theta}_n - \theta_0) - \hat{g}_{n,h}(\theta_0) - \Omega_0\tilde{\lambda}_n = 0_{d^2 p}. \tag{1.24}$$

The equations in (1.24) are stacked as

$$\begin{bmatrix} O_{m\times m} & G_0^\top \\ G_0 & \Omega_0 \end{bmatrix} \begin{pmatrix} \tilde{\theta}_n - \theta_0 \\ \tilde{\lambda}_n \end{pmatrix} = \begin{pmatrix} 0_{d^2 p} \\ -\hat{g}_{n,h}(\theta_0) \end{pmatrix},$$

and it is easy to see that

$$\begin{pmatrix} O_{m\times m} & G^\top \\ G & \Omega \end{pmatrix}^{-1} = \begin{pmatrix} -\Xi & \Xi G_0^\top \Omega_0^{-1} \\ \Omega_0^{-1} G_0 \Xi & \Omega_0^{-1} - \Omega_0^{-1} G_0 \Xi G_0^\top \Omega_0^{-1} \end{pmatrix},$$

where $\Xi := (G_0^\top \Omega_0^{-1} G_0)^{-1}$. Together with the result in Step 2 and Lemma 1.3, we obtain the limiting distribution of $\sqrt{n}(\hat{\theta}_n - \theta_0)$.

Step 4. Based on Steps 2 and 3, we obtain the stochastic expansion

$$T_n = n\hat{g}_{n,h}(\theta_0)^\top (P_R - P)\hat{g}_{n,h}(\theta_0) + o_p(1) = N^\top \Lambda N + o_p(1),$$

where N is a standard normal random vector,

$$P := \Omega_0^{-1} - \Omega_0^{-1} G_0 \Xi G_0^\top \Omega_0^{-1},$$
$$P_R := \Omega_0^{-1} - \Omega_0^{-1} G_0 (\Xi - \Xi \dot{R}^\top (\dot{R}\Xi\dot{R}^\top)^{-1} \dot{R}\Xi) G_0^\top \Omega_0^{-1},$$
$$\dot{R} := \partial R(\theta_0)/\partial\theta^\top \in \mathbb{R}^{s\times m} \text{ and}$$
$$\Lambda := \Omega_0^{-1/2} G_0 \Xi \dot{R}^\top (\dot{R}\Xi\dot{R}^\top)^{-1} \dot{R}\Xi G_0^\top \Omega_0^{-1/2}.$$

Noting that Λ is idempotent with rank s, we obtain the conclusion of the theorem by [31, (3b.4)]. □

Acknowledgements The author would like to express his deepest appreciation to Professor Masanobu Taniguchi for his successive guidance and wholehearted encouragement since the author was an undergraduate student. The author also would like to thank Dr. Junichi Hirukawa, Dr. Kou Fujimori, and Dr. Yuichi Goto for their comments on an earlier draft. The author also thank an anonymous referee for comments and suggestions on the presentation of the draft. The author would like to thank Editage (www.editage.com) for English language editing. This work was supported by JSPS KAKENHI Grant-in-Aid for Early-Career Scientists Grant Number 20K13467.

References

1. AKASHI, F. (2017). Self-weighted generalized empirical likelihood methods for hypothesis testing in infinite variance ARMA models. *Statistical Inference for Stochastic Processes* **20** 291–313.
2. ARCONES, M. A. (1998). Asymptotic theory for M-estimators over a convex kernel. *Econometric Theory* **14** 387–422.
3. BAI, Z., CHEN, X., MIAO, B. AND RADHAKRISHNA RAO, C. (1990). Asymptotic theory of least distances estimate in multivariate linear models. *Statistics* **21** 503–519.
4. BOSE, A. AND CHAUDHURI, P. (1993). On the dispersion of multivariate median. *Annals of the Institute of Statistical Mathematics* **45** 541–550.
5. DAVIS, R. A. AND RESNICK, S. I. (1985). Limit theory for moving averages of random variables with regularly varying tail probabilities. *The Annals of Probability* **13** 179–195.
6. DAVIS, R. A. AND RESNICK, S. I. (1985). More limit theory for the sample correlation function of moving averages. *Stochastic Processes and their Applications* **20** 257–279.
7. DAVIS, R. A. AND RESNICK, S. I. (1986). Limit theory for the sample covariance and correlation functions of moving averages. *The Annals of Statistics* **14** 533–558.
8. HANSEN, L. P., HEATON, J. AND YARON, A. (1996). Finite-sample properties of some alternative GMM estimators. *Journal of Business & Economic Statistics* **14** 262–280.
9. KITAMURA, Y. (1997). Empirical likelihood methods with weakly dependent processes. *The Annals of Statistics* **25** 2084–2102.
10. KITAMURA, Y. AND STUTZER, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica* **65** 861–874.
11. KLŬPPELBERG, C. AND MIKOSCH, T. (1996). The integrated periodogram for stable processes. *The Annals of Statistics* **24** 1855–1879.
12. LI, J., LIANG, W. AND HE, S. (2011). Empirical likelihood for LAD estimators in infinite variance ARMA models. *Statistics & Probability Letters* **81** 212–219.
13. LI, J., LIANG, W. AND HE, S. (2012). Empirical likelihood for AR-ARCH models based on LAD estimation. *Acta Mathematicae Applicatae Sinica, English Series* **28** 371–382.
14. LING, S. (2005). Self-weighted least absolute deviation estimation for infinite variance autoregressive models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67** 381–393.
15. LING, S. (2007). Self-weighted and local quasi-maximum likelihood estimators for ARMA–GARCH/IGARCH models. *Journal of Econometrics* **140** 849–873.
16. MCLEISH, D. L. (1974). Dependent central limit theorems and invariance principles. *The Annals of Probability* **2** 620 – 628.
17. MIKOSCH, T., GADRICH, T., KLUPPELBERG, C. AND ADLER, R. J. (1995). Parameter estimation for ARMA models with infinite variance innovations. *The Annals of Statistics* **23** 305–326.
18. MIKOSCH, T., RESNICK, S. AND SAMORODNITSKY, G. (2000). The maximum of the periodogram for a heavy-tailed sequence. *The Annals of Probability* **28** 885–908.
19. MILASEVIC, P. AND DUCHARME, G. R. (1987). Uniqueness of the spatial median. *The Annals of Statistics* **15** 1332 – 1333.

20. Monti, A. C. (1997). Empirical likelihood confidence regions in time series models. *Biometrika* **84** 395–405.
21. Newey, W. K. (1991). Uniform convergence in probability and stochastic equicontinuity. *Econometrica* **59** 1161–1167.
22. Newey, W. K. and Smith, R. J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* **72** 219–255.
23. Nordman, D. J. and Lahiri, S. N. (2006). A frequency domain empirical likelihood for short-and long-range dependence. *The Annals of Statistics* **34** 3019–3050.
24. Nordman, D. J. and Lahiri, S. N. (2014). A review of empirical likelihood methods for time series. *Journal of Statistical Planning and Inference* **155** 1–18.
25. Ogata, H. and Taniguchi, M. (2010). An empirical likelihood approach for non-Gaussian vector stationary processes and its application to minimum contrast estimation. *Australian & New Zealand Journal of Statistics* **52** 451–468.
26. Oja, H. (2010). *Multivariate Nonparametric Methods with R: An Approach Based on Spatial Signs and Ranks*. Springer Science & Business Media.
27. Owen, A. B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249.
28. Pan, J., Wang, H. and Yao, Q. (2007). Weighted least absolute deviations estimation for ARMA models with infinite variance. *Econometric Theory* **23** 852–879.
29. Parente, P. M. and Smith, R. J. (2011). GEL methods for nonsmooth moment indicators. *Econometric Theory* **27** 74–113.
30. Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *The Annals of Statistics* **22** 300 – 325.
31. Rao, C. R. (1973). *Linear Statistical Inference and its Applications*, vol 2. Wiley New York.
32. Samorodnitsky, G. and Taqqu, M. (1994). *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance*. CRC Press.
33. Small, C. (1990). A survey of multidimensional medians. *International Statistical Review* **58** 263–277.
34. Whang, Y.- J. (2006). Smoothed empirical likelihood methods for quantile regression models. *Econometric Theory* **22** 173–205.
35. Zhang, R. and Ling, S. (2015). Asymptotic inference for AR models with heavy-tailed G-GARCH noises. *Econometric Theory* **31** 880–890.
36. Zhu, K. and Ling, S. (2011). Global self-weighted and local quasi-maximum exponential likelihood estimators for ARMA–GARCH/IGARCH models. *The Annals of Statistics* **39** 2131–2163.
37. Zhu, K. and Ling, S. (2015). LADE-based inference for ARMA models with unspecified and heavy-tailed heteroscedastic noises. *Journal of the American Statistical Association* **110** 784–794.

Chapter 2
Excess Mean of Power Estimator of Extreme Value Index

Ngai Hang Chan, Yuxin Li, and Tony Sit

Abstract We propose a new type of extreme value index (EVI) estimator, namely, excess mean of power (EMP) estimator, which can be regarded as an average of the existing mean of order p (MOP) estimators over different thresholds. The asymptotic normalities of the MOP and EMP estimators for dependent observations are established under some mild conditions. We also develop consistent estimators for the asymptotic variances of the MOP and EMP estimators. Furthermore, the asymptotic normality of the extreme quantile estimator is established for dependent observations from which confidence intervals for the extreme quantile can be constructed. The proposed EMP estimator not only attains the best efficiency among typical EVI estimators under the optimal threshold, but is also more robust with respect to the choice of threshold.

2.1 Introduction

In various disciplines, such as insurance, finance and hydrology, we are concerned about the tail behavior of the underlying distribution for the interest of better decisions. For example, a substantial drawdown in returns poses threats to the stability of financial institutions. Extremely high water levels are dangerous and may incur tremendous losses. Therefore, it is of great significance to investigate the statistical results of such extreme events.

Extreme value theory (EVT) studies extreme events including the extreme quantile of a distribution and the probability of specific rare events, to name but a few. One essential part in extreme value theory is the statistical inference of extreme value

N. H. Chan (✉)
City University of Hong Kong, Tat Chee Ave, Kowloon Tong, Hong Kong SAR, China
e-mail: nhchan@cityu.edu.hk

Y. Li · T. Sit
The Chinese University of Hong Kong, Shatin, Hong Kong SAR, China
e-mail: 1155066153@link.cuhk.edu.hk

T. Sit
e-mail: tonysit@sta.cuhk.edu.hk

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_2

index (EVI) γ. Mathematically, a distribution F belongs to the domain of attraction with γ, if there is a positive function $a(t)$ such that for $x > 0$,

$$\lim_{t\to\infty} \frac{U(tx) - U(t)}{a(t)} = \begin{cases} \frac{x^\gamma - 1}{\gamma}, & \gamma > 0, \\ \log x & \gamma = 0, \end{cases} \tag{2.1}$$

where $U(x) := F^{-1}(1 - x^{-1})$. Throughout this paper, we focus on heavy-tailed distributions, where $\gamma > 0$; in this case, (2.1) is equivalent to $U(x) \in RV_\gamma$, where the symbol RV_γ denotes the class of regularly varying functions at infinity, with an index of regular variation equals $\gamma \in \mathbb{R}$, i.e., positive measurable function $U(\cdot)$ such that for all $x > 0$, $\lim_{t\to\infty} U(tx)/U(t) = x^\gamma$ (see, for example, Bingham et al. [4]).

Instead of the conventional Pareto quantile plots, Beirlant et al. [2] proposed generalized quantile plots. To begin with, we consider the mean excess function $e_U = E(X - U(t) \mid X > U(t))$, which is regularly varying with index γ when $\gamma < 1$. However, if $\gamma \geq 1$, the conditional expectation does not exist. To remedy this limitation, Beirlant et al. [2] replaced the empirical mean excess value evaluated at $t = nj^{-1}$ $E_{j,n} := j^{-1} \sum_{i=1}^{j} \left(X_{n-i+1,n} - X_{n-j,n}\right)$ by approximation $UH_{j,n} := X_{n-j,n} \left(j^{-1} \sum_{i=1}^{j} \log(X_{n-i+1,n}) - \log(X_{n-j,n})\right)$. Correspondingly, Beirlant et al. [2] defined $UH(x) := U(x) \int_1^\infty \{\log(U(wx)) - \log(U(x))\}\, w^{-2}\, dw$, which is also regularly varying with index γ with $UH_{j,n}$ as the sample analog of $UH(x)$ at $x = nj^{-1}$. One may estimate γ based on the slope of the generalized quantile plot, which is defined as $\left(\log\left((n+1)/j\right),\ \log(UH_{j,n})\right)$, for $j = 1, 2, \ldots, k_n$. Echoing the relationship between Hill estimator and the slope of Pareto Quantile Plot, Beirlant et al. [2] proposed a generalized Hill (GH) estimator, given by

$$\hat{\gamma}^{GH}(k_n) := \frac{1}{k_n} \sum_{j=1}^{k_n} \left\{\log(UH_{j,n}) - \log(UH_{k_n+1,n})\right\}.$$

In this paper, we consider an alternative approach to avoid the problem when $\gamma \geq 1$. Define

$$e_p(t) := E\left(X^p \mid X > U(t)\right),$$

where p is a tuning parameter. The existence of $e_p(t)$ can be guaranteed provided that $p\gamma < 1$; see Theorem 2.7 in Sect. 2.6.1. With

$$e_{j,p} = \frac{1}{j} \sum_{i=1}^{j} X_{n-i+1,n}^p,$$

as the sample analog of $e_p(t)$ valued at $t = (n+1)/j$, one can now define the excess mean of order-p Plot:

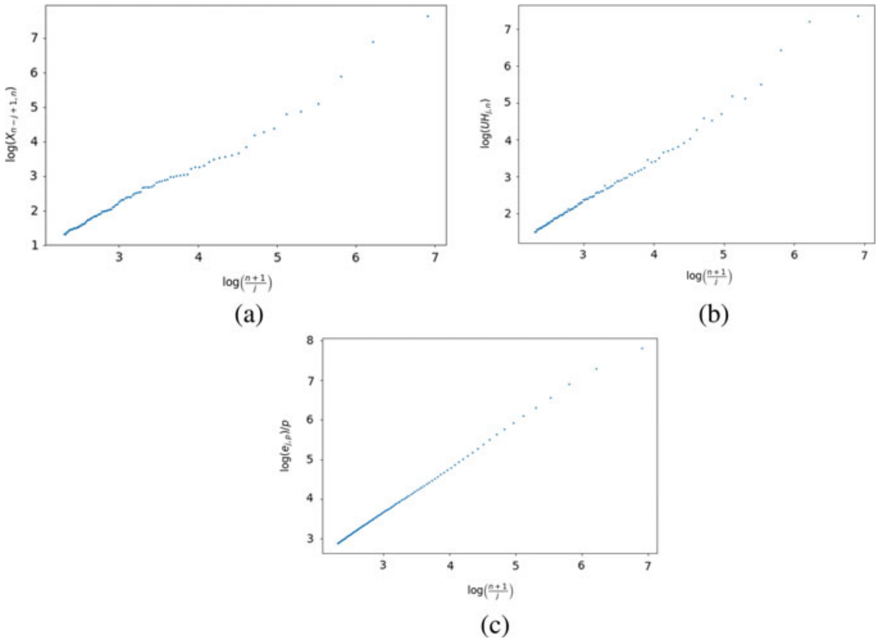


Fig. 2.1 Pareto quantile plot (**a**), generalized quantile plot (**b**) and excess mean of order-p plot (**c**) for $T(1)$: $n = 1000,\ k_n = 100,\ \gamma = 1,\ p = 0.4$

$$\left(\log\left(\frac{n+1}{j}\right),\ \frac{\log(e_{j,p})}{p}\right), \tag{2.2}$$

for $j = 1, 2, \ldots, k_n$. It is clear that the slope of the proposed plot can be used as an estimate of γ as $n \to \infty$, $k_n \to \infty$, and $k_n/n \to 0$. To illustrate the effectiveness of the excess mean of order-p Plot, we compare it with Pareto quantile plot and generalized quantile plot in Fig. 2.1, where we use $T(1)$ as the underlying distribution and take tuning parameter $p = 0.4$, where $T(\nu)$ denotes Student's t distribution with ν degrees of freedom.

As seen in Fig. 2.1, one advantage of the excess mean of order-p Plot is its relative smoothness and robustness with respect to k_n, compared with the Pareto quantile plot as well as the generalized quantile plot. One of the main reasons is that adjacent terms in $\{e_{j,p}\}_{j=1}^{k_n}$ contain similar information, which leads to more continuity and stability of the plot.

It is noteworthy that (2.2) cannot be directly applied to the case $p = 0$. Therefore, we should consider the limit:

$$\lim_{p \to 0} \frac{1}{p} \log\left(e_{j,p}\right) = \frac{1}{j} \sum_{i=1}^{j} \log\left(X_{n-i+1,n}\right). \tag{2.3}$$

Based on (2.3) and (2.21) in Sect. 2.6, we can now define a new extreme value index estimator, namely, excess mean of power (EMP) estimator, which approximates the slope of the excess mean of order-p (MOP) plot, as the estimation of γ. The proposed estimator is given by

$$\hat{\gamma}^{\text{EMP}}_{m_n,p}(k_n) := \begin{cases} \frac{1}{p(k_n-m_n)}\left[\sum_{j=m_n+1}^{k_n}\left\{\log(e_{j,p}) - \log(e_{k_n+1,p})\right\}\right. \\ \left. +m_n\left\{\log(e_{m_n+1,p}) - \log(e_{k_n+1,p})\right\}\right], & p \neq 0, \\ \frac{1}{k_n-m_n}\left[\sum_{j=m_n+1}^{k_n}\left\{\frac{1}{j}\sum_{i=1}^{j}\log\left(X_{n-i+1,n}\right)\right\}\right. \\ \left. +\frac{m_n}{m_n+1}\sum_{i=1}^{m_n}\log\left(X_{n-i+1,n}\right) - \frac{k_n}{k_n+1}\sum_{i=1}^{k_n+1}\log\left(X_{n-i+1,n}\right)\right], & p = 0, \end{cases} \tag{2.4}$$

where $m_n = o(k_n)$ and $e_{j,p} := j^{-1}\sum_{i=1}^{j} X^p_{n-i+1,n}$.

In (2.4), the extreme terms $(e_{1,p}, \dots, e_{m_n,n})$ are not taken into account as their large variances and may introduce unnecessary instability. As we shall see in the sequel, to ensure the asymptotic normality of $\hat{\gamma}^{\text{EMP}}_{m_n,p}(k_n)$, one can choose a sequence m_n such that $m_n = O(k_n^{1-\kappa})$ and $k_n^{1-\kappa} = O(m_n)$, $\forall\, \kappa \in \left(0, 1-(2-2p\gamma)^{-1}\right)$ provided that $p\gamma < 1/2$. We study its properties in this paper.

The rest of this paper is organized as follows: In Sect. 2.2, we investigate the relationship between the EMP and MOP estimators and establish their asymptotic normalities for dependent observations under some mild conditions. In Sect. 2.3, we conduct asymptotic comparison of the EMP estimator with other contenders at optimal thresholds for independent and identically distributed (i.i.d.) observations. In Sect. 2.4, we focus on finite sample properties of the EMP estimator through simulations. Section 2.5 concludes this paper. Technical details are given in Sect. 2.6.

2.2 Asymptotic Properties

2.2.1 Asymptotic Property of the Empirical Tail Process $Q_n(t)$ for Dependent Sequences

To facilitate our subsequent discussion, we first investigate the asymptotic property of the empirical tail process $Q_n(t)$ for a stationary sequence $(X_1, X_2, \dots, X_n)$, where $Q_n(t)$ is defined as

$$Q_n(t) := F_n^{-1}\left(1 - \frac{k_n}{n}t\right) = X_{n-\lceil k_n t\rfloor,n},$$

$F_n(t)$ is the empirical distribution function for $\{X_i\}_{i=1}^n$, $\lceil x \rceil$ represents the integer part of x.

Assume that the β-mixing (absolutely regular) condition holds, namely, the β-mixing coefficient $\beta(k)$ satisfies

$$\beta(k) := \sup_{l \in N} E\left(\sup_{A \in \mathscr{F}_{l+k+1}^{\infty}} \left| P\left(A \middle| \mathscr{F}_1^l\right) - P(A) \right| \right) \to 0 \qquad \text{as } k \to \infty,$$

where $\mathscr{F}_1^l$ and $\mathscr{F}_{l+k+1}^{\infty}$ denote the σ-fields generated by $\{U_i\}_{i=1}^l$ and $\{U_i\}_{i=l+k+1}^{\infty}$ respectively.

As seen in Drees [12], many common time series models, including but not limited to ARMA, ARCH, and GARCH models, can generate β-mixing sequences where $\beta(k)$ vanishes in an exponential rate, provided that some natural conditions are satisfied.

Furthermore, we assume that there exists a constant $\epsilon > 0$, a sequence $(l_n)_{n \in N}$ as well as a function $r(x, y)$ such that the following are satisfied:

$(C1)$ $\lim_{n\to\infty} \frac{\beta(l_n)}{l_n} n + l_n k_n^{-\frac{1}{2}} \log^2 k_n = 0$;
$(C2)$ For any $0 \le x < y \le 1 + \epsilon$,

$$\lim_{n\to\infty} \frac{n}{l_n k_n} Cov\left(\sum_{i=1}^{l_n} I\left(X_i > F^{-1}\left(1 - \frac{k_n}{n} x\right)\right), \sum_{i=1}^{l_n} I\left(X_i > F^{-1}\left(1 - \frac{k_n}{n} y\right)\right)\right) = r(x, y),$$

where $I(x)$ is the indicator function;
$(C3)$ For some constants $C, \epsilon > 0$, $\forall\, 0 \le x < y \le 1 + \epsilon$,

$$\frac{n}{l_n k_n} E\left(\sum_{i=1}^{l_n} I\left(F^{-1}\left(1 - \frac{k_n}{n} y\right) < X_i \le F^{-1}\left(1 - \frac{k_n}{n} x\right)\right)\right)^4 \le C(y - x).$$

Before establishing the asymptotic property of $Q_n(t)$, we first present the results in Drees [11].

Lemma 2.1 (Drees [11]) *Suppose that $\{U_i\}_{i=1}^n$ is a stationary, β-mixing sequence according to the standard uniform distribution, and Conditions $(C1)$–$(C3)$ hold. Then, for any $\gamma \in R$,*

$$\sup_{t \in [\frac{1}{2k_n}, 1]} t^{\gamma + \frac{1}{2}} (1 + |\log t|)^{-\frac{1}{2}} k_n^{\frac{1}{2}} \left| \frac{(V_n(t))^{-\gamma} - t^{-\gamma}}{\gamma} - k_n^{-\frac{1}{2}} t^{-(\gamma+1)} e(t) \right| \to_p 0, \quad (2.5)$$

where $V_n(t)$ is defined as

$$V_n(t) := \frac{n}{k_n}\left(1 - U_{n-[k_n t], n}\right), \qquad t \in [0, 1],$$

and $e(t)$ is a centered Gaussian Process with covariance function $r(x, y)$, satisfying

$$\lim_{\theta \downarrow 0} P\left\{ \sup_{t \in (0, \theta]} \left| \frac{e(t)}{t^{\frac{1}{2}} (1 + |\log t|)^{\frac{1}{2}}} \right| > \delta \right\} = 0, \qquad (2.6)$$

$\forall\, \delta > 0.$

To prove Lemma 2.1, we need to partition the sequence $\{U_i\}_{i=1}^n$ into blocks with length l_n. Condition $(C1)$ not only guarantees that l_n is large enough so that the non-adjacent blocks tend to be independent, but also ensures that l_n does not grow too fast so that the effect of each block on e_n is insignificant. Condition $(C2)$ is mainly used to provide the covariance structure for the limiting process $e(t)$. Condition $(C3)$ restricts the size of the cluster in extreme intervals so that for any extreme interval it is not too clustered. As discussed in Drees [12], finite-order casual ARMA(p, q) and ARCH(1) models satisfy Conditions $(C1)$–$(C3)$ under some general conditions. For further discussions on Conditions $(C1)$–$(C3)$ and more concrete examples, we refer the readers to Resnick and Stărică [19] and Drees [11, 12].

Lemma 2.1 establishes the convergence property of the empirical tail process for the uniform distribution. Furthermore, for an arbitrary heavy-tailed continuous random variable X with $U(x) := F^{-1}(1 - x^{-1})$, because $1 - 1/U^{-1}(X)$ follows Uniform(0,1) distribution, we can extend the convergence property for heavy-tailed distributions in general. Hence, based on Lemma 2.1, we have the following theorem:

Theorem 2.1 *Suppose that $\{X_i\}_{i\in N}$ is a stationary and β-mixing sequence with continuous marginal d.f. $F \in D(G_\gamma)$, $\gamma > 0$, and satisfies the second-order condition, namely,*

$$\lim_{x\to\infty} \frac{\frac{U(tx)}{U(x)} - t^\gamma}{A(x)} = t^\gamma \Psi(t), \tag{2.7}$$

where $\Psi(t)$ is given by

$$\Psi(t) = \begin{cases} \frac{t^\rho - 1}{\rho}, & \rho < 0, \\ \log t, & \rho = 0. \end{cases}$$

Moreover, assume that conditions in Lemma 2.1 hold, and k_n is an intermediate sequence satisfying $k_n^{\frac{1}{2}} A(n/k_n) \to \phi \in \mathbb{R}$. Then, there exists a centered Gaussian process $e(t)$ satisfying (2.6) with covariance function $r(x, y)$ such that

$$\sup_{t\in(0,1]} t^{\gamma+\frac{1}{2}}(1 + |\log t|)^{-\frac{1}{2}} \left| k_n^{\frac{1}{2}} \left(\frac{Q_n(t)}{U(n/k_n)} - t^{-\gamma} \right) - \gamma t^{-(\gamma+1)} e(t) - \phi t^{-\gamma} \Psi_0(t) \right| \to_p 0, \tag{2.8}$$

where $A(x) = \gamma\beta x^\rho$ and $\Psi_0(t)$ is given by

$$\Psi_0(t) = \begin{cases} \frac{t^{-\rho} - 1}{\rho}, & \rho < 0, \\ -\log t, & \rho = 0. \end{cases}$$

In the literature, the second-order condition is widely-adopted, which can lead to the asymptotic normality of EVI estimators. The commonly used Hall–Welsh class of models is a special case satisfying (2.7) and the function $U(t)$ has the following representation:

$$U(t) = Ct^{\gamma}\left(1 + \gamma\beta t^{\rho}/\rho + o(t^{\rho})\right), \qquad \text{as } t \to \infty,$$

with $C > 0$, $\rho < 0$. Note that Pareto, Fréchet, Student-t, Burr, and Extreme Value distributions all belong to this class.

It is noteworthy that Drees ([11], Theorem 3.1) also provided a similar result. However, it does not specify the second-order framework (2.7); a more strict restriction on k_n must be required, namely,

$$k_n^{\frac{1}{2}} \sup_{t\in(0,t_0]} t^{\gamma+\frac{1}{2}}(1+|\log t|)^{-\frac{1}{2}} \left| \frac{U\left(\frac{n}{k_n}t^{-1}\right) - U\left(\frac{n}{k_n}\right)}{a\left(\frac{n}{k_n}\right)} - \frac{t^{-\gamma}-1}{\gamma} \right| \to 0, \quad (2.9)$$

for some $t_0 > 1$. In fact, given the second-order condition, we can immediately derive from (2.9) that $k_n^{\frac{1}{2}} A\left(n/k_n\right) \to 0$, and the term $\phi t^{-\gamma}\Psi_0(t)$ in (2.8) automatically disappears; see De Haan and Ferreira ([8], Corollary 2.3.5). Under the second-order condition, we have

$$\lim_{n\to\infty} \frac{\frac{U\left(\frac{n}{k_n}t^{-1}\right)-U\left(\frac{n}{k_n}\right)}{a\left(\frac{n}{k_n}\right)} - \frac{t^{-\gamma}-1}{\gamma}}{A\left(\frac{n}{k_n}\right)} = \tilde{\Psi}(t^{-1}),$$

$\forall\, t \in (0, 1]$, where

$$\tilde{\Psi}(t) = \begin{cases} \frac{t^{\gamma+\rho}-1}{\gamma+\rho}, & \gamma+\rho \neq 0, \rho < 0, \\ \log t, & \gamma+\rho = 0, \rho < 0, \\ \frac{1}{\gamma}t^{\gamma}\log t, & \rho = 0 \neq \gamma. \end{cases}$$

Hence, for any fixed $t_1 \in (0, 1)$, based on (2.9), we have

$$\begin{aligned} \left| k_n^{\frac{1}{2}} A\left(\frac{n}{k_n}\right) \right| &= \left| k_n^{\frac{1}{2}} \frac{\frac{U\left(\frac{n}{k_n}t_1^{-1}\right)-U\left(\frac{n}{k_n}\right)}{a\left(\frac{n}{k_n}\right)} - \frac{t^{-\gamma}-1}{\gamma}}{\tilde{\Psi}(t_1^{-1})}(1+o^{(n)}(1)) \right| \\ &< \frac{2k_n^{\frac{1}{2}}}{\tilde{\Psi}(t_1^{-1})} \left| \frac{U\left(\frac{n}{k_n}t_1^{-1}\right) - U\left(\frac{n}{k_n}\right)}{a\left(\frac{n}{k_n}\right)} - \frac{t_1^{-\gamma}-1}{\gamma} \right| \\ &\to 0. \end{aligned}$$

Compared with the results in Drees [11], the main advantage of Theorem 2.1 is that by introducing the widely accepted second-order condition (2.7), it paves a way to represent the asymptotic mean of the limiting distribution for $k_n^{\frac{1}{2}}\left(\hat{\gamma}^{\cdot} - \gamma\right)$ as a more explicit expression of the second-order parameter ρ, which has been intensively studied in the literature, see, for instance, Alves et al. [1], Goegebeur et al. [13] and Ciuperca and Mercadier [6]. After obtaining explicit forms of asymptotic mean and variance, we can construct confidence intervals, derive asymptotic RMSE, and compare asymptotic performances among various estimators.

2.2.2 Connections Between EMP and MOP Estimators

In this subsection, we discuss the connection between EMP and MOP estimators to offer another perspective of the proposed estimator. In the case $p > 0$, by using the approximation $\log(1+x) = x + o(x)$ as $x \to 0$, we have

$$
\begin{aligned}
\hat{\gamma}_{m_n,p}^{\mathrm{EMP}}(k_n) &= \frac{1}{p(k_n - m_n)} \left[\sum_{i=m_n+1}^{k_n} \log\left(\frac{e_{i,p}}{e_{k_n+1,n}}\right) + m_n \log\left(\frac{e_{m_n+1,p}}{e_{k_n+1,n}}\right) \right] \\
&= \frac{1}{p(k_n - m_n)} \left[\sum_{i=m_n+1}^{k_n} \sum_{j=i}^{k_n} \log\left(\frac{e_{j,p}}{e_{j+1,p}}\right) + m_n \log\left(\frac{e_{m_n+1,p}}{e_{k_n+1,n}}\right) \right] \\
&= \frac{1}{p(k_n - m_n)} \sum_{j=m_n+1}^{k_n} j \log\left(\frac{e_{j,p}}{e_{j+1,p}}\right) \\
&= \frac{1}{p(k_n - m_n)} \sum_{j=m_n+1}^{k_n} j \left\{ \log\left(1 + \frac{1}{j}\right) - \log\left(1 + \frac{1}{\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p}\right) \right\} \\
&= \frac{1}{k_n - m_n} \sum_{j=m_n+1}^{k_n} \left(\frac{1 - j \Big/ \sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p}{p} \right) (1 + o_p(1)).
\end{aligned}
$$

On the other hand, if $p = 0$, we have

$$
\begin{aligned}
\hat{\gamma}_{m_n,0}^{\mathrm{EMP}}(k_n) &= \frac{1}{k_n - m_n} \sum_{j=m_n+1}^{k_n} \hat{\gamma}^{\mathrm{Hill}}(j) + \frac{1}{k_n - m_n} \left\{ \hat{\gamma}^{\mathrm{Hill}}(k_n+1) - \hat{\gamma}^{\mathrm{Hill}}(m_n+1) \right. \\
&\quad \left. + \log\left(\frac{X_{n-k_n,n}}{X_{n-m_n,n}}\right) \right\}. \\
&= \frac{1}{k_n - m_n} \sum_{j=m_n+1}^{k_n} \hat{\gamma}^{\mathrm{Hill}}(j) + o_p\left(k_n^{-\frac{1}{2}}\right).
\end{aligned}
$$

Meanwhile, the MOP estimator is defined as follows:

$$\hat{\gamma}_p^{\mathrm{MOP}}(k_n) := \begin{cases} \dfrac{1-k_n \Big/ \sum_{i=1}^{k_n} \left(\frac{X_{n-i+1,n}}{X_{n-k_n,n}}\right)^p}{p}, & p \neq 0, \\ \hat{\gamma}^{\mathrm{Hill}}(k_n), & p = 0. \end{cases}$$

Therefore, the EMP estimator may be regarded as an approximation of the average of $\hat{\gamma}_p^{\mathrm{MOP}}(j)$ over different thresholds $j = m_n + 1, \ldots, k_n$, which can lead to more stable estimates.

In the literature, there is usually a tradeoff between bias and variance in terms of the choice of k_n. A smaller k_n can often lead to less bias, but brings more variance. By taking average of MOP estimators over different thresholds, the EMP estimator combines the ones with less bias as well as those with less variance. Hence, it can keep the balance between these two factors within a wide range of choices of thresholds k_n, which can be regarded as the main source of robustness for the EMP estimator. For a comparatively large k_n, the EMP estimator still contains much information from those $\hat{\gamma}_p^{\mathrm{MOP}}(j)$ with j small enough, so the bias term grows more slowly than other EVI estimators.

2.2.3 Asymptotic Normalities of the MOP and the EMP Estimators

Regarding the existing MOP estimator, when $p \neq 0$, we can rewrite it as a functional of $Q_n(t)$:

$$\hat{\gamma}_p^{\mathrm{MOP}}(k_n) = \frac{1 - 1 \Big/ \int_0^1 \left(\frac{Q_n(t)}{Q_n(1)}\right)^p dt}{p}. \tag{2.10}$$

The asymptotic normality of MOP estimator for $p \in (0, (2\gamma)^{-1})$ in the independent case was established in Brilhante et al. [5]. In the following theorem, we utilize (2.10) and the asymptotic property of $Q_n(t)$ established in (2.8) to extend the existing result in two aspects: Firstly, we allow the tuning parameter p to take negative values. Secondly, we generalize the conclusion to dependent observations.

Theorem 2.2 *Suppose that $\{X_i\}_{i\in N}$ is a stationary and β-mixing sequence satisfying conditions in Lemma 2.1 and Theorem 2.1, and k_n is an intermediate sequence satisfying $k_n^{\frac{1}{2}} A(n/k_n) \to \phi \in R$. Furthermore, the tuning parameter p satisfies $p\gamma < 1/2$. Then, we have*

$$k_n^{\frac{1}{2}} \left(\hat{\gamma}_p^{\mathrm{MOP}}(k_n) - \gamma\right) \to_d N\left(\mu_{p,\gamma,\phi,\rho}^{\mathrm{MOP}}, \sigma_{\mathrm{MOP},p,\gamma,r}^2\right),$$

where

$$\mu_{p,\gamma,\phi,\rho}^{\mathrm{MOP}} = \frac{(1-p\gamma)\phi}{1-p\gamma-\rho},$$

and

$$\sigma^2_{\text{MOP},p,\gamma,r} = \gamma^2(1-p\gamma)^4 \int_0^1 \int_0^1 (st)^{-p\gamma-1} r(s,t)\, ds\, dt \\ -2\gamma^2(1-p\gamma)^3 \int_0^1 t^{-p\gamma-1} r(t,1)\, dt + \gamma^2(1-p\gamma)^2\, r(1,1).$$

As a special case, if $\{X_i\}_{i=1}^n$ are independent, then $e(t)$ in (2.8) is a Brownian motion and $r(s,t) = \min(s,t)$. We can deduce that

$$\sigma^2_{\text{MOP},p,\gamma,r} = \frac{\gamma^2(1-p\gamma)^2}{1-2p\gamma},$$

which accords with the existing results in the literature for the MOP estimator.

In terms of the proposed the EMP estimator, recall that

$$\hat{\gamma}^{\text{EMP}}_{m_n,p}(k_n) \approx \frac{1}{k_n - m_n} \sum_{j=m_n+1}^{k_n} \hat{\gamma}^{\text{MOP}}_p(j).$$

Hence, by applying the same techniques that we have used on MOP estimator, we can likewise establish the asymptotic normality for the EMP estimator.

Theorem 2.3 *Suppose that $\{X_i\}_{i\in N}$ is a stationary and β-mixing sequence satisfying conditions in Lemma 2.1 and Theorem 2.2, and k_n is an intermediate sequence satisfying $k_n^{\frac{1}{2}} A(n/k_n) \to \phi \in R$. Furthermore, the tuning parameter p satisfies $p\gamma < 1/2$. Then, for all sequences m_n such that $m_n = O(k_n^{1-\kappa})$ and $k_n^{1-\kappa} = O(m_n)$, $\forall\, \kappa \in \big(0, 1-(2-2p\gamma)^{-1}\big)$, we have*

$$k_n^{\frac{1}{2}} \left(\hat{\gamma}^{\text{EMP}}_{m_n,p}(k_n) - \gamma\right) \to_d N(\mu^{\text{EMP}}_{p,\gamma,\phi,\rho}, \sigma^2_{\text{EMP},p,\gamma,r}),$$

where

$$\mu^{\text{EMP}}_{p,\gamma,\phi,\rho} = \frac{(1-p\gamma)\phi}{(1-p\gamma-\rho)(1-\rho)},$$

and

$$\sigma^2_{\text{EMP},p,\gamma,r} = \gamma^2(1-p\gamma)^2 \int_0^1 \int_0^1 (us)^{-1} r(u,s)\, du\, ds \\ -2\gamma^2(1-p\gamma)^3 \int_0^1 \int_0^1 \int_0^1 (us)^{-1} t^{-p\gamma-1} r(st,u)\, du\, ds\, dt \\ +\gamma^2(1-p\gamma)^4 \int_0^1 \int_0^1 \int_0^1 \int_0^1 (us)^{-1} (vt)^{-p\gamma-1} r(uv,st)\, du\, dv\, ds\, dt.$$

In the case that $\{X_i\}_{i=1}^n$ are independent, we have $r(s,t) = \min(s,t)$, which leads to the following corollary:

Corollary 2.1 *Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. continuous random variables following distribution F, with $F \in D(G_\gamma)$ and $\gamma > 0$. Moreover, $F(x)$ satisfies the second-order condition. Assume that $k_n \to \infty$, $k_n/n \to 0$ and $k_n^{\frac{1}{2}} A(n/k_n) \to \phi \in (-\infty, +\infty)$ as $n \to \infty$, and that the tuning parameter p satisfies $p\gamma < 1/2$. Then, for all sequences m_n such that $m_n = O(k_n^{1-\kappa})$ and $k_n^{1-\kappa} = O(m_n)$, $\forall\ \kappa \in \left(0, 1-(2-2p\gamma)^{-1}\right)$, we have*

$$k_n^{\frac{1}{2}} \left(\hat{\gamma}_{m_n,p}^{\mathrm{EMP}}(k_n) - \gamma\right) \to_d N(\mu_{p,\gamma,\phi,\rho}, \sigma_{p,\gamma}^2),$$

where

$$\mu_{p,\gamma,\phi,\rho} = \frac{(1-p\gamma)\phi}{(1-p\gamma-\rho)(1-\rho)} \quad \text{and} \quad \sigma_{p,\gamma}^2 = \frac{2\gamma^2(1-p\gamma)}{1-2p\gamma}.$$

2.2.4 Consistent Estimators of $\sigma^2_{\mathrm{MOP},p,\gamma,r}$ and $\sigma^2_{\mathrm{EMP},p,\gamma,r}$

Constructions of the confidence interval of either EVI γ or extreme quantile x_q require estimations on variance of the corresponding the EVI estimator. Because the covariance function $r(x,y)$ is involved in both $\sigma^2_{\mathrm{MOP},p,\gamma,r}$ and $\sigma^2_{\mathrm{EMP},p,\gamma,r}$, it is not easy to estimate them in general. Avoiding direct estimation on $r(x,y)$, Drees [12] proposed three types of estimators for asymptotic variance, where stronger conditions rather than $(C1)$–$(C3)$ were assumed. Furthermore, the homogeneity of covariance function $r(x,y)$ was required, namely,

$$r(\lambda x, \lambda y) = \lambda r(x,y), \quad \forall\ \lambda, x, y \in [0,1], \tag{2.11}$$

under which the corresponding Gaussian process $e(t)$ is self-similar:

$$e(\lambda t) =_d \lambda^{\frac{1}{2}} e(t), \quad \forall\ \lambda \in [0,1].$$

However, the conditions in Drees [12] are somehow restrictive in terms of the threshold k_n, which can give rise to

$$k_n^{\frac{1}{2}} \left(\hat{\gamma}^{\boldsymbol{\cdot}} - \gamma\right) \to_d N(0, \sigma^2_{\hat{\gamma}^{\boldsymbol{\cdot}}}),$$

where $\hat{\gamma}^{\boldsymbol{\cdot}}(\cdot)$ denotes $\hat{\gamma}_p^{\mathrm{MOP}}(\cdot)$ or $\hat{\gamma}_p^{\mathrm{EMP}}(\cdot)$ whenever appropriate. Under our settings, namely, the more relaxed conditions regarding k_n in Theorem 2.3, the asymptotic mean of $k_n^{\frac{1}{2}} \left(\hat{\gamma}^{\boldsymbol{\cdot}} - \gamma\right)$ is not necessarily to be 0. Therefore, the conclusions in Drees

[12] cannot directly be applied to our case. In Theorem 2.4, we make an assertion that the estimator

$$\hat{\sigma}^2_{\hat{\gamma}^{\cdot}} := \left(\log \frac{k_n}{j_n}\right)^{-1} \sum_{i=j_n}^{k_n} \left(\hat{\gamma}^{\cdot}(i) - \hat{\gamma}^{\cdot}(k_n)\right)^2,$$

first proposed in Drees [12], is also valid for the MOP and EMP estimators under the conditions in Theorem 2.2 and (2.11).

Theorem 2.4 *Suppose that the conditions in Lemma 2.1 and Theorem 2.2 hold, and $r(x, y)$ satisfies (2.11). Then, there exist sequences $j_n^{(1)} = o(k_n)$ and $j_n^{(2)} = o(k_n)$ such that*

$$\hat{\sigma}^2_{\text{MOP},1} := \left(\log \frac{k_n}{j_n^{(1)}}\right)^{-1} \sum_{i=j_n^{(1)}}^{k_n} \left(\hat{\gamma}_p^{\text{MOP}}(i) - \hat{\gamma}_p^{\text{MOP}}(k_n)\right)^2 \to_p \sigma^2_{\text{MOP},p,\gamma,r},$$

$$\hat{\sigma}^2_{\text{EMP},1} := \left(\log \frac{k_n}{j_n^{(2)}}\right)^{-1} \sum_{i=j_n^{(2)}}^{k_n} \left(\hat{\gamma}_{m_n,p}^{\text{EMP}}(i) - \hat{\gamma}_{m_n,p}^{\text{EMP}}(k_n)\right)^2 \to_p \sigma^2_{\text{EMP},p,\gamma,r}.$$

As discussed by Drees [12], any sequence j_n such that j_n/k_n vanishes not too fast can lead to the consistency. In practice, one may often choose j_n to be rather small. Specifically, one may take $j_n^{(2)} = m_n + 1$ for $\hat{\sigma}^2_{\text{EMP},1}$, and take $j_n^{(1)} = 1$ for $\hat{\sigma}^2_{\text{MOP},1}$.

Furthermore, as seen in Sect. 2.6, the proof of Theorem 2.4 involves a direct elimination of the term $e^{\frac{u}{2}}\tilde{R}(0)$. Though this approximation is correct in a theoretical way, for moderate sample sizes it is too crude, which results in short confidence intervals in practice. To handle this problem, we refer to (30)–(32) in Drees [12] and use the following estimators as the alternatives for $\hat{\sigma}^2_{\text{MOP},1}$ and $\hat{\sigma}^2_{\text{EMP},1}$ respectively, which give rise to more conservative confidence intervals:

$$\hat{\sigma}^2_{\text{MOP},2} := \left(\sum_{i=j_n^{(1)}}^{k_n} \left(i^{-\frac{1}{2}} - k_n^{-\frac{1}{2}}\right)^2\right)^{-1} \sum_{i=j_n^{(1)}}^{k_n} \left(\hat{\gamma}_p^{\text{MOP}}(i) - \hat{\gamma}_p^{\text{MOP}}(k_n)\right)^2,$$

$$\hat{\sigma}^2_{\text{EMP},2} := \left(\sum_{i=j_n^{(2)}}^{k_n} \left(i^{-\frac{1}{2}} - k_n^{-\frac{1}{2}}\right)^2\right)^{-1} \sum_{i=j_n^{(2)}}^{k_n} \left(\hat{\gamma}_{m_n,p}^{\text{EMP}}(i) - \hat{\gamma}_{m_n,p}^{\text{EMP}}(k_n)\right)^2.$$

Notice that since

$$\sum_{i=j_n}^{k_n} \left(i^{-\frac{1}{2}} - k_n^{-\frac{1}{2}}\right)^2 \to \log \frac{k_n}{j_n},$$

as $k_n \to \infty$, $\hat{\sigma}^2_{\text{MOP},2}$ and $\hat{\sigma}^2_{\text{EMP},2}$ are also consistent estimators of $\sigma^2_{\text{MOP},p,\gamma,r}$ and $\sigma^2_{\text{EMP},p,\gamma,r}$, respectively. In the following, we focus on using $\hat{\sigma}^2_{\text{MOP},2}$ and $\hat{\sigma}^2_{\text{EMP},2}$ as the asymptotic variance estimators.

2.3 Asymptotic Comparison of EMP with Existing Contenders for I.I.D. Observations

In addition to the Hill, MOP, and generalized Hill (GH) estimators, moment (Mom) estimator and mixed moment (MM) estimator are also well-known in the literature. Define

$$M^{(j)}_{k_n,n} := \frac{1}{k_n}\sum_{i=1}^{k_n}\left(\log(X_{n-i+1,n}) - \log(X_{n-k_n,n})\right)^j,$$

$$L^{(j)}_{k_n,n} := \frac{1}{k_n}\sum_{i=1}^{k_n}\left(1 - \frac{X_{n-k_n,n}}{X_{n-i+1,n}}\right)^j,$$

where $j \geq 1$. The Mom estimator, denoted as $\hat{\gamma}^{\text{Mom}}(k_n)$, is given by

$$\hat{\gamma}^{\text{Mom}}(k_n) := M^{(1)}_{k_n,n} + \frac{1}{2}\left[1 - \left\{M^{(2)}_{k_n,n} \Big/ \left(M^{(1)}_{k_n,n}\right)^2 - 1\right\}^{-1}\right].$$

The MM estimator, denoted as $\hat{\gamma}^{\text{MM}}(k_n)$, is given by

$$\hat{\gamma}^{\text{MM}}(k_n) := \frac{\hat{\phi}_{k_n,n} - 1}{1 + 2\min(\hat{\phi}_{k_n,n} - 1, 0)},$$

where

$$\hat{\phi}_{k_n,n} := \frac{M^{(1)}_{k_n,n} - L^{(1)}_{k_n,n}}{\left(L^{(1)}_{k_n,n}\right)^2}.$$

To the best of our knowledge, asymptotic comparisons of various EVI estimators in the literature require the independence assumption. On the other hand, because the asymptotic normalities of the EVI estimators established for dependent observations involve an unknown covariance function $r(x, y)$, it is very tedious to conduct asymptotic comparisons in the dependent framework. Therefore, in this section, we focus on comparing asymptotic mean square errors (AMSEs) of different EVI estimators for i.i.d. observations.

Combining the existing results in the literature and Corollary 2.1, we have

Table 2.1 Explicit expressions of $b_{\cdot}$ and $\sigma_{\cdot}^2$ for different estimators

	$b_{\cdot}$	$\sigma_{\cdot}^2$
Hill	$\frac{1}{1-\rho}$	γ^2
Mom	$\frac{\gamma-\gamma\rho+\rho}{\gamma(1-\rho)^2}$	$1+\gamma^2$
GH	$\frac{\gamma-\gamma\rho+\rho}{\gamma(1-\rho)^2}$	$1+\gamma^2$
MM	$\frac{(1+\gamma)(\gamma+\rho)}{\gamma(1-\rho)(1+\gamma-\rho)}$	$(1+\gamma)^2$
MOP	$\frac{1-p\gamma}{1-p\gamma-\rho}$	$\frac{\gamma^2(1-p\gamma)^2}{1-2p\gamma}$
EMP	$\frac{1-p\gamma}{(1-p\gamma-\rho)(1-\rho)}$	$\frac{2\gamma^2(1-p\gamma)}{1-2p\gamma}$

$$k_n^{\frac{1}{2}}\left(\hat{\gamma}^{\cdot}(k_n)-\gamma\right) \to_d N(\phi b_{\cdot}, \sigma_{\cdot}^2), \qquad \text{as } n \to \infty, \tag{2.12}$$

for the Hill, GH, Mom, MM, MOP, and EMP estimators, where the explicit expressions of $b_{\cdot}$ and $\sigma_{\cdot}^2$ are shown in Table 2.1.

We observe that

$$\begin{cases} b_{\text{Hill}} \geq b_{\text{MOP}} \geq b_{\text{EMP}} > 0, & p > 0, \\ b_{\text{Hill}} = b_{\text{MOP}} \geq b_{\text{EMP}} > 0, & p = 0, \\ b_{\text{MOP}} \geq b_{\text{Hill}} \geq b_{\text{EMP}} > 0, & p < 0, \end{cases} \tag{2.13}$$

where any equality in (2.13) holds if and only if $\rho = 0$. Also,

$$\begin{cases} \sigma^2_{\text{EMP}} > \sigma^2_{\text{MOP}} \geq \sigma^2_{\text{Hill}}, & p\gamma > -1, \\ \sigma^2_{\text{EMP}} = \sigma^2_{\text{MOP}} > \sigma^2_{\text{Hill}}, & p\gamma = -1, \\ \sigma^2_{\text{MOP}} > \sigma^2_{\text{EMP}} > \sigma^2_{\text{Hill}}, & p\gamma < -1, \end{cases} \tag{2.14}$$

where the equality in (2.14) holds if and only if $p = 0$.

By (2.12), we can obtain the AMSE for the EVI estimators:

$$\text{AMSE}(\hat{\gamma}^{\cdot}(k_n)) := \frac{\sigma_{\cdot}^2}{k_n} + b_{\cdot}^2 A\left(\frac{n}{k_n}\right).$$

Furthermore, referring to Dekkers and De Haan [10], whenever $b_{\cdot} \neq 0$, there exists some function $\psi(n, \gamma, \rho)$, such that under the optimal threshold k_n^*, we have

$$\lim_{n\to\infty} \psi(n, \gamma, \rho)\text{AMSE}(\hat{\gamma}^{\cdot}(k_n^*)) = \left\{(\sigma_{\cdot}^2)^{-\rho} b_{\cdot}\right\}^{\frac{2}{1-2\rho}} =: \text{LMSE}(\hat{\gamma}^{\cdot}). \tag{2.15}$$

Because the limiting MSE (LMSE) (2.15) may just involve the terms γ, ρ, and tuning parameter p, comparison between different estimators becomes simpler. In the literature, the LMSE is often used as the comparison criterion, see De Haan and Peng [9], Gomes and Martins [15], Gomes and Neves [16], Gomes and Henriques-

Rodrigues [14] and Brilhante et al. [5]. In this section, we also follow this way and focus on the asymptotic comparison at the optimal threshold.

As discussed in Brilhante et al. [5], for the MOP estimator, the optimal LMSE is attained at

$$p^*_{\text{MOP}} = \frac{1 - \frac{\rho}{2} - \frac{\sqrt{\rho^2 - 4\rho + 2}}{2}}{\gamma}.$$

Furthermore, for the entire plane (γ, ρ) with $\gamma > 0$, $\rho \le 0$, it has been shown that $\text{LMSE}(\hat{\gamma}^{\text{MOP}}_{p^*_{\text{MOP}}}) \le \text{LMSE}(\hat{\gamma}^{\text{Hill}})$ and the equality holds if and only if $\rho = 0$. In the following, we mainly compare the EMP estimator with the MOP estimator. To begin with, we give the explicit form of the optimal tuning parameter for the EMP estimator that can minimize the corresponding LMSE.

Theorem 2.5 *For $\gamma > 0$, $\rho \le 0$, the EMP estimator attains the minimum LMSE when $p = p^*_{\text{EMP}} := \rho\gamma^{-1}$.*

While p^*_{MOP} is always positive, p^*_{EMP} always takes non-positive values. After obtaining p^*_{EMP}, we can now compare the LMSE for the EMP and the MOP estimators at optimal tuning parameters p.

Theorem 2.6 *For $\gamma > 0$, $\rho \le 0$, we always have $LMSE(\hat{\gamma}^{\text{EMP}}_{m_n,p} \mid p = p^*_{\text{EMP}}) \le LMSE(\hat{\gamma}^{\text{MOP}}_p \mid p = p^*_{\text{MOP}})$, and the equality holds if and only if $\rho = 0$.*

To get a direct interpretation on the ratio

$$\frac{\text{LMSE}\left(\hat{\gamma}^{\text{MOP}}_p \mid p = p^*_{\text{MOP}}\right)}{\text{LMSE}\left(\hat{\gamma}^{\text{EMP}}_{m_n,p} \mid p = p^*_{\text{EMP}}\right)},$$

we can refer to Fig. 2.2. Numerically, the ratio attains the maximum value 1.0430 at $\rho_0 = -1.0869$. For any $\rho < 0$, we have $\text{LMSE}\left(\hat{\gamma}^{\text{MOP}}_p \mid p = p^*_{\text{MOP}}\right) > \text{LMSE}\left(\hat{\gamma}^{\text{EMP}}_{m_n,p} \mid p = p^*_{\text{EMP}}\right)$. Though the gain in efficiency is not very obvious, it is acceptable because any little improvement in terms of LMSE is not easy, see De Haan and Peng [9], Gomes and Henriques-Rodrigues [14] and Brilhante et al. [5]. Furthermore, for any fixed ρ, we take $p = p^*_{\text{MOP}}$ for the MOP estimator, while for the EMP estimator, we set different values of p. The corresponding contour map of the ratio $\text{LMSE}\left(\hat{\gamma}^{\text{MOP}}_p \mid p = p^*_{\text{MOP}}\right) / \text{LMSE}\left(\hat{\gamma}^{\text{EMP}}_{m_n,p}\right)$ is shown in Fig. 2.3. We can see that within a rather wide range of $(\rho, p\gamma)$ for the EMP estimator, it can still outperform the MOP estimator at p^*_{MOP}.

We reproduce the overall comparison procedures in Sect. 4.2 of Brilhante et al. [5]. Meanwhile, the EMP estimator is also considered. For each (γ, ρ), the one with the lowest LMSE among Hill, maximum likelihood (ML), Mom, MM, MOP, and EMP estimators is listed in Fig. 2.4. As expected, the EMP estimator now dominates all the cases that the MOP originally does in Brilhante et al. [5]. Furthermore, when ρ is close to 0, the problem becomes much more challenging because $A(t)$ in (2.7) vanishes more slowly and the bias term $b_{\cdot}$ is larger. We can find that the EMP estimator also dominates a large proportion of these most challenging cases—the top right part of Fig. 2.4, where the MOP estimator is beaten by the Mom estimator.

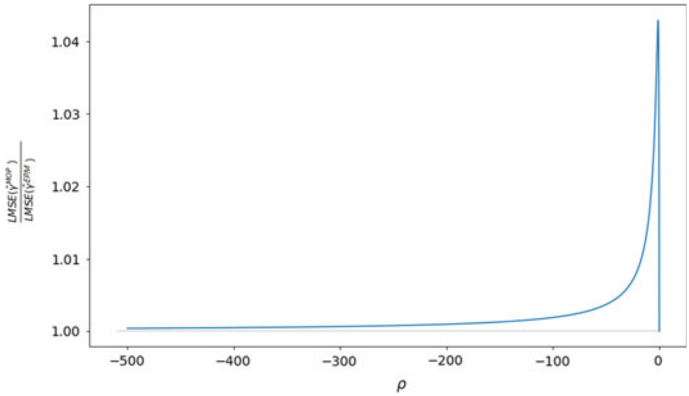


Fig. 2.2 The ratio of LMSE $\left(\hat{\gamma}_p^{\text{MOP}} \mid p = p^*_{\text{MOP}}\right)$ /LMSE $\left(\hat{\gamma}_{m_n,p}^{\text{EMP}} \mid p = p^*_{\text{EMP}}\right)$ for $\rho \in [-500, 0]$

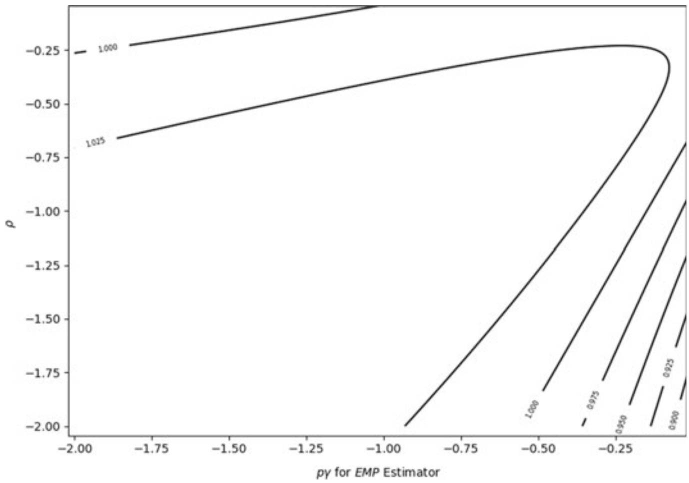


Fig. 2.3 The ratio LMSE $\left(\hat{\gamma}_p^{\text{MOP}} \mid p = p^*_{\text{MOP}}\right)$ /LMSE $\left(\hat{\gamma}_{m_n,p}^{\text{EMP}}\right)$ under different ρ and $p\gamma$ for the EMP estimator

2.4 Simulations

Simulations are implemented based on the following underlying distributions:

1. Burr(a, b) distribution, where $\gamma = (ab)^{-1}$, $\rho = -b^{-1}$. The cumulative distribution function (c.d.f.) is given by

$$F(x; a, b) = 1 - \left(1 + x^a\right)^{-b}, \qquad x \geq 0,$$

ρ \ γ	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1	1.1	1.2	1.3	1.4	1.5	1.6	1.7	1.8	1.9	2
0	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP
-0.1	ML	MM	MM	MM	MM	MM	MM	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP
-0.2	EMP	ML	MM	MM	MM	MM	MM	MM	MM	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP
-0.3	EMP	Momm	ML	MM	MM	MM	MM	MM	MM	MM	MM	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP
-0.4	EMP	EMP	Mom	ML	MM	MM	MM	MM	MM	MM	MM	MM	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP
-0.5	EMP	EMP	Mom	Mom	ML	MM	MM	MM	MM	MM	MM	MM	MM	MM	EMP	EMP	EMP	EMP	EMP	EMP
-0.6	EMP	EMP	EMP	Mom	MM	ML	MM	MM	MM	MM	MM	MM	MM	MM	MM	MM	EMP	EMP	EMP	EMP
-0.7	EMP	EMP	EMP	Mom	Mom	MM	ML	MM	MM	MM	MM	MM	MM	MM	MM	MM	MM	EMP	EMP	EMP
-0.8	EMP	EMP	EMP	Mom	Mom	Mom	MM	ML	MM	MM	MM	MM	MM	MM	MM	MM	MM	MM	MM	EMP
-0.9	EMP	EMP	EMP	EMP	Mom	Mom	Mom	MM	ML	MM	MM	MM	MM	MM	MM	MM	MM	MM	MM	MM
-1	EMP	EMP	EMP	EMP	Mom	Mom	Mom	MM	MM	ML	MM	MM	MM	MM	MM	MM	MM	MM	MM	MM
-1.1	EMP	EMP	EMP	EMP	Mom	Mom	Mom	EMP	MM	MM	ML	MM	MM	MM	MM	MM	MM	MM	MM	MM
-1.2	EMP	EMP	EMP	EMP	Mom	Mom	Mom	EMP	EMP	MM	MM	ML	MM	MM	MM	MM	MM	MM	MM	MM
-1.3	EMP	EMP	EMP	EMP	EMP	Mom	Mom	EMP	EMP	EMP	MM	MM	ML	MM	MM	MM	MM	MM	MM	MM
-1.4	EMP	EMP	EMP	EMP	EMP	Mom	Mom	EMP	EMP	EMP	EMP	MM	MM	ML	MM	MM	MM	MM	MM	MM
-1.5	EMP	EMP	EMP	EMP	EMP	Mom	Mom	EMP	EMP	EMP	EMP	EMP	MM	MM	ML	MM	MM	MM	MM	MM
-1.6	EMP	EMP	EMP	EMP	EMP	Mom	Mom	EMP	EMP	EMP	EMP	EMP	MM	MM	MM	ML	MM	MM	MM	MM
-1.7	EMP	EMP	EMP	EMP	EMP	Mom	Mom	EMP	EMP	EMP	EMP	EMP	EMP	MM	MM	MM	ML	MM	MM	MM
-1.8	EMP	EMP	EMP	EMP	EMP	Mom	Mom	EMP	EMP	EMP	EMP	EMP	EMP	EMP	MM	MM	MM	ML	MM	MM
-1.9	EMP	EMP	EMP	EMP	EMP	EMP	Mom	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	MM	MM	MM	ML	MM
-2	EMP	EMP	EMP	EMP	EMP	EMP	Mom	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	EMP	MM	MM	MM	ML

Fig. 2.4 Asymptotic comparison at optimal thresholds for different estimators under various (γ, ρ)

or equivalently, we can represent the c.d.f. in terms of parameters γ and ρ directly:

$$F(x; \gamma, \rho) = 1 - \left(1 + x^{-\frac{\rho}{\gamma}}\right)^{\frac{1}{\rho}}, \qquad x \geq 0.$$

2. Pareto(μ, σ, α)-Type II distribution, where $\gamma = \alpha^{-1}$ and $\rho = -\gamma = -\alpha^{-1}$ provided that $\mu \neq \sigma$. The c.d.f is given by

$$F(x; \mu, \sigma, \alpha) = 1 - \left(1 + \frac{x - \mu}{\sigma}\right)^{-\alpha}, \qquad x \geq \mu.$$

3. Student-t distribution with ν degrees of freedom, $T(\nu)$, where $\gamma = \nu^{-1}$, $\rho = -2\nu^{-1}$.
4. Fréchet(α) distribution, where $\gamma = \alpha^{-1}$, $\rho = -1$. The c.d.f. is given by

$$F(x; \alpha) = e^{-x^{-\alpha}}, \qquad x > 0.$$

5. Extreme Value distribution EV(γ), where $\rho = -\gamma$, and the c.d.f. is given by

$$F(x; \gamma) = e^{-(1+\gamma x)^{-\frac{1}{\gamma}}}, \qquad x > -\frac{1}{\gamma}.$$

In the following, we first implement sensitivity analysis on the term m_n for $\hat{\gamma}^{\text{EMP}}_{m_n,p}(k_n)$. Then, we compare the performance of the EMP estimator with other classical estimators under different choices of the thresholds as well as the tuning parameters p. Finally, we investigate the potential efficiencies of the EMP and the

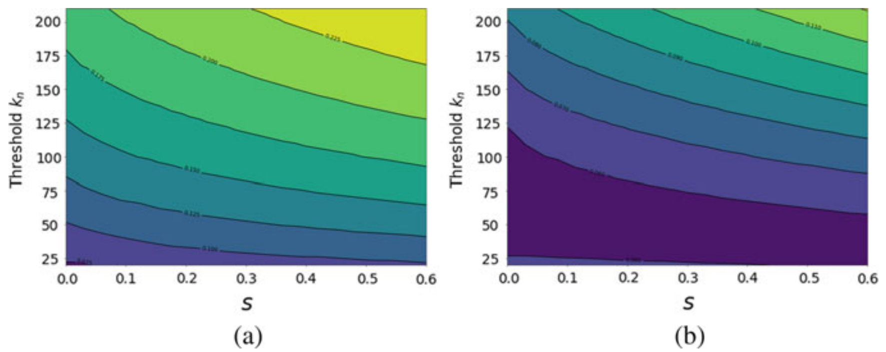


Fig. 2.5 **a** Burr(1, 4), $\gamma = 0.25$, $\rho = -0.25$, $n = 500$, $p = 1.6$; **b** Burr(2, 2), $\gamma = 0.25$, $\rho = -0.50$, $n = 500$, $p = 1.6$

MOP estimators at the simulated optimal thresholds under different tuning parameters and sample sizes.

2.4.1 Sensitivity Analysis of m_n for $\hat{\gamma}^{\text{EMP}}_{m_n,p}(k_n)$

Theoretically, to guarantee the asymptotic normality of $\hat{\gamma}^{\text{EMP}}_{m_n,p}(k_n)$, the term m_n should be large enough so that $k_n^{\kappa_0} = o(m_n)$, where $\kappa_0 = (2 - 2p\gamma)^{-1}$. Intuitively, a comparatively larger choice of m_n may help to avoid the instability coming from the most extreme terms $e_{j,p}$ when j is rather small. However, more information contained in the extreme tail fraction may also be lost. In the following, we investigate the effects of various choices of m_n on the root mean squared error $\text{RMSE}\big(\hat{\gamma}^{\text{EMP}}_{m_n,p}(k_n)\big)$ for finite samples. In each case, we obtain RMSE under different thresholds k_n as well as $m_n(s) := sk_n$ for $s \in (0, 0.6]$, based on $J = 1000$ runs. The contour maps are shown in Figs. 2.5, 2.6 and 2.7. Due to space limitations, we only show a part of cases as illustrations, where sample size $n = 500$ and $p = 0.4\gamma^{-1}$. For other simulation settings, similar results are obtained.

In contour maps, the areas with darker colors correspond to relatively lower RMSE. Generally speaking, RMSE is less sensitive to m_n compared with the threshold k_n, since the contour lines usually change smoothly in the horizontal direction. For example, in Fig. 2.5a, where the underlying model is Burr(1,4), if we fixed $k_n = 50$, then the RMSE changes around 30% when m_n rises from 0 up to $0.6k_n$. On the other hand, if we fixed $s = 0.1$, then the RMSE increases more than 80% when k_n changes from 50 to 200, or equivalently, from $0.1\,n$ to $0.4\,n$.

Since the effects of m_n are comparatively less significant, we restrict ourselves to the rule that

$$m_n = \min\left(\lceil k_n^{\tilde{\kappa}+c_0} \rfloor, k_n - 1\right) \quad \text{with} \quad \tilde{\kappa} := (2 - 2p\gamma)^{-1},$$

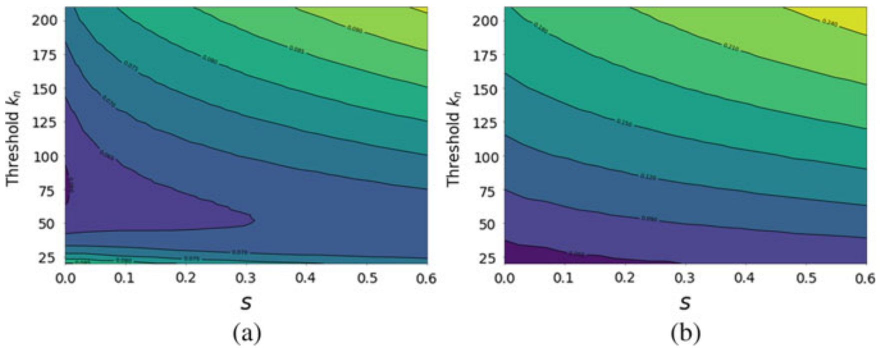


Fig. 2.6 **a** Pareto(1, 2, 3), $\gamma = 1/3$, $\rho = -1/3$, $n = 500$, $p = 1.2$; **b** $T(5)$, $\gamma = 0.2$, $\rho = -0.4$, $n = 500$, $p = 2$

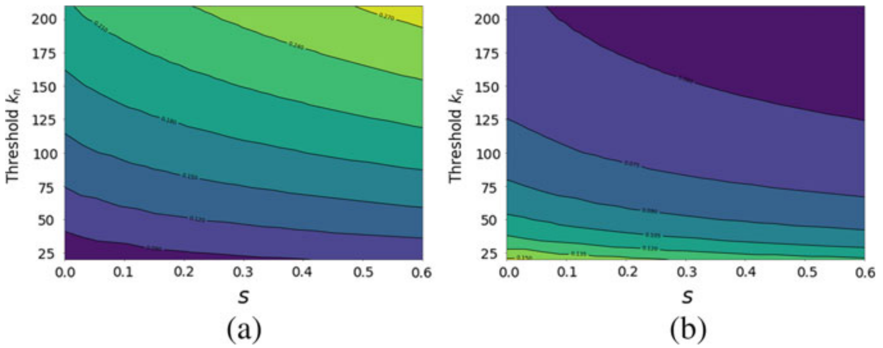


Fig. 2.7 **a** EV(0.5), $\gamma = 0.5$, $\rho = -0.5$, $n = 500$, $p = 0.8$; **b** Fréchet(2), $\gamma = 0.5$, $\rho = -1$, $n = 500$, $p = 0.8$

where c_0 is a positive constant so that the asymptotic normality of $\hat{\gamma}^{\text{EMP}}_{m_n,p}(k_n)$ is guaranteed. In practice, one may choose c_0 to be rather small. Hereinafter, we focus on the specification that $c_0 = 0.01$.

2.4.2 Simulation Results under Various Thresholds

We obtain estimates of γ based on stationary sequences generated by ARMA(1, 1) and ARCH(1) models, where Conditions $(C1)$–$(C3)$ as well as (2.11) have already been verified in Drees [12]. To be specific, the models are listed as follows:

1. ARMA(1, 1) model:

$$X_t = 0.85\ X_{t-1} + Y_t + 0.8\ Y_{t-1}, \tag{2.16}$$

$$X_t = 0.85\ X_{t-1} + Y_t + 0.3\ Y_{t-1}, \tag{2.17}$$

$$X_t = 0.85\ X_{t-1} + Y_t - 0.3\ Y_{t-1}, \tag{2.18}$$

$$X_t = 0.3\ X_{t-1} + Y_t + 0.8\ Y_{t-1}, \tag{2.19}$$

where $Y_t \sim T(4)$ distribution, with $\gamma = 0.25$ and $\rho = -0.5$.

2. ARCH(1) model:

$$X_t = \sigma_t Z_t, \quad \sigma_t^2 = 0.1 + 105^{-\frac{1}{4}}\ X_t^2, \quad Z_t \sim N(0, 1). \tag{2.20}$$

For finite-order casual ARMA(p, q) models with heavy-tailed innovations, the observations and innovations share the same extreme value index γ. Hence, in Models (2.16)–(2.19), the true γ for X_t is 0.25. Notice that for ARMA(1, 1) model, the decaying rate of dependence is mainly determined by the AR part. Therefore, relatively slower decaying rates of dependence exist in Models (2.16)–(2.18), whereas Model (2.19) has a very short-memory but strong local dependence because of the large MA parameter 0.8. Moreover, as the MA parameters decline in the first three models, the degrees of dependence also decrease to a certain extent.

On the other hand, for ARCH(1) model

$$X_t = \sigma_t Z_t, \quad \sigma_t^2 = \alpha_0 + \alpha_1\ X_t^2, \quad Z_t \sim N(0, 1),$$

where $\alpha_0 > 0, \alpha_1 \in (0, 1)$, there exists a constant $\theta > 0$ such that $\alpha_1^\theta E(Z_t^{2\theta}) = 1$. With this constant θ, X_t is heavy-tailed with EVI $\gamma = \theta^{-1}$. In Model (2.20), we choose $\alpha_1 = 105^{-\frac{1}{4}} \approx 0.3124$ so that the corresponding EVI for X_t is 0.25.

In Figs. 2.8, 2.9, 2.10, 2.11 and 2.12, we show the simulation results under Models (2.16)–(2.20) respectively, based on $J = 1000$ runs. For each setting, we choose sample size $n = 1000$ and threshold $k_n = 10, 11, \ldots, 300$. Red lines and blue lines relate to the EMP estimator and the MOP estimator with different p, respectively. Green, cyan, and gold lines correspond to the Mom, GH, and MM estimators, respectively. For these five models, the EMP estimator has much lower RMSE compared with the MOP estimator for moderately large choices of the thresholds k_n regardless of the tuning parameter p, whereas the MOP estimator can slightly outperform the EMP estimator when k_n is small. Roughly speaking, when $k_n = 10, 11, \ldots, 100$, the RMSE of the MOP estimator is 0–0.04 lower than that of the EMP estimator, whereas the EMP estimator can outperform MOP estimators by 0–0.5 regarding RMSE for $k_n = 101, 102, \ldots, 300$. Further, the RMSE and biases of the EMP estimator with $p = 0.4\gamma^{-1}$ are no more than 0.17 and 0.12 respectively, which demonstrates the overall effectiveness of the EMP estimator for dependent observations within the whole range where $k_n = 100, 101, \ldots, 300$. Regarding the Mom, MM, and GH estimators, they seriously underestimate γ especially when k_n is smaller than 50 and even give rise to negative values. In these five models, the EMP estimator with $p = 0.4\gamma^{-1}$ has lower RMSE than these three types of estimators for nearly all choices of thresholds k_n, except when $k_n \geq 293$ in Model (2.18) and $k_n \geq 244$ in Model (2.19).

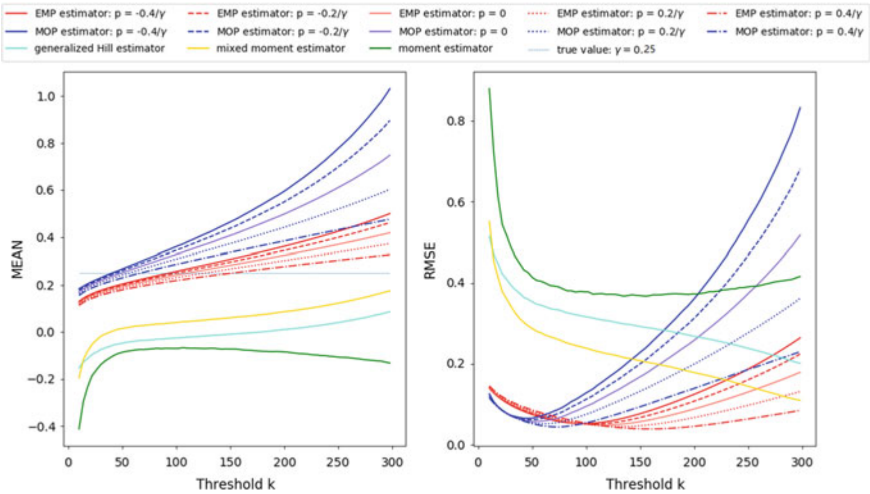


Fig. 2.8 Mean (left) and RMSE (right) of EVI estimators for Model (2.16), $\gamma = 0.25$

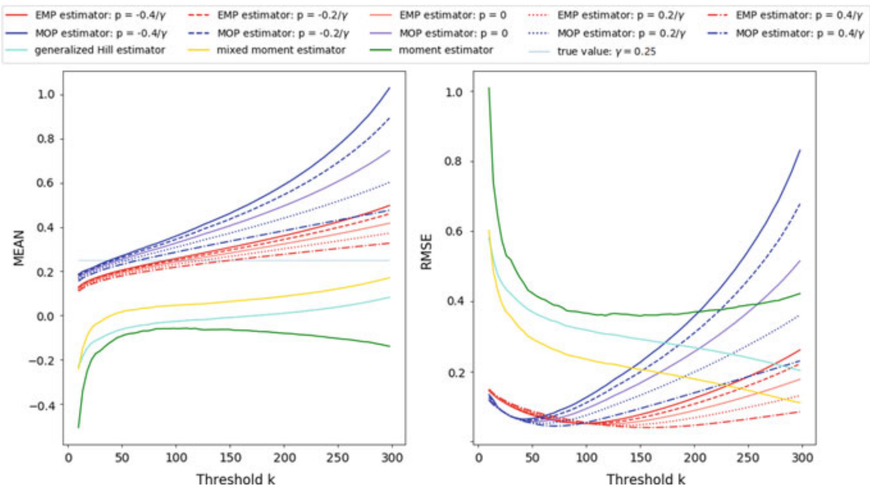


Fig. 2.9 Mean (left) and RMSE (right) of EVI estimators for Model (2.17), $\gamma = 0.25$

2.4.3 Simulation Results at Optimal Threshold

The performance of the EVI estimators at the simulated optimal thresholds using RMSE is studied in this section. Though this is not relevant in practice, it can provide information of the potential efficiencies of the EVI estimators. To measure the performance, one common indicator is the relative efficiency (REFF), which is defined as

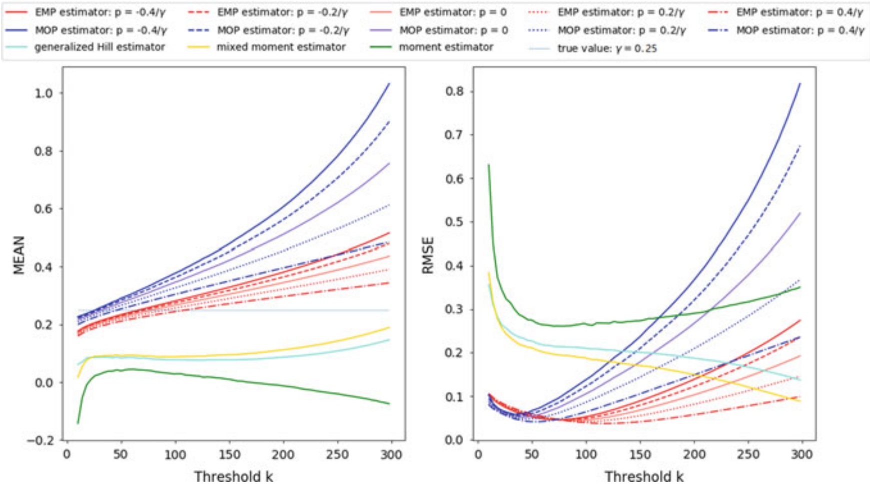


Fig. 2.10 Mean (left) and RMSE (right) of EVI estimators for Model (2.18), $\gamma = 0.25$

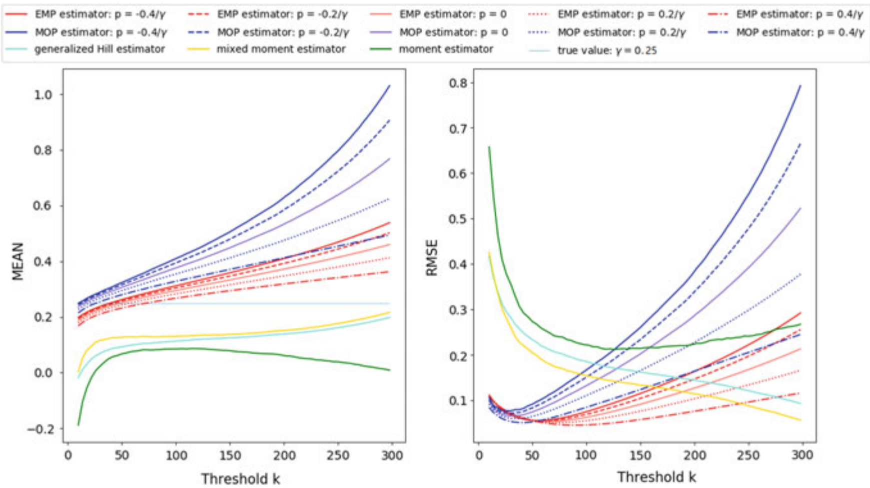


Fig. 2.11 Mean (left) and RMSE (right) of EVI estimators for Model (2.19), $\gamma = 0.25$

$$\text{REFF}(\hat{\gamma}^{\text{target}} \mid \hat{\gamma}_0) := \frac{\text{RMSE}(\hat{\gamma}_0 \mid k_0^*)}{\text{RMSE}(\hat{\gamma}^{\text{target}} \mid k_{\text{target}}^*)},$$

where $\hat{\gamma}^{\text{target}}$ is the estimator of interest, $\hat{\gamma}_0$ is another estimator used as a benchmark, and k_0^* as well as k_{target}^* represent the optimal thresholds for these two estimators respectively. Notice that the larger REFF is, the better the associated estimator performs. It is common to adopt the Hill estimator as the benchmark. Hereinafter, we will follow this practice, namely,

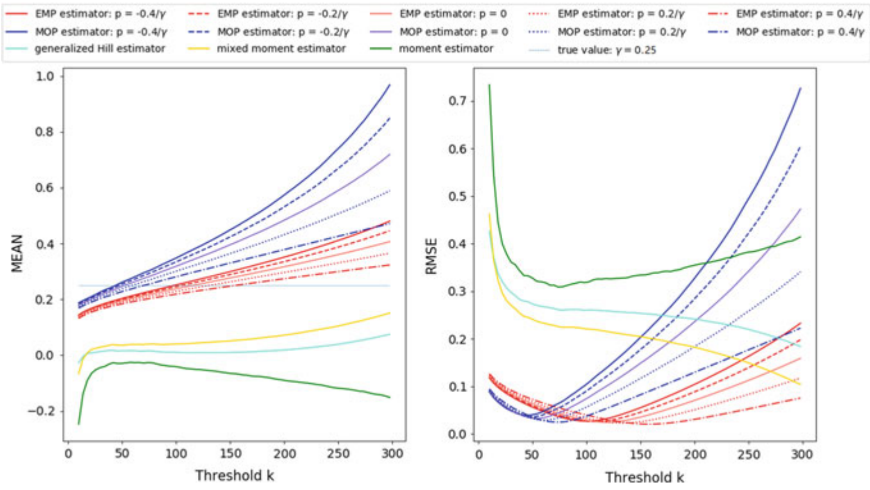


Fig. 2.12 Mean (left) and RMSE (right) of EVI estimators for Model (2.20), $\gamma = 0.25$

$$\mathrm{REFF}(\hat{\gamma}^{\text{target}}) := \frac{\mathrm{RMSE}(\hat{\gamma}^{\text{Hill}} \mid k^*_{\text{Hill}})}{\mathrm{RMSE}(\hat{\gamma}^{\text{target}} \mid k^*_{\text{target}})}.$$

On the other hand, the distributional properties of the indicators of performance, such as mean values and REFF, are not easy to estimate in general. To tackle this problem, multi-sample Monte Carlo simulations are often conducted; see Gomes and Oliveira [17], Gomes and Pestana [18] and Brilhante et al. [5] to name but a few. Suppose we are concerned about the parameter θ, like REFF. Then, in $r \times m$ multi-sample simulations, we obtain r independent observations of θ, denoted as $\hat{\theta}_1, \ \hat{\theta}_2, \ \ldots, \ \hat{\theta}_r$, with each computed based on m runs. Under mild conditions, we have

$$\frac{\bar{\hat{\theta}} - \theta}{\hat{\sigma}_\theta / \sqrt{r}} \approx_d N(0, 1),$$

for r large enough, where

$$\bar{\hat{\theta}} := \frac{1}{r} \sum_{i=1}^{r} \hat{\theta}_i \quad \text{and} \quad \hat{\sigma}_\theta := \sqrt{\frac{1}{r-1} \sum_{i=1}^{r} \left(\hat{\theta}_i - \bar{\hat{\theta}} \right)^2}.$$

Hence, we can construct the confidence interval for θ as

$$\left(\bar{\hat{\theta}} - z_{1-\frac{\alpha}{2}} \hat{\sigma}_\theta / \sqrt{r}, \bar{\hat{\theta}} + z_{1-\frac{\alpha}{2}} \hat{\sigma}_\theta / \sqrt{r} \right).$$

Table 2.2 Simulated mean values and associated confidence intervals at optimal thresholds for Model (2.16), $\gamma = 0.25$

	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
EMP: $p\gamma = -0.4$	0.2599 ± 0.00215	0.2676 ± 0.00173	0.2613 ± 0.00112	0.2489 ± 0.00087	0.2476 ± 0.00070
MOP: $p\gamma = -0.4$	0.2785 ± 0.00312	0.2912 ± 0.00212	0.2625 ± 0.00149	0.2481 ± 0.00111	0.2449 ± 0.00078
EMP: $p\gamma = -0.2$	0.2585 ± 0.00206	0.2658 ± 0.00168	0.2603 ± 0.00108	0.2495 ± 0.00086	0.2477 ± 0.00070
MOP: $p\gamma = -0.2$	0.2730 ± 0.00307	0.2813 ± 0.00202	0.2738 ± 0.00139	0.2447 ± 0.00098	0.2473 ± 0.00070
EMP: $p = 0$	0.2573 ± 0.00192	0.2663 ± 0.00158	0.2603 ± 0.00103	0.2494 ± 0.00084	0.2479 ± 0.00069
MOP: $p = 0$	0.2583 ± 0.00263	0.2698 ± 0.00188	0.2672 ± 0.00125	0.2536 ± 0.00093	0.2465 ± 0.00070
EMP: $p\gamma = 0.2$	0.2525 ± 0.00180	0.2627 ± 0.00146	0.2597 ± 0.00098	0.2494 ± 0.00081	0.2482 ± 0.00067
MOP: $p\gamma = 0.2$	0.2545 ± 0.00241	0.2729 ± 0.00152	0.2656 ± 0.00115	0.2477 ± 0.00092	0.2494 ± 0.00061
EMP: $p\gamma = 0.4$	0.2564 ± 0.00162	0.2641 ± 0.00128	0.2591 ± 0.00095	0.2501 ± 0.00076	0.2489 ± 0.00062
MOP: $p\gamma = 0.4$	0.2653 ± 0.00156	0.2630 ± 0.00132	0.2610 ± 0.00100	0.2473 ± 0.00083	0.2490 ± 0.00064

For more details, we refer the readers to Gomes and Oliveira [17]. We investigate the potential efficiencies of the MOP and EMP estimators at the simulated optimal thresholds in Models (2.16)–(2.20), with sample size $n = 200, 500, 1000, 5000$ and tuning parameter $p = -0.4\gamma^{-1}, -0.2\gamma^{-1}, 0, 0.2\gamma^{-1}$ and $0.4\gamma^{-1}$, respectively. The 20×5000 multi-sample Monte Carlo simulations are carried out to obtain simulated mean values, RMSE, REFF as well as their associated 95% confidence intervals at simulated optimal thresholds. The results are reported in Tables 2.2, 2.3, 2.4, 2.5, 2.6, 2.7, 2.8, 2.9, 2.10 and 2.11.

We observe that the EMP estimator has lower biases and higher REFF than MOP estimator for nearly all settings. The EMP estimator can mostly outperform the Hill estimator at the optimal thresholds in the sense of lower RMSE, except the case with $p = -0.4\gamma^{-1}$ and $n = 5000$ under Model (2.17). Regarding the advantage of the EMP estimator over the MOP estimator, it is more significant as the sample size or the tuning parameter decreases. For instance, in Model (2.19) with sample size $n = 200$, the REFF for the EMP and the MOP estimators are 1.7333 and 1.5402 when $p = 0.4\gamma^{-1}$, whereas theses pairs of values become 1.2341 and 0.7872 when p decreases to $-0.4\gamma^{-1}$. Finally, for the ARCH(1) model, the EMP estimator can outperform the MOP estimator by 8%–50% in the sense of the REFF under various thresholds and tuning parameters.

Table 2.3 Simulated mean values and associated confidence intervals at optimal thresholds for Model (2.17), $\gamma = 0.25$

	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
EMP: $p\gamma = -0.4$	0.2658 ± 0.00196	0.2629 ± 0.00118	0.2621 ± 0.00145	0.2524 ± 0.00108	0.2492 ± 0.00089
MOP: $p\gamma = -0.4$	0.2956 ± 0.00220	0.2823 ± 0.00219	0.2665 ± 0.00169	0.2496 ± 0.00109	0.2521 ± 0.00107
EMP: $p\gamma = -0.2$	0.2640 ± 0.00189	0.2631 ± 0.00114	0.2620 ± 0.00143	0.2520 ± 0.00107	0.2494 ± 0.00088
MOP: $p\gamma = -0.2$	0.2820 ± 0.00199	0.2730 ± 0.00197	0.2701 ± 0.00157	0.2515 ± 0.00111	0.2487 ± 0.00102
EMP: $p = 0$	0.2622 ± 0.00175	0.2625 ± 0.00111	0.2611 ± 0.00139	0.2517 ± 0.00105	0.2505 ± 0.00085
MOP: $p = 0$	0.2668 ± 0.00177	0.2675 ± 0.00151	0.2640 ± 0.00148	0.2528 ± 0.00106	0.2506 ± 0.00099
EMP: $p\gamma = 0.2$	0.2632 ± 0.00156	0.2620 ± 0.00104	0.2605 ± 0.00129	0.2523 ± 0.00099	0.2498 ± 0.00083
MOP: $p\gamma = 0.2$	0.2618 ± 0.00153	0.2687 ± 0.00118	0.2688 ± 0.00129	0.2519 ± 0.00111	0.2508 ± 0.00087
EMP: $p\gamma = 0.4$	0.2575 ± 0.00132	0.2616 ± 0.00083	0.2598 ± 0.00115	0.2523 ± 0.00090	0.2501 ± 0.00078
MOP: $p\gamma = 0.4$	0.2709 ± 0.00139	0.2635 ± 0.00085	0.2598 ± 0.00117	0.2527 ± 0.00097	0.2505 ± 0.00077

2.5 Conclusion

We have proposed a new type of EVI estimator—EMP estimator. The asymptotic normalities of the MOP and the EMP estimators have been established for dependent observations under some mild conditions. We also have developed consistent estimators for the asymptotic variances of the MOP and the EMP estimators. Furthermore, the asymptotic normality of the extreme quantile estimator has been established for dependent observations, based on which we can construct the confidence interval for the extreme quantile. Generally speaking, the proposed EMP estimator demonstrates the following advantages: In a theoretical way, it outperforms the MOP and the Hill estimators at the optimal thresholds in terms of the LMSE. For finite samples, our numerical results reveal that the EMP estimator typically yields lower RMSE than other classical estimators at optimal thresholds. It can give rise to more satisfactory finite sample performances under a wide range of choices of thresholds, especially in the case that $\gamma \in (0, 1/2]$. Last, but not least, our proposal is more robust with respect to the threshold k_n. Compared with the MOP estimator, the new EMP estimator is also more robust with respect to the choice of the tuning parameter p.

Table 2.4 Simulated mean values and associated confidence intervals at optimal thresholds for Model (2.18), $\gamma = 0.25$

	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
EMP: $p\gamma = -0.4$	0.2457 ± 0.00217	0.2554 ± 0.00158	0.2587 ± 0.00139	0.2546 ± 0.00078	0.2522 ± 0.00064
MOP: $p\gamma = -0.4$	0.2716 ± 0.00331	0.2541 ± 0.00228	0.2637 ± 0.00161	0.2590 ± 0.00087	0.2559 ± 0.00077
EMP: $p\gamma = -0.2$	0.2597 ± 0.00220	0.2587 ± 0.00155	0.2590 ± 0.00136	0.2544 ± 0.00078	0.2524 ± 0.00064
MOP: $p\gamma = -0.2$	0.2616 ± 0.00296	0.2590 ± 0.00160	0.2671 ± 0.00157	0.2586 ± 0.00083	0.2525 ± 0.00074
EMP: $p = 0$	0.2545 ± 0.00209	0.2605 ± 0.00149	0.2590 ± 0.00131	0.2547 ± 0.00077	0.2525 ± 0.00064
MOP: $p = 0$	0.2507 ± 0.00262	0.2720 ± 0.00183	0.2708 ± 0.00151	0.2536 ± 0.00081	0.2533 ± 0.00074
EMP: $p\gamma = 0.2$	0.2594 ± 0.00197	0.2616 ± 0.00142	0.2592 ± 0.00124	0.2546 ± 0.00073	0.2526 ± 0.00061
MOP: $p\gamma = 0.2$	0.2560 ± 0.00218	0.2635 ± 0.00170	0.2620 ± 0.00141	0.2541 ± 0.00070	0.2553 ± 0.00059
EMP: $p\gamma = 0.4$	0.2598 ± 0.00179	0.2597 ± 0.00130	0.2595 ± 0.00114	0.2543 ± 0.00066	0.2528 ± 0.00055
MOP: $p\gamma = 0.4$	0.2594 ± 0.00183	0.2580 ± 0.00137	0.2596 ± 0.00118	0.2575 ± 0.00063	0.2535 ± 0.00051

2.6 Technical Details

This section includes all the technical proofs of the results in Sects. 2.2 and 2.3.

2.6.1 Details on Existence of $e_p(t)$

Theorem 2.7 *Suppose that $X_1, X_2, \ldots, X_n$ are i.i.d. continuous random variables with the EVI, $\gamma > 0$. If $p\gamma < 1$, then $e_p(t)$ is regularly varying with index $p\gamma$. In other words, as $t \to \infty$, we have*

$$\log(e_p(t)) = \log\left(E\left(X^p \mid X > U(t)\right)\right) = p\gamma \log(t) + o(\log(t)). \quad (2.21)$$

Proof Note that

$$1 - F(U(t)) = 1 - F\left(F^{-1}\left(1 - \frac{1}{t}\right)\right) = \frac{1}{t}.$$

Table 2.5 Simulated mean values and associated confidence intervals at optimal thresholds for Model (2.19), $\gamma = 0.25$

	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
EMP: $p\gamma = -0.4$	0.2807 ± 0.00257	0.2716 ± 0.00197	0.2692 ± 0.00132	0.2604 ± 0.00130	0.2604 ± 0.00063
MOP: $p\gamma = -0.4$	0.3295 ± 0.00348	0.2945 ± 0.00264	0.2863 ± 0.00190	0.2754 ± 0.00163	0.2652 ± 0.00059
EMP: $p\gamma = -0.2$	0.2791 ± 0.00251	0.2725 ± 0.00191	0.2682 ± 0.00130	0.2634 ± 0.00125	0.2606 ± 0.00065
MOP: $p\gamma = -0.2$	0.3169 ± 0.00317	0.2872 ± 0.00252	0.2810 ± 0.00182	0.2716 ± 0.00155	0.2626 ± 0.00060
EMP: $p = 0$	0.2780 ± 0.00238	0.2678 ± 0.00188	0.2676 ± 0.00126	0.2642 ± 0.00121	0.2612 ± 0.00066
MOP: $p = 0$	0.3089 ± 0.00288	0.2785 ± 0.00237	0.2747 ± 0.00171	0.2713 ± 0.00131	0.2598 ± 0.00066
EMP: $p\gamma = 0.2$	0.2761 ± 0.00212	0.2716 ± 0.00169	0.2672 ± 0.00119	0.2654 ± 0.00112	0.2610 ± 0.00068
MOP: $p\gamma = 0.2$	0.2908 ± 0.00248	0.2779 ± 0.00191	0.2705 ± 0.00151	0.2677 ± 0.00123	0.2613 ± 0.00066
EMP: $p\gamma = 0.4$	0.2646 ± 0.00183	0.2675 ± 0.00140	0.2648 ± 0.00106	0.2614 ± 0.00103	0.2604 ± 0.00069
MOP: $p\gamma = 0.4$	0.2744 ± 0.00204	0.2692 ± 0.00150	0.2722 ± 0.00107	0.2626 ± 0.00104	0.2604 ± 0.00069

Through a transformation $s := U^{-1}(x)$, we have

$$\begin{aligned} e_p(t) &= E\left(X^p \mid X > U(t)\right) \\ &= \frac{1}{1 - F(U(t))} \int_{U(t)}^{+\infty} x^p \, dF(x) \\ &= t \int_t^{+\infty} U^p(s) \, d\{F(U(s))\} \\ &= -t \int_t^{+\infty} U^p(s) \, d\{1 - F(U(s))\} \\ &= t \int_t^{+\infty} U^p(s) s^{-2} \, ds. \end{aligned} \tag{2.22}$$

Because $U(s) \in RV_\gamma$, $U^p(s)s^{-2} \in RV_{p\gamma-2}$. Since $p\gamma - 2 < -1$, based on De Haan and Ferreira ([8], Proposition B.1.9(4)), we know that

$$\int_t^{+\infty} U^p(s)s^{-2} \, ds$$

Table 2.6 Simulated mean values and associated confidence intervals at optimal thresholds for Model (2.20), $\gamma = 0.25$

	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
EMP: $p\gamma = -0.4$	0.2575 ± 0.00137	0.2563 ± 0.00103	0.2547 ± 0.00066	0.2494 ± 0.00052	0.2492 ± 0.00034
MOP: $p\gamma = -0.4$	0.2741 ± 0.00248	0.2571 ± 0.00189	0.2489 ± 0.00092	0.2508 ± 0.00064	0.2488 ± 0.00035
EMP: $p\gamma = -0.2$	0.2580 ± 0.00128	0.2564 ± 0.00098	0.2545 ± 0.00064	0.2497 ± 0.00051	0.2490 ± 0.00034
MOP: $p\gamma = -0.2$	0.2714 ± 0.00249	0.2593 ± 0.00157	0.2572 ± 0.00095	0.2492 ± 0.00067	0.2490 ± 0.00035
EMP: $p = 0$	0.2525 ± 0.00127	0.2574 ± 0.00093	0.2537 ± 0.00061	0.2497 ± 0.00050	0.2493 ± 0.00033
MOP: $p = 0$	0.2596 ± 0.00223	0.2506 ± 0.00140	0.2537 ± 0.00092	0.2504 ± 0.00057	0.2478 ± 0.00038
EMP: $p\gamma = 0.2$	0.2586 ± 0.00117	0.2580 ± 0.00090	0.2546 ± 0.00056	0.2496 ± 0.00048	0.2493 ± 0.00032
MOP: $p\gamma = 0.2$	0.2651 ± 0.00152	0.2585 ± 0.00103	0.2520 ± 0.00076	0.2495 ± 0.00051	0.2508 ± 0.00037
EMP: $p\gamma = 0.4$	0.2578 ± 0.00096	0.2552 ± 0.00083	0.2541 ± 0.00052	0.2496 ± 0.00042	0.2496 ± 0.00030
MOP: $p\gamma = 0.4$	0.2586 ± 0.00102	0.2594 ± 0.00096	0.2571 ± 0.00056	0.2514 ± 0.00046	0.2488 ± 0.00033

exists for t large enough and is regularly varying with index $(p\gamma - 1)$. Hence, based on (2.22), we can conclude that $e_p(t) \in RV_{p\gamma}$. □

2.6.2 *Proof of Theorem 2.1*

Hereinafter, we use the notation $o_p^{(n)}(f(n))$ to denote a term $g(n)$ such that $\forall\, \epsilon > 0$,

$$\lim_{n\to\infty} P\left(\left|\frac{g(n)}{f(n)}\right| < \epsilon\right) = 1.$$

Also, use $o_p^{(t)}(f(t))$ to denote a term $g(t)$ such that $\forall\, \epsilon > 0$,

$$\lim_{t\downarrow 0} P\left(\left|\frac{g(t)}{f(t)}\right| < \epsilon\right) = 1.$$

Define $U_i := 1 - 1\big/U^{-1}(X_i)$. Then,

Table 2.7 Simulated REFF and RMSE$(\hat{\gamma}^{\text{Hill}})$, together with confidence intervals for Model (2.16), $\gamma = 0.25$

	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
Hill (RMSE)	0.0988 ± 0.00276	0.0734 ± 0.00237	0.0559 ± 0.00111	0.0412 ± 0.00096	0.0281 ± 0.00062
EMP: $p\gamma = -0.4$	1.0967 ± 0.01883	1.0723 ± 0.01341	1.0639 ± 0.01453	1.0170 ± 0.00975	1.0167 ± 0.00677
MOP: $p\gamma = -0.4$	0.7654 ± 0.01582	0.8065 ± 0.01005	0.8934 ± 0.01154	0.8939 ± 0.01375	0.9206 ± 0.01056
EMP: $p\gamma = -0.2$	1.1379 ± 0.01987	1.1061 ± 0.01344	1.0826 ± 0.01757	1.0295 ± 0.01140	1.0207 ± 0.00811
MOP: $p\gamma = -0.2$	0.8636 ± 0.00734	0.8915 ± 0.00606	0.9159 ± 0.00759	0.9150 ± 0.01294	0.9731 ± 0.00846
EMP: $p = 0$	1.2250 ± 0.02200	1.1492 ± 0.01359	1.1162 ± 0.02114	1.0488 ± 0.01455	1.0280 ± 0.01161
MOP: $p = 0$	1.0	1.0	1.0	1.0	1.0
EMP: $p\gamma = 0.2$	1.3579 ± 0.02409	1.2562 ± 0.01515	1.1882 ± 0.02376	1.0925 ± 0.01899	1.0485 ± 0.02071
MOP: $p\gamma = 0.2$	1.1858 ± 0.01934	1.1161 ± 0.01200	1.0976 ± 0.01201	1.0376 ± 0.01214	1.0545 ± 0.01561
EMP: $p\gamma = 0.4$	1.5769 ± 0.03452	1.4404 ± 0.02496	1.3449 ± 0.02490	1.2067 ± 0.02004	1.1145 ± 0.02788
MOP: $p\gamma = 0.4$	1.4630 ± 0.03299	1.3843 ± 0.02406	1.2944 ± 0.02284	1.1546 ± 0.02016	1.1022 ± 0.02763

$$P(U_i \le u) = P\left(U^{-1}(x_i) \le \frac{1}{1-u}\right) = P\left(X_i \le U\left(\frac{1}{1-u}\right)\right) = P(X_i \le F^{-1}(u)) = u,$$

for $u \in [0, 1]$. So $U_i \sim \text{Uniform}(0, 1)$. Denoting $U_{1,n} \le U_{2,n} \le \cdots \le U_{n,n}$ as the order statistics, we have

$$X_{k,n} = U\left(\frac{1}{1 - U_{k,n}}\right).$$

Under the second-order condition, we have

$$\begin{aligned}\frac{Q_n(t)}{U(\frac{n}{k_n})} &= \frac{U\left(\frac{1}{1-U_{n-[k_n t],n}}\right)}{U(\frac{n}{k_n})} \\ &= \left(\frac{n}{k_n}\left(1 - U_{n-[k_n t],n}\right)\right)^{-\gamma} \left\{1 + A\left(\frac{n}{k_n}\right) \Psi\left(\frac{k_n}{n\left(1 - U_{n-[k_n t],n}\right)}\right)\left(1 + o_p^{(n)}(1)\right)\right\}.\end{aligned}$$

Denote

$$V_n(t) := \frac{n}{k_n}\left(1 - U_{n-[k_n t],n}\right),$$

Table 2.8 Simulated REFF and RMSE$(\hat{\gamma}^{\text{Hill}})$, together with confidence intervals for Model (2.17), $\gamma = 0.25$

	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
Hill (RMSE)	0.0995 ± 0.00214	0.0725 ± 0.00227	0.0553 ± 0.00150	0.0408 ± 0.00097	0.0270 ± 0.00062
EMP: $p\gamma = -0.4$	1.0927 ± 0.01258	1.0749 ± 0.01107	1.0493 ± 0.01191	1.0311 ± 0.00789	0.9978 ± 0.00819
MOP: $p\gamma = -0.4$	0.7574 ± 0.01074	0.8317 ± 0.01573	0.8830 ± 0.01706	0.8972 ± 0.01377	0.9279 ± 0.01245
EMP: $p\gamma = -0.2$	1.1353 ± 0.01262	1.1042 ± 0.01419	1.0661 ± 0.01260	1.0400 ± 0.00975	1.0019 ± 0.00939
MOP: $p\gamma = -0.2$	0.8589 ± 0.00621	0.9099 ± 0.00917	0.9240 ± 0.00945	0.9464 ± 0.01131	0.9526 ± 0.00948
EMP: $p = 0$	1.2243 ± 0.01362	1.1499 ± 0.01840	1.0998 ± 0.01413	1.0590 ± 0.01310	1.0155 ± 0.01302
MOP: $p = 0$	1.0	1.0	1.0	1.0	1.0
EMP: $p\gamma = 0.2$	1.3438 ± 0.01396	1.2493 ± 0.01881	1.1699 ± 0.01480	1.1068 ± 0.01709	1.0295 ± 0.02386
MOP: $p\gamma = 0.2$	1.1896 ± 0.00862	1.1244 ± 0.01262	1.0733 ± 0.00983	1.0633 ± 0.01210	1.0332 ± 0.01740
EMP: $p\gamma = 0.4$	1.5929 ± 0.02002	1.4428 ± 0.02264	1.3261 ± 0.01690	1.2216 ± 0.01729	1.0923 ± 0.02805
MOP: $p\gamma = 0.4$	1.4559 ± 0.02353	1.3733 ± 0.02520	1.2814 ± 0.01524	1.1932 ± 0.01676	1.0827 ± 0.02829

and rewrite $Q_n(t)\big/U(\frac{n}{k_n})$ as

$$\frac{Q_n(t)}{U(\frac{n}{k_n})} = (V_n(t))^{-\gamma} + (V_n(t))^{-\gamma}\Psi\left(\frac{1}{V_n(t)}\right)A\left(\frac{n}{k_n}\right)\left(1 + o_p^{(n)}(1)\right).$$

Denote

$$q(t) := t^{\frac{1}{2}}(1 + |\log t|)^{\frac{1}{2}}.$$

By (2.5), we know that for $(2k_n)^{-1} \le t \le 1$ uniformly,

$$\{V_n(t)\}^{-\gamma} = t^{-\gamma} + k_n^{-\frac{1}{2}}\gamma\left\{t^{-(1+\gamma)}e(t) + o_p^{(n)}(1)t^{-\gamma-1}q(t)\right\}. \tag{2.23}$$

Similarly, we have

$$\{V_n(t)\}^{-\rho} = t^{-\rho} + k_n^{-\frac{1}{2}}\rho\left\{t^{-(1+\rho)}e(t) + o_p^{(n)}(1)t^{-\rho-1}q(t)\right\}.$$

Hence,

Table 2.9 Simulated REFF and RMSE$(\hat{\gamma}^{\text{Hill}})$, together with confidence intervals for Model (2.18), $\gamma = 0.25$

	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
Hill (RMSE)	0.0871 ± 0.00337	0.0649 ± 0.00209	0.0506 ± 0.00182	0.0372 ± 0.00088	0.0246 ± 0.00074
EMP: $p\gamma = -0.4$	1.1834 ± 0.02806	1.1500 ± 0.01247	1.1290 ± 0.01502	1.0798 ± 0.00898	1.0357 ± 0.01071
MOP: $p\gamma = -0.4$	0.7807 ± 0.01829	0.8094 ± 0.02727	0.9068 ± 0.01642	0.9235 ± 0.01232	0.9333 ± 0.01141
EMP: $p\gamma = -0.2$	1.2138 ± 0.02721	1.1583 ± 0.01304	1.1337 ± 0.01723	1.0798 ± 0.00918	1.0290 ± 0.01155
MOP: $p\gamma = -0.2$	0.8803 ± 0.01067	0.9111 ± 0.01736	0.9574 ± 0.01034	0.9666 ± 0.00852	0.9682 ± 0.00918
EMP: $p = 0$	1.2836 ± 0.02522	1.1940 ± 0.01465	1.1515 ± 0.02045	1.0852 ± 0.01118	1.0209 ± 0.01396
MOP: $p = 0$	1.0	1.0	1.0	1.0	1.0
EMP: $p\gamma = 0.2$	1.3857 ± 0.02642	1.2683 ± 0.01547	1.1979 ± 0.02149	1.1050 ± 0.01470	1.0141 ± 0.02134
MOP: $p\gamma = 0.2$	1.1851 ± 0.02202	1.1343 ± 0.00916	1.1042 ± 0.01077	1.0496 ± 0.01167	1.0119 ± 0.01631
EMP: $p\gamma = 0.4$	1.5934 ± 0.04052	1.4387 ± 0.02347	1.3188 ± 0.01883	1.1805 ± 0.01614	1.0320 ± 0.02809
MOP: $p\gamma = 0.4$	1.4680 ± 0.03588	1.3567 ± 0.02497	1.2642 ± 0.01731	1.1475 ± 0.01571	1.0195 ± 0.02833

$$\frac{\{V_n(t)\}^{-\rho} - 1}{\rho} = \frac{t^{-\rho} - 1}{\rho} + k_n^{-\frac{1}{2}} \left\{t^{-(1+\rho)} e(t) + o_p^{(n)}(1) t^{-\rho-1} q(t)\right\}. \quad (2.24)$$

Letting $\rho \uparrow 0$ in (2.24), we obtain

$$\log\left(V_n(t)\right) = \log t + k_n^{-\frac{1}{2}} \left\{t^{-1} e(t) + o_p^{(n)}(1) t^{-1} q(t)\right\}. \quad (2.25)$$

So, combining (2.6) and (2.23)–(2.25), we have

$$\{V_n(t)\}^{-\gamma} \Psi\left(\frac{1}{V_n(t)}\right) = t^{-\gamma} \Psi_0(t) + o_p^{(n)}(1) t^{-\gamma-1} q(t)$$

uniformly for $(2k_n)^{-1} \leq t \leq 1$. Hence,

$$\frac{Q_n(t)}{U(\frac{n}{k_n})} = t^{-\gamma} + \gamma k_n^{-\frac{1}{2}} \left\{t^{-(1+\gamma)} e(t) + o_p^{(n)}(1) t^{-\gamma-1} q(t)\right\}$$
$$+ A\left(\frac{n}{k_n}\right) \left\{t^{-\gamma} \Psi_0(t) + o_p^{(n)}(1) t^{-\gamma-1} q(t)\right\}.$$

Table 2.10 Simulated REFF and RMSE$(\hat{\gamma}^{\text{Hill}})$, together with confidence intervals for Model (2.19), $\gamma = 0.25$

	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
Hill (RMSE)	0.1148 ± 0.00341	0.0827 ± 0.00194	0.0634 ± 0.00179	0.0485 ± 0.00111	0.0342 ± 0.00080
EMP: $p\gamma = -0.4$	1.2341 ± 0.01545	1.1957 ± 0.01747	1.1533 ± 0.01194	1.1142 ± 0.01065	1.1036 ± 0.01477
MOP: $p\gamma = -0.4$	0.7872 ± 0.01167	0.8462 ± 0.01005	0.8675 ± 0.00610	0.8835 ± 0.00982	0.9165 ± 0.00722
EMP: $p\gamma = -0.2$	1.2487 ± 0.01659	1.2103 ± 0.01675	1.1616 ± 0.01092	1.1224 ± 0.00988	1.0991 ± 0.01551
MOP: $p\gamma = -0.2$	0.8837 ± 0.01078	0.9186 ± 0.00625	0.9315 ± 0.00405	0.9376 ± 0.00802	0.9621 ± 0.00543
EMP: $p = 0$	1.3070 ± 0.01852	1.2379 ± 0.01671	1.1765 ± 0.01032	1.1256 ± 0.01077	1.0957 ± 0.01672
MOP: $p = 0$	1.0	1.0	1.0	1.0	1.0
EMP: $p\gamma = 0.2$	1.4335 ± 0.02290	1.3058 ± 0.01624	1.2129 ± 0.01344	1.1408 ± 0.01594	1.0938 ± 0.02108
MOP: $p\gamma = 0.2$	1.1921 ± 0.00943	1.1368 ± 0.01344	1.0890 ± 0.01029	1.0585 ± 0.01372	1.0475 ± 0.01511
EMP: $p\gamma = 0.4$	1.7333 ± 0.03230	1.4956 ± 0.01897	1.3438 ± 0.01947	1.2147 ± 0.02332	1.1194 ± 0.02810
MOP: $p\gamma = 0.4$	1.5402 ± 0.02629	1.3782 ± 0.01896	1.2586 ± 0.01890	1.1591 ± 0.02142	1.0885 ± 0.02662

Combining with the assumption $k_n^{\frac{1}{2}} A(n/k_n) \to \phi \in (-\infty, +\infty)$, we have

$$\sup_{t\in[\frac{1}{2k_n},1]} \frac{t^{\gamma+1}}{q(t)} \left| k_n^{\frac{1}{2}} \left(\frac{Q_n(t)}{U(n/k_n)} - t^{-\gamma} \right) - \gamma t^{-(\gamma+1)} e(t) - \phi t^{-\gamma} \Psi_0(t) \right| \to 0 \quad \text{in probability.}$$

If $t \in (0, (2k_n)^{-1}]$, $Q_n(t) = Q_n((2k_n)^{-1})$. Hence,

$$\begin{aligned}
&\sup_{0<t\le\frac{1}{2k_n}} \frac{t^{\gamma+1}}{q(t)} \left| k_n^{\frac{1}{2}} \left(\frac{Q_n(t)}{U(\frac{n}{k_n})} - t^{-\gamma} \right) - \gamma t^{-(\gamma+1)} e(t) - \phi t^{-\gamma} \Psi_0(t) \right| \\
&\le \sup_{0<t\le\frac{1}{2k_n}} \frac{t^{\gamma+1}}{q(t)} \left| k_n^{\frac{1}{2}} \left(\frac{Q_n\left(\frac{1}{2k_n}\right)}{U(\frac{n}{k_n})} - \left(\frac{1}{2k_n}\right)^{-\gamma} \right) - \gamma \left(\frac{1}{2k_n}\right)^{-(\gamma+1)} e\left(\frac{1}{2k_n}\right) \right. \\
&\left. -\phi \left(\frac{1}{2k_n}\right)^{-\gamma} \Psi_0\left(\frac{1}{2k_n}\right) \right| \\
&+ \sup_{0<t\le\frac{1}{2k_n}} \frac{t^{\gamma+1}}{q(t)} \left| k_n^{\frac{1}{2}} \left\{ t^{-\gamma} - (2k_n)^{\gamma} \right\} + \gamma \left\{ t^{-(1+\gamma)} e(t) - (2k_n)^{\gamma+1} e\left(\frac{1}{2k_n}\right) \right\} \right.
\end{aligned}$$

Table 2.11 Simulated REFF and RMSE$(\hat{\gamma}^{\text{Hill}})$, together with confidence intervals for Model (2.20), $\gamma = 0.25$

	$n = 200$	$n = 500$	$n = 1000$	$n = 2000$	$n = 5000$
Hill (RMSE)	0.0693 ± 0.00138	0.0445 ± 0.00107	0.0316 ± 0.00044	0.0225 ± 0.00030	0.0146 ± 0.00033
EMP: $p\gamma = -0.4$	1.2177 ± 0.01726	1.2015 ± 0.01704	1.1765 ± 0.01581	1.1663 ± 0.01279	1.1726 ± 0.01326
MOP: $p\gamma = -0.4$	0.7700 ± 0.01479	0.8142 ± 0.01327	0.8180 ± 0.01654	0.8315 ± 0.01293	0.8498 ± 0.01289
EMP: $p\gamma = -0.2$	1.2855 ± 0.01916	1.2384 ± 0.01757	1.2061 ± 0.01619	1.1945 ± 0.01353	1.1962 ± 0.01374
MOP: $p\gamma = -0.2$	0.8676 ± 0.00511	0.8971 ± 0.00531	0.8952 ± 0.00789	0.9013 ± 0.00866	0.9237 ± 0.01001
EMP: $p = 0$	1.3554 ± 0.0195	1.2882 ± 0.01844	1.2555 ± 0.01716	1.2352 ± 0.01429	1.2370 ± 0.01528
MOP: $p = 0$	1.0	1.0	1.0	1.0	1.0
EMP: $p\gamma = 0.2$	1.4516 ± 0.01868	1.3606 ± 0.01961	1.3194 ± 0.01781	1.3007 ± 0.01451	1.3012 ± 0.01698
MOP: $p\gamma = 0.2$	1.1660 ± 0.01507	1.1437 ± 0.01838	1.1307 ± 0.01236	1.1211 ± 0.01062	1.1424 ± 0.01580
EMP: $p\gamma = 0.4$	1.6205 ± 0.02659	1.5056 ± 0.02366	1.4321 ± 0.01646	1.4005 ± 0.01323	1.4024 ± 0.01902
MOP: $p\gamma = 0.4$	1.4454 ± 0.02433	1.3565 ± 0.02534	1.3024 ± 0.02074	1.3094 ± 0.01595	1.2963 ± 0.01851

$$
\begin{aligned}
&+\phi\left\{t^{-\gamma}\Psi_0(t) - \left(\frac{1}{2k_n}\right)^{-\gamma}\Psi_0\left(\frac{1}{2k_n}\right)\right\}\Bigg| \\
&= o_p^{(n)}(1) + O_p^{(n)}\left(\sup_{0<t\le\frac{1}{2k_n}} \frac{t}{q(t)}k_n^{\frac{1}{2}} + \sup_{0<t\le\frac{1}{2k_n}}\left|\frac{e(t)}{q(t)}\right| + \phi \sup_{0<t\le\frac{1}{2k_n}}\left|\frac{t\Psi_0(t)}{q(t)}\right|\right) \\
&= o_p^{(n)}(1),
\end{aligned}
$$

due to (2.6) and

$$
\Psi_0(t) = o^{(t)}(t^{-\delta}) \quad \forall\, \delta > 0. \tag{2.26}
$$

This completes the proof of Theorem 2.1.

2.6.3 Proof of Theorem 2.2

First, we consider the case $p \neq 0$. We can rewrite the MOP estimator as the following form:

$$\hat{\gamma}_p^{\text{MOP}}(k_n) = \frac{1 - 1\Big/ \int_0^1 \left(\frac{Q_n(t)}{Q_n(1)}\right)^p dt}{p}. \tag{2.27}$$

By (2.8), we have

$$\frac{Q_n(t)}{U\left(\frac{n}{k_n}\right)} = t^{-\gamma} + k_n^{-\frac{1}{2}} \left\{ o_p^{(n)}(1) t^{-\gamma-1} q(t) + \gamma t^{-\gamma-1} e(t) + \phi t^{-\gamma} \Psi_0(t) \right\} := t^{-\gamma} + k_n^{-\frac{1}{2}} y_n(t),$$

where

$$q(t) := t^{\frac{1}{2}} (1 + |\log t|)^{\frac{1}{2}}, \tag{2.28}$$

$$y_n(t) := k_n^{\frac{1}{2}} \left(\frac{Q_n(t)}{U\left(\frac{n}{k_n}\right)} - t^{-\gamma} \right) = o_p^{(n)}(1) t^{-\gamma-1} q(t) + \gamma t^{-\gamma-1} e(t) + \phi t^{-\gamma} \Psi_0(t).$$

By (2.6) and (2.26), we have $y_n(t) = o_p^{(t)}\left(t^{-\gamma-1} q(t)\right)$.

To begin with, consider the term:

$$\begin{aligned} \Delta_n &:= \int_0^1 \left(\frac{Q_n(t)}{Q_n(1)}\right)^p dt \\ &= k_n^{-1} \left(\frac{Q_n\left(\frac{1}{2k_n}\right)}{Q_n(1)} \right)^p + \int_{k_n^{-1}}^{k_n^{-1+\alpha}} \left(\frac{Q_n(t)}{Q_n(1)}\right)^p dt + \int_{k_n^{-1+\alpha}}^1 \left(\frac{Q_n(t)}{Q_n(1)}\right)^p dt \\ &=: \Delta_n^{(1)} + \Delta_n^{(2)} + \Delta_n^{(3)}, \end{aligned}$$

where α is a constant such that

$$(1 - p\gamma)(1 - \alpha) > \frac{1}{2} \quad \text{and} \quad \alpha > 0. \tag{2.29}$$

Because

$$y_n(t) = o_p^{(t)}\left(t^{-\gamma-\frac{1}{2}-\epsilon}\right), \quad \forall \epsilon > 0,$$

we have

$$k_n^{-\frac{1}{2}} t^{\gamma} y_n(t) \to_p 0,$$

uniformly for $t \in [k_n^{-1+\alpha}, 1]$. Hence, we can deduce that

$$
\begin{aligned}
\Delta_n^{(3)} &= \int_{k_n^{-1+\alpha}}^{1} \left(\frac{Q_n(t)/U\left(\frac{n}{k_n}\right)}{Q_n(1)/U\left(\frac{n}{k_n}\right)} \right)^p dt \\
&= \int_{k_n^{-1+\alpha}}^{1} \left(\frac{t^{-\gamma} + k_n^{-\frac{1}{2}} y_n(t)}{1 + k_n^{-\frac{1}{2}} y_n(1)} \right)^p dt \\
&= \int_{k_n^{-1+\alpha}}^{1} t^{-p\gamma} \left(1 + k_n^{-\frac{1}{2}} t^{\gamma} y_n(t) - k_n^{-\frac{1}{2}} y_n(1) + o_p^{(n)}(k_n^{-\frac{1}{2}})\right)^p dt \\
&= \int_{k_n^{-1+\alpha}}^{1} t^{-p\gamma} \left(1 + k_n^{-\frac{1}{2}} p t^{\gamma} y_n(t) - k_n^{-\frac{1}{2}} p y_n(1) + o_p^{(n)}(k_n^{-\frac{1}{2}})\right) dt \\
&= \left(1 + o_p^{(n)}(k_n^{-\frac{1}{2}})\right) \int_{k_n^{-1+\alpha}}^{1} t^{-p\gamma}\, dt \\
&\quad + k_n^{-\frac{1}{2}} p \int_{k_n^{-1+\alpha}}^{1} \left(o_p^{(n)}(1) t^{-p\gamma-1} q(t) + \gamma t^{-p\gamma-1} e(t) + \phi t^{-p\gamma} \Psi_0(t)\right) dt \\
&\quad - k_n^{-\frac{1}{2}} p \int_{k_n^{-1+\alpha}}^{1} \left(o_p^{(n)}(1) t^{-p\gamma} + \gamma t^{-p\gamma} e(1)\right) dt.
\end{aligned}
$$

Because (2.29), we have

$$
\int_{k_n^{-1+\alpha}}^{1} t^{-p\gamma}\, dt = \frac{1}{1-p\gamma}\left(1 - k_n^{(-1+\alpha)(1-p\gamma)}\right) = \frac{1}{1-p\gamma}\left(1 + o\left(k_n^{-\frac{1}{2}}\right)\right).
$$

On the other hand, note that

$$
\int_{k_n^{-1+\alpha}}^{1} t^{-p\gamma} \Psi_0(t)\, dt = \begin{cases} \int_{k_n^{-1+\alpha}}^{1} t^{-p\gamma} \frac{t^{-\rho}-1}{\rho}\, dt = \frac{1}{(1-p\gamma)(1-p\gamma-\rho)} + o(1), & \rho < 0, \\ \int_{k_n^{-1+\alpha}}^{1} -t^{-p\gamma} \log t\, dt = \frac{1}{(1-p\gamma)^2} + o(1), & \rho = 0. \end{cases}
$$

So, we conclude that

$$
\int_{k_n^{-1+\alpha}}^{1} t^{-p\gamma} \Psi_0(t)\, dt = \frac{1}{(1-p\gamma)(1-p\gamma-\rho)} + o(1).
$$

Furthermore, it can be shown that

$$
\int_0^1 t^{-p\gamma-1} q(t)\, dt \quad \text{and} \quad \int_0^1 \gamma t^{-p\gamma-1} e(t)\, dt
$$

are integrable. Hence,

$$
\begin{aligned}
\Delta_n^{(3)} &= \frac{1}{1-p\gamma}\left(1+o\left(k_n^{-\frac{1}{2}}\right)\right)+k_n^{-\frac{1}{2}}p\gamma\int_{k_n^{-1+\alpha}}^{1}t^{-p\gamma}\left(t^{-1}e(t)-e(1)\right)dt \\
&\quad +k_n^{-\frac{1}{2}}p\phi\int_{k_n^{-1+\alpha}}^{1}t^{-p\gamma}\Psi_0(t)\,dt+o_p^{(n)}(k_n^{-\frac{1}{2}}) \\
&= \frac{1}{1-p\gamma}+k_n^{-\frac{1}{2}}\left\{p\gamma\int_0^1 t^{-p\gamma}\left(t^{-1}e(t)-e(1)\right)dt\right. \\
&\quad \left.+\frac{p\phi}{(1-p\gamma)(1-p\gamma-\rho)}+o_p^{(n)}(1)\right\}. \qquad (2.30)
\end{aligned}
$$

For the term $\Delta_n^{(2)}$, note that for $t\in[k_n^{-1},k_n^{-1+\alpha}]$, $\forall\,\epsilon>0$,

$$
\left(\frac{Q_n(t)}{Q_n(1)}\right)^p=\left(\frac{t^{-\gamma}+k_n^{-\frac{1}{2}}y_n(t)}{1+k_n^{-\frac{1}{2}}y_n(1)}\right)^p=O_p^{(t)}\left(t^{-p\gamma-p\epsilon}\right).
$$

By (2.29), we have

$$
\Delta_n^{(2)}=O_p\left(\int_0^{k_n^{-1+\alpha}}t^{-p\gamma-p\epsilon}\,dt\right)=O_p\left(k_n^{(-1+\alpha)(1-p\gamma-p\epsilon)}\right)=O_p\left(k_n^{-\frac{1}{2}}\right). \qquad (2.31)
$$

For the remaining term $\Delta_n^{(1)}$, we have for any $\epsilon>0$,

$$
y_n\left(\frac{1}{2k_n}\right)=O_p\left(k_n^{\gamma+\frac{1}{2}+\epsilon}\right).
$$

So,

$$
\Delta_n^{(1)}=k_n^{-1}\left(\frac{\left(\frac{1}{2k_n}\right)^{-\gamma}+k_n^{-\frac{1}{2}}y_n\left(\frac{1}{2k_n}\right)}{1+k_n^{-\frac{1}{2}}y_n(1)}\right)^p=O_p\left(k_n^{-1+p\gamma+p\epsilon}\right)=O_p\left(k_n^{-\frac{1}{2}}\right). \qquad (2.32)
$$

Combining (2.30)–(2.32), we have

$$
\begin{aligned}
\Delta_n &= \frac{1}{1-p\gamma}+k_n^{-\frac{1}{2}}\left\{p\gamma\int_0^1 t^{-p\gamma}\left(t^{-1}e(t)-e(1)\right)dt\right. \\
&\quad \left.+\frac{p\phi}{(1-p\gamma)(1-p\gamma-\rho)}+o_p^{(n)}(1)\right\}. \qquad (2.33)
\end{aligned}
$$

Plugging (2.33) into (2.27), we have

$$\begin{aligned}
&\hat{\gamma}_p^{\mathrm{MOP}}(k_n)\\
&= p^{-1}\left(1 - \frac{1}{\frac{1}{1-p\gamma} + k_n^{-\frac{1}{2}}\left[p\gamma\int_0^1 t^{-p\gamma}(t^{-1}e(t)-e(1))\,dt + \frac{p\phi}{(1-p\gamma)(1-p\gamma-\rho)} + o_p^{(n)}(1)\right]}\right)\\
&= \gamma + k_n^{-\frac{1}{2}}\left[\gamma(1-p\gamma)^2\int_0^1 t^{-p\gamma}(t^{-1}e(t)-e(1))\,dt + \frac{(1-p\gamma)\phi}{1-p\gamma-\rho}\right] + o_p^{(n)}(k_n^{-\frac{1}{2}}).
\end{aligned}$$

Second, we consider the case $p = 0$; the MOP is degenerated into the Hill estimator:

$$\hat{\gamma}_0^{\mathrm{MOP}}(k_n) := \frac{1}{k_n}\sum_{i=1}^{k_n}\left\{\log\left(X_{n-i+1,n}\right) - \log\left(X_{n-k_n,n}\right)\right\}.$$

Divide $\hat{\gamma}_0^{\mathrm{MOP}}(k_n)$ into two parts:

$$\begin{aligned}
\hat{\gamma}_0^{\mathrm{MOP}}(k_n) &= \int_0^1 \log\left(\frac{Q_n(t)}{Q_n(1)}\right)dt\\
&= \int_0^{k_n^{-1+\alpha}} \log\left(\frac{Q_n(t)}{Q_n(1)}\right)dt + \int_{k_n^{-1+\alpha}}^1 \log\left(\frac{Q_n(t)}{Q_n(1)}\right)dt\\
&=: \tilde{\Delta}_n^{(1)} + \tilde{\Delta}_n^{(2)},
\end{aligned}$$

where $\alpha \in (0, 1/2)$. Note that

$$\begin{aligned}
\tilde{\Delta}_n^{(2)} &= \int_{k_n^{-1+\alpha}}^1 \log\left(\frac{t^{-\gamma} + k_n^{-\frac{1}{2}}y_n(t)}{1 + k_n^{-\frac{1}{2}}y_n(1)}\right)dt\\
&= \int_{k_n^{-1+\alpha}}^1 \left[-\gamma\log t + \log\left\{1 + k_n^{-\frac{1}{2}}\left(t^{\gamma}y_n(t) - y_n(1)\right) + o_p^{(n)}(k_n^{-\frac{1}{2}})\right\}\right]dt\\
&= \gamma + k_n^{-\frac{1}{2}}\int_0^1 \left(\gamma t^{-1}e(t) - \gamma e(1) + \phi\Psi_0(t)\right)dt + o_p^{(n)}(k_n^{-\frac{1}{2}}).
\end{aligned}$$

Further, we have

$$\int_0^1 \Psi_0(t)\,dt = \begin{cases}\int_0^1 \frac{t^{-\rho}-1}{\rho}\,dt = \frac{1}{1-\rho}, & \rho < 0,\\ \int_0^1 -\log t\,dt = 1, & \rho = 0.\end{cases}$$

Hence,

$$\tilde{\Delta}_n^{(2)} = \gamma + k_n^{-\frac{1}{2}}\int_0^1 \left(\gamma t^{-1}e(t) - \gamma e(1)\right)dt + \frac{\phi}{1-\rho} + o_p^{(n)}(k_n^{-\frac{1}{2}}). \quad (2.34)$$

For the term $\tilde{\Delta}_n^{(1)}$, note that for $t \in (0, k_n^{-1+\alpha}]$,

$$\log\left(\frac{Q_n(t)}{Q_n(1)}\right) = \log\left(\frac{t^{-\gamma} + k_n^{-\frac{1}{2}} y_n(t)}{1 + k_n^{-\frac{1}{2}} y_n(1)}\right) = O_p^{(t)}\left(\log(t)\right).$$

Hence,

$$\tilde{\Delta}_n^{(1)} = O_p\left(\int_0^{k_n^{-1+\alpha}} \log(t)\, dt\right) = O(k_n^{-\frac{1}{2}}). \tag{2.35}$$

Combining (2.34) with (2.35), we have

$$\hat{\gamma}_0^{\text{MOP}}(k_n) = \gamma + k_n^{-\frac{1}{2}} \int_0^1 \left(\gamma t^{-1} e(t) - \gamma e(1)\right) dt + \frac{\phi}{1-\rho} + o_p^{(n)}(k_n^{-\frac{1}{2}}). \tag{2.36}$$

In conclusion, for any p such that $p\gamma < 1/2$, we have

$$k_n^{\frac{1}{2}}\left(\hat{\gamma}_p^{\text{MOP}}(k_n) - \gamma\right) \to_d \gamma(1-p\gamma)^2 \int_0^1 t^{-p\gamma}(t^{-1}e(t) - e(1))\, dt + \frac{(1-p\gamma)\phi}{1-p\gamma-\rho}.$$

On the other hand, we have

$$\begin{aligned} & E\left\{\gamma(1-p\gamma)^2 \int_0^1 t^{-p\gamma}(t^{-1}e(t) - e(1))\, dt\right\} \\ & = \gamma(1-p\gamma)^2 \int_0^1 t^{-p\gamma} E\left(t^{-1}e(t) - e(1)\right) dt = 0 \end{aligned}$$

and

$$\begin{aligned} & Var\left\{\gamma(1-p\gamma)^2 \int_0^1 t^{-p\gamma}\left(t^{-1}e(t) - e(1)\right) dt\right\} \\ & = \gamma^2(1-p\gamma)^4 cov\left(\int_0^1 t^{-p\gamma}\left(t^{-1}e(t) - e(1)\right) dt,\ \int_0^1 s^{-p\gamma}\left(s^{-1}e(s) - e(1)\right) ds\right) \\ & = \gamma^2(1-p\gamma)^4 \int_0^1 \int_0^1 (st)^{-p\gamma-1} r(s,t)\, ds\, dt \\ & -\gamma^2(1-p\gamma)^4 \int_0^1 s^{-p\gamma} ds \int_0^1 t^{-p\gamma-1} r(t,1)\, dt \\ & -\gamma^2(1-p\gamma)^4 \int_0^1 t^{-p\gamma}\, dt \int_0^1 s^{-p\gamma-1} r(s,1)\, ds \\ & +\gamma^2(1-p\gamma)^4\, r(1,1) \int_0^1 s^{-p\gamma}\, ds \int_0^1 t^{-p\gamma}\, dt \end{aligned}$$

$$= \gamma^2(1-p\gamma)^4 \int_0^1 \int_0^1 (st)^{-p\gamma-1} r(s,t)\, ds\, dt$$
$$-2\gamma^2(1-p\gamma)^3 \int_0^1 t^{-p\gamma-1} r(t,1)\, dt + \gamma^2(1-p\gamma)^2\, r(1,1).$$

This completes the proof of Theorem 2.2.

2.6.4 Proof of Theorem 2.3

First, we consider the case $p \neq 0$. Recall that

$$\begin{aligned} &\hat{\gamma}^{\text{EMP}}_{m_n,p}(k_n) \\ &= \frac{1}{p(k_n-m_n)} \left[\sum_{i=m_n+1}^{k_n} \sum_{j=i}^{k_n} \log\left(\frac{e_{j,p}}{e_{j+1,p}}\right) + m_n \log\left(\frac{e_{m_n+1,p}}{e_{k_n+1,n}}\right) \right] \\ &= \frac{1}{p(k_n-m_n)} \sum_{j=m_n+1}^{k_n} j \log\left(\frac{e_{j,p}}{e_{j+1,p}}\right) \\ &= \frac{1}{p(k_n-m_n)} \sum_{j=m_n+1}^{k_n} j \left(\log\left(1+\frac{1}{j}\right) - \log\left(1 + \frac{1}{\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p}\right)\right), \end{aligned} \tag{2.37}$$

where $m_n = O\left(k_n^{1-\kappa}\right)$ with $\kappa \in \left(0, 1-(2-2p\gamma)^{-1}\right)$. To start with, consider the following approximation form of (2.37):

$$\begin{aligned} &\frac{1}{p(k_n-m_n)} \sum_{j=m_n+1}^{k_n} j \left(\frac{1}{j} - \frac{1}{\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p}\right) \\ &= \frac{1}{k_n-m_n} \sum_{j=m_n+1}^{k_n} \hat{\gamma}^{\text{MOP}}_p(j), \end{aligned} \tag{2.38}$$

where $\hat{\gamma}^{\text{MOP}}_p(j)$ is the MOP estimator with the threshold chosen as j. At first, we focus on the term

$$\Gamma_{n,j} := \frac{1}{j} \sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p = \int_0^1 \left(\frac{Q_n(u_j t)}{Q_n(u_j)}\right)^p dt,$$

where $u_j = j/k_n$ with $j = m_n+1, \ldots, k_n$. By (2.8), we have

$$\frac{Q_n(u_jt)}{U(\frac{n}{k_n})} = u_j^{-\gamma}t^{-\gamma} + k_n^{\frac{1}{2}}\Big(\gamma u_j^{-(\gamma+1)}t^{-(\gamma+1)}e(u_jt) + \phi u_j^{-\gamma}t^{-\gamma}\Psi_0(u_jt)$$
$$+o_p^{(n)}(1)u_j^{-\gamma-1}q(u_j)t^{-\gamma-1}q(t)\Big)$$
$$=: u_j^{-\gamma}\left(t^{-\gamma} + k_n^{-\frac{1}{2}}y_n^{(j)}(t)\right),$$

where $q(t)$ is defined in (2.28), and

$$y_n^{(j)}(t) := k_n^{\frac{1}{2}}\left(\frac{u_j^{\gamma}Q_n(u_jt)}{U\left(\frac{n}{k_n}\right)} - t^{-\gamma}\right)$$
$$= \gamma u_j^{-1}t^{-(\gamma+1)}e(u_jt) + \phi t^{-\gamma}\Psi_0(u_jt) + o_p^{(n)}(1)u_j^{-1}q(u_j)t^{-\gamma-1}q(t)$$
$$= O_p\left(t^{-\frac{1}{2}-\gamma}|\log(t)|^{\frac{1}{2}}u_j^{-\frac{1}{2}}|\log(u_j)|^{\frac{1}{2}}\right).$$

Take any $\delta \in \left(\kappa, 1-(2-2p\gamma)^{-1}\right)$, and divide $\Gamma_{n,j}$ into three parts:

$$\Gamma_{n,j} = (u_jk_n)^{-1}\left(\frac{Q_n\left(\frac{1}{2k_n}\right)}{Q_n(u_j)}\right)^p + \int_{(u_jk_n)^{-1}}^{k_n^{-1+\delta}}\left(\frac{Q_n(u_jt)}{Q_n(u_j)}\right)^p dt + \int_{k_n^{-1+\delta}}^{1}\left(\frac{Q_n(u_jt)}{Q_n(u_j)}\right)^p dt$$
$$=: \Gamma_{n,j}^{(1)} + \Gamma_{n,j}^{(2)} + \Gamma_{n,j}^{(3)}.$$

Note that for $t \in [k_n^{-1+\delta}, 1]$ and $u_j \in [(m_n+1)/k_n, 1]$ uniformly, we have

$$k_n^{-\frac{1}{2}}t^{\gamma}y_n^{(j)}(t) = O_p\left(k_n^{-\frac{1}{2}}t^{-\frac{1}{2}}|\log(t)|^{\frac{1}{2}}u_j^{-\frac{1}{2}}|\log(u_j)|^{\frac{1}{2}}\right) = O_p\left(k_n^{\frac{\kappa}{2}-\frac{\delta}{2}}|\log(k_n)|\right) = o_p(1).$$

Hence, for $u_j \in [(m_n+1)/k_n, 1]$ uniformly,

$$\Gamma_{n,j}^{(3)} = \int_{k_n^{-1+\delta}}^{1}\left(\frac{u_j^{\gamma}Q_n(u_jt)/U(\frac{n}{k_n})}{u_j^{\gamma}Q_n(u_j)/U(\frac{n}{k_n})}\right)^p dt$$
$$= \int_{k_n^{-1+\delta}}^{1} t^{-p\gamma}\left(1 + k_n^{-\frac{1}{2}}pt^{\gamma}y_n^{(j)}(t) - k_n^{-\frac{1}{2}}py_n^{(j)}(1) + o_p^{(n)}(k_n^{-\frac{1}{2}})\right) dt$$
$$= \left(1 + o_p^{(n)}(k_n^{-\frac{1}{2}})\right)\int_{k_n^{-1+\delta}}^{1} t^{-p\gamma}\,dt$$
$$+k_n^{-\frac{1}{2}}p\int_{k_n^{-1+\delta}}^{1}\Big(\gamma u_j^{-1}t^{-(p\gamma+1)}e(u_jt) + \phi t^{-p\gamma}\Psi_0(u_jt)$$
$$+o_p^{(n)}(1)u_j^{-1}q(u_j)t^{-p\gamma-1}q(t)\Big)\,dt$$
$$-k_n^{-\frac{1}{2}}p\int_{k_n^{-1+\delta}}^{1}\left(\gamma u_j^{-1}t^{-p\gamma}e(u_j) + \phi t^{-p\gamma}\Psi_0(u_j) + o_p^{(n)}(1)t^{-p\gamma}u_j^{-1}q(u_j)\right)dt.$$

Because $p\gamma < 1/2$, we can see that

$$\int_0^1 t^{-(p\gamma+1)} e(u_j t)\, dt \quad \text{and} \quad \int_0^1 \gamma t^{-p\gamma-1} q(t)\, dt$$

are integrable, and

$$\int_\epsilon^1 t^{-(p\gamma+1)} e(u_j t)\, dt \to_p 0, \qquad \int_\epsilon^1 \gamma t^{-p\gamma-1} q(t)\, dt \to_p 0,$$

as $\epsilon \to 0$. On the other hand, note that

$$\int_0^1 t^{-p\gamma} (\Psi_0(u_j t) - \Psi_0(u_j))\, dt = \begin{cases} \int_0^1 t^{-p\gamma} \frac{u_j^{-\rho} t^{-\rho} - u_j^{-\rho}}{\rho}\, dt = \frac{u_j^{-\rho}}{(1-p\gamma)(1-p\gamma-\rho)}, & \rho < 0, \\ \int_0^1 -t^{-p\gamma} \log t\, dt = \frac{1}{(1-p\gamma)^2}, & \rho = 0, \end{cases}$$

which can be summarized as

$$\int_0^1 t^{-p\gamma} (\Psi_0(u_j t) - \Psi_0(u_j))\, dt = \frac{u_j^{-\rho}}{(1 - p\gamma)(1 - p\gamma - \rho)}.$$

Furthermore, because $\delta < 1 - (2 - 2p\gamma)^{-1}$,

$$\int_{k_n^{-1+\delta}}^1 t^{-p\gamma}\, dt = \frac{1}{1 - p\gamma} \left(1 - k_n^{(-1+\delta)(1-p\gamma)}\right) = \frac{1}{1 - p\gamma} \left(1 + o\left(k_n^{-\frac{1}{2}}\right)\right).$$

Hence,

$$\begin{aligned} \Gamma_{n,j}^{(3)} &= \frac{1}{1 - p\gamma} + k_n^{-\frac{1}{2}} \left\{ p\gamma u_j^{-1} \int_0^1 t^{-p\gamma} \left(t^{-1} e(u_j t) - e(u_j)\right)\, dt \right. \\ &\quad \left. + p\phi \int_0^1 t^{-p\gamma} (\Psi_0(u_j t) - \Psi_0(u_j))\, dt \right\} + o_p^{(n)}(k_n^{-\frac{1}{2}}) \\ &= \frac{1}{1 - p\gamma} + k_n^{-\frac{1}{2}} \left\{ p\gamma u_j^{-1} \int_0^1 t^{-p\gamma} \left(t^{-1} e(u_j t) - e(u_j)\right)\, dt \right. \\ &\quad \left. + \frac{u_j^{-\rho} p\phi}{(1 - p\gamma)(1 - p\gamma - \rho)} + o_p^{(n)}(1) \right\}. \end{aligned} \tag{2.39}$$

For the term $\Gamma_{n,j}^{(2)}$, note that for $t \in [(u_j k_n)^{-1}, k_n^{-1+\delta}]$ and $u_j \in [(m_n + 1)/k_n, 1]$ uniformly,

$$\left(\frac{Q_n(u_j t)}{Q_n(u_j)}\right)^p = \left(\frac{t^{-\gamma} + k_n^{-\frac{1}{2}} y_n^{(j)}(t)}{1 + k_n^{-\frac{1}{2}} y_n^{(j)}(1)}\right)^p$$
$$= O_p\left(\left(k_n^{-\frac{1}{2}} t^{-\frac{1}{2}-\gamma} \left|\log(t)\right|^{\frac{1}{2}} u_j^{-\frac{1}{2}} \left|\log(u_j)\right|^{\frac{1}{2}}\right)^p\right)$$
$$= O\left(\left|\log(k_n)\right|^p t^{-p\gamma}\right).$$

So, we have

$$\Gamma_{n,j}^{(2)} = \int_{(u_j k_n)^{-1}}^{k_n^{-1+\delta}} \left(\frac{Q_n(u_j t)}{Q_n(u_j)}\right)^p dt$$
$$= O_p\left(\left|\log(k_n)\right|^p \int_0^{k_n^{-1+\delta}} t^{-p\gamma}\, dt\right)$$
$$= o(k_n^{-\frac{1}{2}}). \tag{2.40}$$

Finally, we consider the remaining term $\Gamma_{n,j}^{(1)}$. Note that

$$y_n^{(j)}\left(\frac{1}{2u_j k_n}\right) = O_p\left(k_n^{\gamma+\frac{1}{2}} \left|\log(k_n)\right|^{\frac{1}{2}} u_j^{\gamma} \left|\log(u_j)\right|^{\frac{1}{2}}\right).$$

So, for $u_j \in [(m_n+1)/k_n, 1]$ uniformly, we have

$$\Gamma_{n,j}^{(1)} = (u_j k_n)^{-1} \left(\frac{\left(\frac{1}{2k_n}\right)^{-\gamma} + k_n^{-\frac{1}{2}} y_n^{(j)}\left(\frac{1}{2u_j k_n}\right)}{1 + k_n^{-\frac{1}{2}} y_n^{(j)}(1)}\right)^p$$
$$= O_p\left(k_n^{-1+p\gamma} \left|\log(k_n)\right|^p u_j^{-1+p\gamma}\right)$$
$$= O\left(k_n^{-\frac{1}{2}}\right). \tag{2.41}$$

Combining (2.39)–(2.41), we can conclude that

$$\int_0^1 \left(\frac{Q_n(u_j t)}{Q_n(u_j)}\right)^p dt = \frac{1}{1-p\gamma} + k_n^{-\frac{1}{2}}\Bigg\{p\gamma u_j^{-1}\int_0^1 t^{-p\gamma}\left(t^{-1}e(u_j t) - e(u_j)\right) dt$$
$$+\frac{u_j^{-\rho}\, p\phi}{(1-p\gamma)(1-p\gamma-\rho)} + o_p^{(n)}(1)\Bigg\}, \tag{2.42}$$

which leads to the following result:

$$
\begin{aligned}
&\hat{\gamma}_p^{\mathrm{MOP}}(j)\\
&= p^{-1}\left(1-\frac{1}{\frac{1}{j}\sum_{i=1}^{j}\left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p}\right)\\
&= p^{-1}\left(1-\frac{1}{\int_0^1\left(\frac{Q_n(u_j t)}{Q_n(u_j)}\right)^p dt}\right)\\
&= p^{-1}\Bigg(1-(1-p\gamma)\Bigg[1+k_n^{-\frac{1}{2}}(1-p\gamma)\Bigg\{p\gamma u_j^{-1}\int_0^1 t^{-p\gamma}\left(t^{-1}e(u_j t)-e(u_j)\right)dt\\
&+\frac{u_j^{-\rho}p\phi}{(1-p\gamma)(1-p\gamma-\rho)}+o_p^{(n)}(1)\Bigg\}\Bigg]^{-1}\Bigg)\\
&= p^{-1}\Bigg[1-(1-p\gamma)\Bigg\{1-k_n^{-\frac{1}{2}}(1-p\gamma)\Bigg(p\gamma u_j^{-1}\int_0^1 t^{-p\gamma}\left(t^{-1}e(u_j t)-e(u_j)\right)dt\\
&\quad+\frac{u_j^{-\rho}p\phi}{(1-p\gamma)(1-p\gamma-\rho)}+o_p^{(n)}(1)\Bigg)\Bigg\}\Bigg]\\
&= \gamma+k_n^{-\frac{1}{2}}\Bigg(\gamma(1-p\gamma)^2 u_j^{-1}\int_0^1 t^{-p\gamma}\left(t^{-1}e(u_j t)-e(u_j)\right)dt\\
&+\frac{(1-p\gamma)\phi u_j^{-\rho}}{1-p\gamma-\rho}+o_p^{(n)}(1)\Bigg),
\end{aligned}
\tag{2.43}
$$

uniformly for $j \in [m_n+1, k_n]$. Note that the term $\sum_{j=m_n+1}^{k_n}\hat{\gamma}_p^{\mathrm{MOP}}(j)$ is the summation of the function

$$
P(u) := p^{-1}\left(1-1\bigg/\int_0^1\left(\frac{Q_n(ut)}{Q_n(u)}\right)^p dt\right),
$$

over $u=(m_n+1)/k_n, (m_n+2)/k_n, \ldots, 1$. So by the definition of the integral, we have

$$
\begin{aligned}
&\frac{1}{k_n-m_n}\sum_{j=m_n+1}^{k_n}\hat{\gamma}_p^{\mathrm{MOP}}(j)\\
&= \frac{k_n}{k_n-m_n}\left(\frac{1}{k_n}\sum_{j=m_n+1}^{k_n}\hat{\gamma}_p^{\mathrm{MOP}}(j)\right)\\
&= \gamma+\frac{k_n^{-\frac{1}{2}}}{1-m_n k_n^{-1}}\Bigg(\gamma(1-p\gamma)^2\int_{m_n/k_n}^1 u^{-1}\int_0^1 t^{-p\gamma}\left(t^{-1}e(ut)-e(u)\right)dt\,du\\
&+\frac{(1-p\gamma)\phi}{1-p\gamma-\rho}\int_{m_n/k_n}^1 u^{-\rho}\,du+o_p^{(n)}(1)\Bigg)
\end{aligned}
$$

$$
\begin{aligned}
&= \gamma + k_n^{-\frac{1}{2}} \left(\gamma(1-p\gamma)^2 \int_0^1 u^{-1} \int_0^1 t^{-p\gamma} \left(t^{-1} e(ut) - e(u) \right) dt\, du \right. \\
&\left. + \frac{(1-p\gamma)\phi}{(1-p\gamma-\rho)(1-\rho)} + o_p^{(n)}(1) \right), \qquad (2.44)
\end{aligned}
$$

since $m_n/k_n \to 0$.

In the next step, we need to verify that (2.37) and (2.38) are asymptotically equivalent, namely,

$$
\begin{aligned}
&\frac{1}{p(k_n-m_n)} \left| \sum_{j=m_n+1}^{k_n} j \left\{ \log\left(1+\frac{1}{j}\right) - \log\left(1 + \frac{1}{\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p}\right) \right\} \right. \\
&\left. - \sum_{j=m_n+1}^{k_n} j \left\{ \frac{1}{j} - \frac{1}{\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p} \right\} \right| \\
&= o_p^{(n)}\left(k_n^{-\frac{1}{2}}\right). \qquad (2.45)
\end{aligned}
$$

Because $|x - \log(1+x)| < x^2/2$ for $x > 0$, we have

$$
\begin{aligned}
&\frac{1}{p(k_n-m_n)} \left| \sum_{j=m_n+1}^{k_n} j \left\{ \log\left(1+\frac{1}{j}\right) - \log\left(1 + \frac{1}{\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p}\right) \right\} \right. \\
&\left. - \sum_{j=m_n+1}^{k_n} j \left\{ \frac{1}{j} - \frac{1}{\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p} \right\} \right| \\
&\le \frac{1}{p(k_n-m_n)} \sum_{j=m_n+1}^{k_n} j \left| \log\left(1+\frac{1}{j}\right) - \frac{1}{j} \right| \\
&+ \frac{1}{p(k_n-m_n)} \left| \sum_{j=m_n+1}^{k_n} j \left\{ \frac{1}{\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p} - \log\left(1 + \frac{1}{\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p}\right) \right\} \right| \\
&\le \frac{1}{p(k_n-m_n)} \sum_{j=m_n+1}^{k_n} \frac{1}{2j} \\
&+ \frac{1}{p(k_n-m_n)} \left| \sum_{j=m_n+1}^{k_n} j \left\{ \frac{1}{\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p} - \log\left(1 + \frac{1}{\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p}\right) \right\} \right| \\
&=: A_n^{(1)} + A_n^{(2)}.
\end{aligned}
$$

Note that

$$\begin{aligned}
A_n^{(1)} &= \frac{1}{p(k_n - m_n)} \sum_{j=m_n+1}^{k_n} \frac{1}{2j} \\
&< \frac{1}{2p(k_n - m_n)} \int_1^{k_n} \frac{1}{t}\, dt = \frac{1}{2p(k_n - m_n)} \log(k_n) = o(k_n^{-\frac{1}{2}}).
\end{aligned}$$

Hence, it suffices to verify that $A_n^{(2)} = o_p^{(n)}\left(k_n^{-\frac{1}{2}}\right)$.

If $p > 0$, we can easily see that

$$\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p \geq j.$$

So

$$\begin{aligned}
A_n^{(2)} &\leq \frac{1}{p(k_n - m_n)} \sum_{j=m_n+1}^{k_n} j \frac{1}{2\left(\sum_{i=1}^{j}\left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p\right)^2} \\
&\leq \frac{1}{p(k_n - m_n)} \sum_{j=m_n+1}^{k_n} j \frac{1}{2j^2} = o(k_n^{-\frac{1}{2}}).
\end{aligned}$$

If $p < 0$,

$$\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p \leq j. \tag{2.46}$$

Define

$$\begin{aligned}
\tilde{A}_n^{(2)} &:= \frac{k_n - m_n}{k_n} A_n^{(2)} \\
&= \frac{1}{pk_n} \left| \sum_{j=m_n+1}^{k_n} j \left\{ \frac{1}{\sum_{i=1}^{j}\left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p} - \log\left(1 + \frac{1}{\sum_{i=1}^{j}\left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p}\right)\right\}\right|.
\end{aligned}$$

Because $A_n^{(2)} = \tilde{A}_n^{(2)}(1 + o(1))$, it suffices to show that $\tilde{A}_n^{(2)} = o_p^{(n)}\left(k_n^{-\frac{1}{2}}\right)$. Divide $\tilde{A}_n^{(2)}$ into three parts:

$$\begin{aligned}\tilde{A}_n^{(2)} &= \frac{1}{pk_n}\sum_{j=m_n+1}^{[k_n^{-\frac{1}{2}-\delta_0}]} j\left\{\frac{1}{\sum_{i=1}^{j}\left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p} - \log\left(1+\frac{1}{\sum_{i=1}^{j}\left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p}\right)\right\}\\
&\quad+\frac{1}{pk_n}\sum_{j=[k_n^{-\frac{1}{2}-\delta_0}]+1}^{[k_n^{-\frac{1}{2}+\delta_0}]} j\left\{\frac{1}{\sum_{i=1}^{j}\left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p} - \log\left(1+\frac{1}{\sum_{i=1}^{j}\left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p}\right)\right\}\\
&\quad+\frac{1}{pk_n}\sum_{j=[k_n^{-\frac{1}{2}+\delta_0}]+1}^{k_n} j\left\{\frac{1}{\sum_{i=1}^{j}\left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p} - \log\left(1+\frac{1}{\sum_{i=1}^{j}\left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p}\right)\right\}\\
&=: B_n^{(1)} + B_n^{(2)} + B_n^{(3)},\end{aligned}$$

$\forall\, \delta_0 \in (0, 0.125)$. By (2.42), we know that for $j = m_n + 1, \ldots, k_n$,

$$\begin{aligned}&\frac{j}{\sum_{i=1}^{j}\left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p}\\
&= (1-p\gamma) - k_n^{-\frac{1}{2}}(1-p\gamma)^2\left\{p\gamma u_j^{-1}\int_0^1 t^{-p\gamma}\left(t^{-1}e(u_j t) - e(u_j)\right)dt\right.\\
&\quad\left.+\frac{u_j^{-\rho}p\phi}{(1-p\gamma)(1-p\gamma-\rho)} + o_p^{(n)}(1)\right\}\\
&=: (1-p\gamma) - k_n^{-\frac{1}{2}} z_1(u_j) + o_p^{(n)}\left(k_n^{-\frac{1}{2}}\right),\end{aligned}$$

where $z_1(u)$ is defined as

$$\begin{aligned}z_1(u) := (1-p\gamma)^2&\left\{p\gamma u^{-1}\int_0^1 t^{-p\gamma}\left(t^{-1}e(ut) - e(u)\right)dt\right.\\
&\left.+\frac{u^{-\rho}p\phi}{(1-p\gamma)(1-p\gamma-\rho)}\right\}.\end{aligned}$$

Note that $\forall\, \epsilon > 0$,

$$t^{-p\gamma}\left(t^{-1}e(ut) - e(u)\right) = o_p\left(t^{-p\gamma-\frac{1}{2}-\epsilon}u^{\frac{1}{2}-\epsilon}\right),$$

so we have

$$z_1(u) = o_p\left(u^{-\frac{1}{2}-\epsilon}\right).$$

For the term $B_n^{(1)}$, by (2.46) and $\log(1+x) < x$, we have

$$
\begin{aligned}
B_n^{(1)} &\le \frac{1}{pk_n} \sum_{j=m_n+1}^{[k_n^{\frac{1}{2}-\delta_0}]} \frac{j}{\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p} \\
&= \frac{1}{p} \int_{m_n/k_n}^{k_n^{-\frac{1}{2}-\delta_0}} \left\{(1-p\gamma) - k_n^{-\frac{1}{2}} z_1(u)\right\} dt + o_p^{(n)}\left(k_n^{-\frac{1}{2}}\right) \\
&\le \frac{1-p\gamma}{p} k_n^{-\frac{1}{2}-\delta_0} - \frac{k_n^{-\frac{1}{2}}}{p} \int_0^{k_n^{-\frac{1}{2}-\delta_0}} z_1(u)\, dt + o_p^{(n)}(k_n^{-\frac{1}{2}}) \\
&= o_p^{(n)}(k_n^{-\frac{1}{2}}). \qquad (2.47)
\end{aligned}
$$

We then study the two remaining terms: $B_n^{(2)}$ and $B_n^{(3)}$. Firstly note that $k_n^{-\frac{1}{2}} < u_j^{\frac{1}{2}}$, so we have

$$
\begin{aligned}
&\left| \frac{j}{\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p} \right| \\
&< (1-p\gamma) + (1-p\gamma)^2 p\gamma u_j^{-\frac{1}{2}} \int_0^1 t^{-p\gamma} \left(t^{-1} e(u_j t) - e(u_j)\right) dt \\
&\quad + \frac{k_n^{-\frac{1}{2}} (1-p\gamma)^2 u_j^{-\rho} p\phi}{(1-p\gamma)(1-p\gamma-\rho)} + o_p^{(n)}(k_n^{-\frac{1}{2}}) \\
&=: (1-p\gamma) + z_2(u_j) + \frac{k_n^{-\frac{1}{2}} (1-p\gamma)^2 u_j^{-\rho} p\phi}{(1-p\gamma)(1-p\gamma-\rho)} + o_p^{(n)}(k_n^{-\frac{1}{2}}),
\end{aligned}
$$

where

$$
z_2(u) := (1-p\gamma)^2 p\gamma u^{-\frac{1}{2}} \int_0^1 t^{-p\gamma} \left(t^{-1} e(ut) - e(u)\right) dt = o_p(u^{-\epsilon}), \quad \forall\, \epsilon > 0.
$$

Hence, for $B_n^{(2)}$, taking ϵ small enough, we have

$$
\begin{aligned}
B_n^{(2)} &\le \frac{1}{2pk_n} \sum_{j=[k_n^{-\frac{1}{2}-\delta_0}]+1}^{[k_n^{-\frac{1}{2}+\delta_0}]} \frac{1}{j} \left(\frac{j}{\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p} \right)^2 \\
&< \frac{1}{2pk_n^{\frac{1}{2}-\delta_0}} \int_0^{k_n^{-\frac{1}{2}+\delta_0}} \left\{ (1-p\gamma) + z_2(u) + \frac{k_n^{-\frac{1}{2}} (1-p\gamma)^2 u_j^{-\rho} p\phi}{(1-p\gamma)(1-p\gamma-\rho)} \right. \\
&\qquad \left. + o_p^{(n)}(k_n^{-\frac{1}{2}}) \right\}^2 du
\end{aligned}
$$

$$
\begin{aligned}
&= \frac{1}{2pk_n^{\frac{1}{2}-\delta_0}} \left\{ O_p^{(n)} \left(\int_0^{k_n^{-\frac{1}{2}+\delta_0}} u^{-\epsilon}\, du \right) + O_p^{(n)}(k_n^{-\frac{1}{2}+\delta_0}) \right\} \\
&= O_p^{(n)} \left(k_n^{\left(-\frac{1}{2}+\delta_0\right)(2-\epsilon)} \right) \\
&= o_p^{(n)} \left(k_n^{-\frac{1}{2}} \right). \qquad (2.48)
\end{aligned}
$$

For $B_n^{(3)}$, we have

$$
\begin{aligned}
B_n^{(3)} &\le \frac{1}{2pk_n} \sum_{j=[k_n^{-\frac{1}{2}+\delta_0}]+1}^{k_n} \frac{1}{j} \left(\frac{j}{\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}} \right)^p} \right)^2 \\
&\le \frac{1}{2pk_n^{\frac{3}{2}+\delta_0}} \sum_{j=[k_n^{-\frac{1}{2}+\delta_0}]+1}^{k_n} \left(\frac{j}{\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}} \right)^p} \right)^2 \\
&= \frac{1}{2pk_n^{\frac{1}{2}+\delta_0}} \int_{k_n^{-\frac{1}{2}}+\delta_0}^{1} \left\{ (1-p\gamma) + z_2(u) + \frac{k_n^{-\frac{1}{2}}(1-p\gamma)^2 u_j^{-\rho} p\phi}{(1-p\gamma)(1-p\gamma-\rho)} \right. \\
&\qquad \left. + o_p^{(n)}(k_n^{-\frac{1}{2}}) \right\}^2 du \\
&= \frac{1}{2pk_n^{\frac{1}{2}+\delta_0}} O_p^{(n)}(1) \\
&= o_p^{(n)} \left(k_n^{-\frac{1}{2}} \right). \qquad (2.49)
\end{aligned}
$$

Combining (2.47)–(2.49), we have $A_n^{(2)} = o_p^{(n)} \left(k_n^{-\frac{1}{2}} \right)$. By (2.37), (2.44) and (2.45), we have

$$
\begin{aligned}
&k_n^{\frac{1}{2}} (\hat{\gamma}_{m_n,p}^{\mathrm{EMP}}(k_n) - \gamma) \\
&= \gamma(1-p\gamma)^2 \int_0^1 u^{-1} \int_0^1 t^{-p\gamma} \left(t^{-1} e(ut) - e(u) \right) dt\, du + \frac{(1-p\gamma)\phi}{(1-p\gamma-\rho)(1-\rho)}, \\
&+ o_p^{(n)}(1) \qquad (2.50)
\end{aligned}
$$

for $p\gamma < 1/2$ and $p \neq 0$.

Second, we consider the case $p = 0$; $\hat{\gamma}_{m_n,0}^{\mathrm{EMP}}(k_n)$ is defined as

$$
\begin{aligned}
&\hat{\gamma}^{\mathrm{EMP}}_{m_n,0}(k_n) \\
&= \frac{1}{k_n - m_n}\left[\sum_{j=m_n+1}^{k_n}\left\{\frac{1}{j}\sum_{i=1}^{j}\log\left(X_{n-i+1,n}\right)\right\} + \frac{m_n}{m_n+1}\sum_{i=1}^{m_n+1}\log\left(X_{n-i+1,n}\right)\right. \\
&\quad \left. -\frac{k_n}{k_n+1}\sum_{i=1}^{k_n+1}\log\left(X_{n-i+1,n}\right)\right] \\
&= \frac{1}{k_n - m_n}\sum_{j=m_n+1}^{k_n}\hat{\gamma}^{\mathrm{Hill}}(j) \\
&\quad +\frac{1}{k_n - m_n}\left\{\hat{\gamma}^{\mathrm{Hill}}(k_n+1) - \hat{\gamma}^{\mathrm{Hill}}(m_n+1) + \log\left(\frac{X_{n-k_n,n}}{X_{n-m_n,n}}\right)\right\} \\
&=: C_n^{(1)} + C_n^{(2)},
\end{aligned}
$$

where $m_n = O\left(k_n^{1-\kappa}\right)$ with $\kappa \in (0, 1/2)$. Take any $\delta \in (\kappa, 1/2)$ and divide $\hat{\gamma}^{\mathrm{Hill}}(j)$ into two parts:

$$
\hat{\gamma}^{\mathrm{Hill}}(j) = \int_0^{k_n^{-1+\delta}} \log\left(\frac{Q\left(u_j t\right)}{Q\left(u_j\right)}\right) dt + \int_{k_n^{-1+\delta}}^{1} \log\left(\frac{Q\left(u_j t\right)}{Q\left(u_j\right)}\right) dt =: \tilde{\Gamma}^{(1)}_{n,j} + \tilde{\Gamma}^{(2)}_{n,j}.
$$

As in $\Gamma^{(2)}_{n,j}$, we can similarly show that for $u_j \in [(m_n+1)/k_n, 1]$ uniformly,

$$
\begin{aligned}
\tilde{\Gamma}^{(2)}_{n,j} &= \int_{k_n^{-1+\delta}}^{1} \log\left(\frac{t^{-\gamma} + k_n^{-\frac{1}{2}} y_n^{(j)}(t)}{1 + k_n^{-\frac{1}{2}} y_n^{(j)}(1)}\right) dt \\
&= \int_{k_n^{-1+\delta}}^{1} -\gamma \log(t) + \log\left(1 + k_n^{-\frac{1}{2}} t^{\gamma} y_n^{(j)}(t) - k_n^{-\frac{1}{2}} y_n^{(j)}(1) + o_p^{(n)}(k_n^{-\frac{1}{2}})\right) dt \\
&= \int_{k_n^{-1+\delta}}^{1} -\gamma \log(t)\, dt + k_n^{-\frac{1}{2}} \int_{k_n^{-1+\delta}}^{1} \left(\gamma u_j^{-1} t^{-1} e(u_j t) + \phi \Psi_0(u_j t)\right. \\
&\quad \left. + o_p^{(n)}(1) u_j^{-1} q(u_j) t^{-1} q(t)\right) dt \\
&\quad - k_n^{-\frac{1}{2}} \int_{k_n^{-1+\delta}}^{1} \left(\gamma u_j^{-1} e(u_j) + \phi \Psi_0(u_j) + o_p^{(n)}(1) u_j^{-1} q(u_j)\right) dt \\
&= \gamma + k_n^{-\frac{1}{2}} \left\{ u_j^{-1}\gamma \int_0^1 \left\{t^{-1} e(u_j t) - e(u_j)\right\} dt + \frac{u_j^{-\rho}\phi}{1-\rho} + o_p^{(n)}(1) \right\}.
\end{aligned}
$$

Furthermore, take $\epsilon > 0$ small enough, for $t \in (0, k_n^{-1+\delta}]$ and $u_j \in [(m_n+1)/k_n, 1]$ uniformly,

$$
k_n^{-\frac{1}{2}} y_n^{(j)}(t) = O_p\left(k_n^{-\frac{1}{2}} t^{-\frac{1}{2}-\gamma-\epsilon} u_j^{-\frac{1}{2}-\epsilon}\right) = O\left(t^{-\frac{1}{2}-\gamma-\epsilon}\right).
$$

Since $\delta < 1/2$, we have

$$\tilde{\Gamma}^{(1)}_{n,j} = \int_0^{k_n^{-1+\delta}} \log\left(\frac{t^{-\gamma} + k_n^{-\frac{1}{2}} y_n^{(j)}(t)}{1 + k_n^{-\frac{1}{2}} y_n^{(j)}(1)}\right) dt = O_p\left(\int_0^{k_n^{-1+\delta}} \log(t)\, dt\right) = o\left(k_n^{-\frac{1}{2}}\right).$$

In conclusion, for $u_j \in [(m_n+1)/k_n, 1]$ uniformly,

$$\hat{\gamma}^{\text{Hill}}(j) = \gamma + k_n^{-\frac{1}{2}}\left[u_j^{-1}\gamma \int_0^1 \left\{t^{-1}e(u_j t) - e(u_j)\right\} dt + \frac{u_j^{-\rho}\phi}{1-\rho} + o_p^{(n)}(1)\right], \quad (2.51)$$

which gives rise to

$$\begin{aligned} C_n^{(1)} &= \gamma + k_n^{-\frac{1}{2}}\gamma\left[\int_{m_n/k_n}^1 u^{-1}\int_0^1 \left\{t^{-1}e(ut) - e(u)\right\} dt\, du \right. \\ &\quad \left. + \frac{\phi}{1-\rho}\int_{m_n/k_n}^1 u^{-\rho}\, du + o_p^{(n)}(1)\right] \\ &= \gamma + k_n^{-\frac{1}{2}}\gamma\left[\int_0^1 u^{-1}\int_0^1 \left\{t^{-1}e(ut) - e(u)\right\} dt\, du + \frac{\phi}{(1-\rho)^2} + o_p^{(n)}(1)\right]. \end{aligned}$$

For the term $C_n^{(2)}$, by (2.51), we have

$$\hat{\gamma}^{\text{Hill}}(m_n+1) = \gamma + o_p^{(n)}(1), \quad \hat{\gamma}^{\text{Hill}}(k_n+1) = \gamma + o_p^{(n)}(1).$$

Moreover, $\forall\, \epsilon > 0$,

$$\log\left(\frac{X_{n-m_n,n}}{X_{n-k_n,n}}\right) = \gamma \log\left(\frac{k_n}{m_n}\right)(1 + o_p(1)) = o\left(k_n^{\epsilon}\right).$$

In addition,

$$\frac{1}{k_n - m_n} = \frac{1}{k_n}(1 + o(1)).$$

Hence,

$$C_n^{(2)} = o_p\left(k_n^{-\frac{1}{2}}\right).$$

We conclude that for $m_n = O\left(k_n^{1-\kappa}\right)$ with $\kappa \in (0, 1/2)$,

$$\hat{\gamma}^{\text{EMP}}_{m_n,0}(k_n) = \gamma + k_n^{-\frac{1}{2}}\gamma\left[\int_0^1 u^{-1}\int_0^1 \left\{t^{-1}e(ut) - e(u)\right\} dt\, du + \frac{\phi}{(1-\rho)^2} + o_p^{(n)}(1)\right]. \quad (2.52)$$

Summarizing, in view of (2.50) and (2.52), for $\forall\ p < (2\gamma)^{-1}$, given $m_n = O\left(k_n^{1-\kappa}\right)$ with $\kappa \in \left(0, 1-(2-2p\gamma)^{-1}\right)$, we have

$$
\begin{aligned}
&k_n^{\frac{1}{2}}(\hat{\gamma}_{m_n,p}^{\text{EMP}}(k_n) - \gamma)\\
&= \gamma(1-p\gamma)^2 \int_0^1 u^{-1} \int_0^1 t^{-p\gamma}\left(t^{-1}e(ut) - e(u)\right) dt\, du\\
&+\frac{(1-p\gamma)\phi}{(1-p\gamma-\rho)(1-\rho)} + o_p^{(n)}(1).
\end{aligned}
$$

On the other hand, we have

$$
\begin{aligned}
&E\left\{\gamma(1-p\gamma)^2 \int_0^1 \int_0^1 u^{-1}t^{-p\gamma}\left(t^{-1}e(ut) - e(u)\right) dt\, du\right\}\\
&= \gamma(1-p\gamma)^2 \int_0^1 \int_0^1 u^{-1}t^{-p\gamma} E\left(t^{-1}e(ut) - e(u)\right) dt\, du\\
&= 0,
\end{aligned}
$$

and

$$
\begin{aligned}
&Var\left\{\gamma(1-p\gamma)^2 \int_0^1 \int_0^1 u^{-1}t^{-p\gamma}\left(t^{-1}e(ut) - e(u)\right) dt\, du\right\}\\
&= \gamma^2(1-p\gamma)^4 \int_0^1\int_0^1\int_0^1\int_0^1 Cov\left(u^{-1}t^{-p\gamma-1}e(ut) - u^{-1}t^{-p\gamma}e(u),\right.\\
&\left. s^{-1}v^{-p\gamma-1}e(vs) - s^{-1}v^{-p\gamma}e(s)\right) du\, dv\, ds\, dt\\
&= \gamma^2(1-p\gamma)^4 \int_0^1\int_0^1\int_0^1\int_0^1 (us)^{-1}(vt)^{-p\gamma-1} r(ut, vs)\, du\, dv\, ds\, dt\\
&-\gamma^2(1-p\gamma)^4 \int_0^1 v^{-p\gamma} dv \int_0^1\int_0^1\int_0^1 (us)^{-1}t^{-p\gamma-1} r(ut, s)\, du\, ds\, dt\\
&-\gamma^2(1-p\gamma)^4 \int_0^1 t^{-p\gamma} dt \int_0^1\int_0^1\int_0^1 (us)^{-1}v^{-p\gamma-1} r(vs, u)\, du\, dv\, ds\\
&+\gamma^2(1-p\gamma)^4 \int_0^1 v^{-p\gamma} dv \int_0^1 t^{-p\gamma} dt \int_0^1\int_0^1\int_0^1 (us)^{-1} r(u, s)\, du\, ds\\
&= \gamma^2(1-p\gamma)^2 \int_0^1\int_0^1 (us)^{-1} r(u, s)\, du\, ds\\
&-2\gamma^2(1-p\gamma)^3 \int_0^1\int_0^1\int_0^1 (us)^{-1}t^{-p\gamma-1} r(st, u)\, du\, ds\, dt\\
&+\gamma^2(1-p\gamma)^4 \int_0^1\int_0^1\int_0^1\int_0^1 (us)^{-1}(vt)^{-p\gamma-1} r(uv, st)\, du\, dv\, ds\, dt.
\end{aligned}
$$

This completes the proof of Theorem 2.3.

2.6.5 Proof of Theorem 2.4

To start with, we concentrate on the EMP estimator; this case is more complicated. Take $m_n = O\left(k_n^{1-\kappa}\right)$ with $\kappa \in \left(0, 1-(2-2p\gamma)^{-1}\right)$. By (2.43) and (2.44), for $s_0 \in [\nu, 1]$ uniformly, where $\nu \in (0, 1)$, we have

$$
\begin{aligned}
&\frac{1}{\lceil s_0 k_n \rfloor - m_n} \sum_{j=m_n+1}^{\lceil s_0 k_n \rfloor} \hat{\gamma}_p^{\mathrm{MOP}}(j) \\
&= \gamma + \frac{k_n^{\frac{1}{2}}}{\lceil s_0 k_n \rfloor - m_n} \left(\gamma(1-p\gamma)^2 \int_{m_n/k_n}^{s_0} u^{-1} \int_0^1 t^{-p\gamma} \left\{ t^{-1} e(ut) - e(u) \right\} dt\, du \right. \\
&\quad \left. + \frac{(1-p\gamma)\phi}{1-p\gamma-\rho} \int_{m_n/k_n}^{s_0} u^{-\rho}\, du + o_p^{(n)}(1) \right) \\
&= \gamma + \frac{k_n^{\frac{1}{2}}}{\lceil s_0 k_n \rfloor - m_n} \left(\gamma(1-p\gamma)^2 \int_{m_n/(k_n s_0)}^{1} w^{-1} \int_0^1 t^{-p\gamma} \left\{ t^{-1} e(ut) - e(u) \right\} dt\, dw \right. \\
&\quad \left. + \frac{(1-p\gamma)\phi s_0^{1-\rho}}{1-p\gamma-\rho} \int_{m_n/(k_n s_0)}^{1} w^{-\rho}\, dw + o_p^{(n)}(1) \right) \\
&\qquad\qquad\qquad\qquad \text{(use a change of variables, } w = \frac{u}{s_0}\text{)} \\
&= \gamma + s_0^{-1} k_n^{-\frac{1}{2}} \left(\gamma(1-p\gamma)^2 \int_0^1 w^{-1} \int_0^1 t^{-p\gamma} \left\{ t^{-1} e(ut) - e(u) \right\} dt\, dw \right. \\
&\quad \left. + \frac{(1-p\gamma)\phi s_0^{1-\rho}}{(1-p\gamma-\rho)(1-\rho)} + o_p^{(n)}(1) \right),
\end{aligned}
$$

since $m_n/(k_n s_0) \to 0$. Furthermore, by (2.45), we can similarly show that

$$
\begin{aligned}
&\frac{1}{p(\lceil s_0 k_n \rfloor - m_n)} \left| \sum_{j=m_n+1}^{\lceil s_0 k_n \rfloor} j \left\{ \log\left(1+\frac{1}{j}\right) - \log\left(1 + \frac{1}{\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p}\right) \right\} \right. \\
&\quad \left. - \sum_{j=m_n+1}^{\lceil s_0 k_n \rfloor} j \left\{ \frac{1}{j} - \frac{1}{\sum_{i=1}^{j} \left(\frac{X_{n-i+1,n}}{X_{n-j,n}}\right)^p} \right\} \right| \\
&= o_p^{(n)}\left((s_0 k_n)^{-\frac{1}{2}} \right).
\end{aligned}
$$

Hence, for all $\nu \in (0, 1)$ and $s_0 \in [\nu, 1]$ uniformly, we have

$$\begin{aligned}
&\hat{\gamma}_{m_n,p}^{\text{EMP}}(\lceil s_0 k_n \rfloor) \\
&= \gamma + s_0^{-1} k_n^{-\frac{1}{2}} \left(\gamma(1-p\gamma)^2 \int_0^1 w^{-1} \int_0^1 t^{-p\gamma} \left\{ t^{-1} e(s_0 wt) - e(s_0 w) \right\} dt\, dw \right. \\
&\quad \left. + \frac{(1-p\gamma)\phi s_0^{1-\rho}}{(1-p\gamma-\rho)(1-\rho)} + o_p^{(n)}(1) \right) \\
&=: \gamma + k_n^{-\frac{1}{2}} R(s_0) + k_n^{-\frac{1}{2}} b_{\text{EMP}} \phi s_0^{-\rho} + o_p^{(n)}\left(k_n^{-\frac{1}{2}}\right),
\end{aligned}$$

where

$$\begin{aligned}
R(s) &:= s^{-1}\gamma(1-p\gamma)^2 \int_0^1 w^{-1} \int_0^1 t^{-p\gamma} \left\{ t^{-1} e(swt) - e(sw) \right\} dt\, dw, \\
b_{\text{EMP}} &:= \frac{1-p\gamma}{(1-p\gamma-\rho)(1-\rho)}.
\end{aligned}$$

By a standard diagonal argument, there exists a sequence ν_n satisfying $\nu_n \downarrow 0$ and $m_n/(\nu_n k_n) \downarrow 0$, such that

$$\sup_{\nu_n \le s_0 \le 1} \left| k_n^{\frac{1}{2}} \left(\hat{\gamma}_{m_n,p}^{\text{EMP}}(\lceil s_0 k_n \rfloor) - \gamma \right) - R(s_0) - b_{\text{EMP}} \phi s_0^{-\rho} \right| \to_p 0.$$

Denoting $j_n^{(1)} := \lceil \nu_n k_n \rfloor$, we have

$$\begin{aligned}
&\left(\log \frac{k_n}{j_n^{(1)}} \right)^{-1} \sum_{i=j_n^{(1)}}^{k_n} \left(\hat{\gamma}_{m_n,p}^{\text{EMP}}(i) - \hat{\gamma}_{m_n,p}^{\text{EMP}}(k_n) \right)^2 \\
&\xrightarrow{p} \left(\log \frac{k_n}{j_n^{(1)}} \right)^{-1} \int_{j_n^{(1)}/k_n}^1 (R(s) - R(1))^2\, ds \\
&\quad + \left(\log \frac{k_n}{j_n^{(1)}} \right)^{-1} b_{\text{EMP}}^2 \phi^2 \int_{j_n^{(1)}/k_n}^1 \left(s^{-\rho} - 1 \right)^2 ds \\
&\quad + 2 \left(\log \frac{k_n}{j_n^{(1)}} \right)^{-1} \int_{j_n^{(1)}/k_n}^1 \left(s^{-\rho} - 1 \right) (R(s) - R(1))\, ds \\
&=: D_n^{(1)} + D_n^{(2)} + D_n^{(3)}.
\end{aligned}$$

Define

$$\tilde{R}(x) := e^{\frac{x}{2}} R(e^x), \qquad x \in (-\infty, 0].$$

Based on the homogeneity of covariance function $r(x, y)$: $r(\lambda x, \lambda y) = \lambda r(x, y)$ for $\lambda, x, y \in [0, 1]$, we have

$$
\begin{aligned}
&Var\left(\tilde{R}(x), \tilde{R}(y)\right) \\
&= e^{-\frac{x+y}{2}} \gamma^2 (1-p\gamma)^4 Cov\left(\int_0^1 w^{-1} \int_0^1 t^{-p\gamma} \left\{t^{-1} e(e^x wt) - e(e^x w)\right\} dt\, dw,\right. \\
&\left. \int_0^1 v^{-1} \int_0^1 s^{-p\gamma} \left\{s^{-1} e(e^y vs) - e(e^y v)\right\} ds\, dv\right) \\
&= e^{-\frac{y-x}{2}} \gamma^2 (1-p\gamma)^4 \left\{\int_0^1\int_0^1\int_0^1\int_0^1 (wv)^{-1}(st)^{-p\gamma-1} r\left(wt, e^{y-x} vs\right) dt\, ds\, dw\, dv\right. \\
&-2\int_0^1\int_0^1\int_0^1\int_0^1 (wv)^{-1} s^{-p\gamma-1} t^{-p\gamma} r\left(w, e^{y-x} vs\right) dt\, ds\, dw\, dv \\
&\left.+\int_0^1\int_0^1\int_0^1\int_0^1 (wv)^{-1}(st)^{-p\gamma} r\left(w, e^{y-x} v\right) dt\, ds\, dw\, dv\right\} \\
&= e^{-\frac{y-x}{2}} \left\{\gamma^2 (1-p\gamma)^4 \int_0^1\int_0^1\int_0^1\int_0^1 (wv)^{-1}(st)^{-p\gamma-1} r\left(wt, e^{y-x} vs\right) dt\, ds\, dw\, dv\right. \\
&-2\gamma^2(1-p\gamma)^3 \int_0^1\int_0^1\int_0^1 (wv)^{-1} s^{-p\gamma-1} r\left(w, e^{y-x} vs\right) ds\, dw\, dv \\
&\left.+\gamma^2(1-p\gamma)^2 \int_0^1\int_0^1 (wv)^{-1} r\left(w, e^{y-x} v\right) dw\, dv\right\},
\end{aligned} \tag{2.53}
$$

depending only on $(y-x)$. Hence, $\tilde{R}(x)$ is a strictly stationary centered Gaussian process, with variance

$$
Var\left(\tilde{R}(x)\right) = \sigma^2_{\mathrm{EMP},p,\gamma,r}, \qquad \forall\, x \in [0, +\infty).
$$

Referring to the proof of Drees ([12], Theorem 2.3), we have

$$
\begin{aligned}
D_n^{(1)} &= \left(\log \frac{k_n}{j_n^{(1)}}\right)^{-1} \int_{j_n^{(1)}/k_n}^1 \left(R(s) - R(1)\right)^2 ds \\
&= \left(\log \frac{k_n}{j_n^{(1)}}\right)^{-1} \int_{\log(j_n^{(1)}/k_n)}^0 \left(\tilde{R}(u) - e^{\frac{u}{2}} \tilde{R}(0)\right)^2 du \\
&\to \left(\log \frac{k_n}{j_n^{(1)}}\right)^{-1} \int_{\log(j_n^{(1)}/k_n)}^0 \tilde{R}^2(u)\, du,
\end{aligned} \tag{2.54}
$$

because

$$
\int_{\log(j_n^{(1)}/k_n)}^0 \left(e^{\frac{u}{2}} \tilde{R}(0)\right)^2 du = O(1).
$$

Moreover, the ergodic theorem in [7, Sect. 5.5] gives rise to

$$
\left(\log \frac{k_n}{j_n^{(1)}}\right)^{-1} \int_{\log(j_n^{(1)}/k_n)}^0 \tilde{R}^2(u)\, du \to E\tilde{R}^2(0) = \sigma^2_{\mathrm{EMP},p,\gamma,r} \quad \text{almost surely.}
$$

Therefore,

$$D_n^{(1)} \to_p \sigma^2_{\mathrm{EMP},p,\gamma,r} \qquad \text{as} \quad j_n^{(1)}/k_n \to 0. \tag{2.55}$$

For the term $D_n^{(3)}$, based on the ergodic theorem, we have

$$\begin{aligned} D_n^{(3)} &= 2\left(\log\frac{k_n}{j_n^{(1)}}\right)^{-1}\int_{j_n^{(1)}/k_n}^{1}\left(s^{-\rho}-1\right)(R(s)-R(1))\,ds \\ &= 2\left(\log\frac{k_n}{j_n^{(1)}}\right)^{-1}\int_{\log(j_n^{(1)}/k_n)}^{0}\left(e^{-u\rho}-1\right)\left(e^{\frac{u}{2}}\tilde{R}(u)-e^{u}\tilde{R}(0)\right)du \\ &\to 0 \quad \text{almost surely.} \end{aligned} \tag{2.56}$$

Furthermore, notice that

$$D_n^{(2)} \to 0 \quad \text{as} \quad j_n^{(1)}/k_n \to 0. \tag{2.57}$$

So, combining (2.55)–(2.57), we have

$$\hat{\sigma}^2_{\mathrm{EMP},1} := \left(\log\frac{k_n}{j_n^{(1)}}\right)^{-1}\sum_{i=j_n^{(1)}}^{k_n}\left(\hat{\gamma}^{\mathrm{EMP}}_{m_n,p}(i)-\hat{\gamma}^{\mathrm{EMP}}_{m_n,p}(k_n)\right)^2 \to_p \sigma^2_{\mathrm{EMP},p,\gamma,r}.$$

Next, we verify the results for the MOP estimator. By (2.43), we have, for all $\nu \in (0,1)$ and $s_0 \in [\nu, 1]$ uniformly,

$$\begin{aligned} \hat{\gamma}_p^{\mathrm{MOP}}(\lceil s_0 k_n \rfloor) &= \gamma + k_n^{-\frac{1}{2}}\left(\gamma(1-p\gamma)^2 s_0^{-1}\int_0^1 t^{-p\gamma}\left(t^{-1}e(s_0t)-e(s_0)\right)dt\right. \\ &\qquad \left. + \frac{(1-p\gamma)\phi s_0^{-\rho}}{1-p\gamma-\rho} + o_p^{(n)}(1)\right) \\ &=: \gamma + k_n^{-\frac{1}{2}}P(s_0) + k_n^{-\frac{1}{2}}b_{\mathrm{MOP}}\phi s_0^{-\rho} + o_p^{(n)}\left(k_n^{-\frac{1}{2}}\right), \end{aligned}$$

where

$$\begin{aligned} P(s) &:= s^{-1}\gamma(1-p\gamma)^2\int_0^1 t^{-p\gamma}\left\{t^{-1}e(st)-e(s)\right\}dt, \\ b_{\mathrm{MOP}} &:= \frac{1-p\gamma}{1-p\gamma-\rho}. \end{aligned}$$

Similarly, there exists a sequence $\tilde{\nu}_n$ satisfying $\tilde{\nu}_n \downarrow 0$, such that

$$\sup_{\tilde{\nu}_n \le s_0 \le 1} \left| k_n^{\frac{1}{2}} \left(\hat{\gamma}_p^{\text{MOP}}(\lceil s_0 k_n \rfloor) - \gamma \right) - P(s_0) - b_{\text{MOP}} \phi s_0^{-\rho} \right| \to_p 0.$$

Define

$$\tilde{P}(x) := e^{\frac{x}{2}} P\left(e^x\right), \quad x \in (-\infty, 0].$$

Similar as (2.53), we can show that

$$\begin{aligned} & Cov\left(\tilde{P}(x), \tilde{P}(y)\right) \\ &= e^{-\frac{y-x}{2}} \left\{ \gamma^2 (1-p\gamma)^4 \int_0^1 \int_0^1 (st)^{-p\gamma-1} r(s, e^{y-x} t)\, ds\, dt \right. \\ & \left. -2\gamma^2 (1-p\gamma)^3 \int_0^1 t^{-p\gamma-1} r(t, e^{y-x})\, dt + \gamma^2 (1-p\gamma)^2\, r(1, e^{y-x}) \right\}, \end{aligned}$$

which is only determined by $(y-x)$. Hence, $\tilde{P}(x)$ is also a strictly stationary centered Gaussian process with variance

$$Var\left(\tilde{P}(x)\right) = \sigma^2_{\text{MOP},p,\gamma,r}, \qquad \forall\, x \in [0, +\infty).$$

The remaining of the proof is omitted; see the proof of the EMP estimator. Finally, denoting $j_n^{(2)} := \tilde{\nu}_n k_n$, we can show that

$$\hat{\sigma}^2_{\text{MOP},1} := \left(\log \frac{k_n}{j_n^{(2)}} \right)^{-1} \sum_{i=j_n^{(2)}}^{k_n} \left(\hat{\gamma}_p^{\text{MOP}}(i) - \hat{\gamma}_p^{\text{MOP}}(k_n) \right)^2 \to_p \sigma^2_{\text{MOP},p,\gamma,r}.$$

2.6.6 Proof of Theorem 2.5

Since $2(1-2\rho)^{-1} > 0$, we just need to minimize the term $(\sigma^2_{\centerdot})^{-\rho} b_{\centerdot}$. Note that

$$\begin{aligned} (\sigma^2_{\text{EMP}})^{-\rho} b_{\text{EMP}} &= \frac{2^{-\rho}\gamma^{-2\rho}(1-p\gamma)^{-\rho}}{(1-2p\gamma)^{-\rho}} \frac{1-p\gamma}{(1-p\gamma-\rho)(1-\rho)} \\ &= \frac{2^{-\rho}\gamma^{-2\rho}}{1-\rho} \frac{(1-p\gamma)^{1-\rho}}{(1-p\gamma-\rho)(1-2p\gamma)^{-\rho}} \\ &:= \frac{2^{-\rho}\gamma^{-2\rho}}{1-\rho} D(p). \end{aligned}$$

Taking the derivative of $D(p)$, we have

$$\begin{aligned}
&\frac{dD(p)}{dp}\\
&=\frac{-\gamma(1-\rho)(1-p\gamma)^{-\rho}(1-p\gamma-\rho)(1-2p\gamma)^{-\rho}+\gamma(1-p\gamma)^{1-\rho}(1-2p\gamma)^{-\rho}}{(1-p\gamma-\rho)^2(1-2p\gamma)^{-2\rho}}\\
&+\frac{-2\gamma\rho(1-2p\gamma)^{-1-\rho}(1-p\gamma-\rho)(1-p\gamma)^{1-\rho}}{(1-p\gamma-\rho)^2(1-2p\gamma)^{-2\rho}}\\
&=\frac{-\gamma(1-p\gamma)^{-\rho}(1-2p\gamma)^{-\rho-1}(1-\rho)(1-p\gamma-\rho)(1-2p\gamma)}{(1-p\gamma-\rho)^2(1-2p\gamma)^{-2\rho}}\\
&+\frac{\gamma(1-p\gamma)^{-\rho}(1-2p\gamma)^{-\rho-1}\{(1-p\gamma)(1-2p\gamma)-2\rho(1-p\gamma-\rho)(1-p\gamma)\}}{(1-p\gamma-\rho)^2(1-2p\gamma)^{-2\rho}}\\
&=\frac{\gamma(1-p\gamma)^{-\rho}(1-2p\gamma)^{-\rho-1}(\rho^2-p\gamma\rho)}{(1-p\gamma-\rho)^2(1-2p\gamma)^{-2\rho}}.
\end{aligned}$$

We see that $dD(p)/dp < 0$ if $p < \rho\gamma^{-1}$, and $dD(p)/dp > 0$ if $p > \rho\gamma^{-1}$. Hence, the EMP estimator attains the minimum LMSE at $p^*_{\text{EMP}} := \rho\gamma^{-1}$.

Acknowledgements The authors would like to thank the Editors for their kind invitation to contribute to this important treatise in honor of Professor Taniguchi, and in particular, to an anonymous referee for meticulous and critical readings and suggestions, which improve this manuscript tremendously. Research supported in part by grants from the Research Grants Council of Hong Kong, HKSAR-RGC-GRF Nos 14308218 and 14307921 (NH Chan), and Nos 14301618, 14301920, and 14307221 (T Sit).

References

1. Alves, M. F., Gomes, M. I. and De Haan, L. (2003). A new class of semi-parametric estimators of the second order parameter. *Portugaliae Mathematica* **60** 193–214.
2. Beirlant, J., Vynckier, P. and Teugels, J. L. (1996a). Excess functions and estimation of the extreme-value index. *Bernoulli* **2** 293–318.
3. Beirlant, J., Vynckier, P. and Teugels, J. L. (1996b). Tail index estimation, Pareto quantile plots regression diagnostics. *Journal of the American statistical Association* **91** 1659–1667.
4. Bingham, N., Goldie, C. and Teugels, J. (1984). *Regular Variation*. Cambridge University Press.
5. Brilhante, M. F., Gomes, M. I. and Pestana, D. (2013). A simple generalisation of the Hill estimator. *Computational Statistics and Data Analysis* **57** 518–535.
6. Ciuperca, G. and Mercadier, C. (2010). Semi-parametric estimation for heavy tailed distributions. *Extremes* **13** 55–87.
7. Cramér, H. and Leadbetter, M. R. (1967). *Stationary and Related Stochastic Processes: Sample Function Properties and Their Applications*. John Wiley & Sons, Inc.
8. De Haan, L. and Ferreira, A. (2007). *Extreme Value Theory: An Introduction*. Springer Science & Business Media.
9. De Haan, L. and Peng, L. (1998). Comparison of tail index estimators. *Statistica Neerlandica* **52** 60–70.

10. Dekkers, A. L. and De Haan, L. (1993). Optimal choice of sample fraction in extreme-value estimation. *Journal of Multivariate Analysis* **47** 173–195.
11. Drees, H. (2000). Weighted approximations of tail processes for β-mixing random variables. *Annals of Applied Probability* **10** 1274–1301.
12. Drees, H. (2003). Extreme quantile estimation for dependent data, with applications to finance. *Bernoulli* **9** 617–657.
13. Goegebeur, Y., Beirlant, J. and De Wet, T. (2008). Linking Pareto-tail kernel goodness-of-fit statistics with tail index at optimal threshold and second order estimation. *Revstat* **6** 51–69.
14. Gomes, M. I. and Henriques- Rodrigues, L. (2010). Comparison at optimal levels of classical tail index estimators: a challenge for reduced-bias estimation? *Discussiones Mathematicae Probability and Statistics* **30** 35–51.
15. Gomes, M. I. and Martins, M. J. (2001). Generalizations of the Hill estimator–asymptotic versus finite sample behaviour. *Journal of Statistical Planning and Inference* **93** 161–180.
16. Gomes, M. I. and Neves, C. (2008). Asymptotic comparison of the mixed moment and classical extreme value index estimators. *Statistics and Probability Letters* **78** 643–653.
17. Gomes, M. I. and Oliveira, O. (2001). The bootstrap methodology in statistics of extremes – choice of the optimal sample fraction. *Extremes* **4** 331–358.
18. Gomes, M. I. and Pestana, D. (2007). A sturdy reduced-bias extreme quantile (var) estimator. *Journal of the American Statistical Association* **102** 280–292.
19. Resnick, S. and Starica, C. (1998). Tail index estimation for dependent data. *The Annals of Applied Probability* **8** 1156–1183.

Chapter 3
Exclusive Topic Model

Hao Lei, Kailiang Liu, and Ying Chen

Abstract Digital documents are generated, disseminated, and disclosed in books, research papers, newspapers, online feedback, and other content containing large amounts of information, for which discovering topics becomes important but challenging. In the current field of topic modeling, there are limited techniques available to deal with (1) the predominance of frequently occurring words in the estimated topics and (2) the overlap of commonly used words in different topics. We propose exclusive topic modeling (ETM) to identify field-specific keywords that typically occur less frequently but are important for representing certain topics, and to provide well-structured exclusive terms for topics within each topic. Specifically, we impose a weighted LASSO penalty to automatically reduce the dominance of frequently occurring but less relevant words and a pairwise Kullback–Leibler divergence penalty to achieve topic separation. Numerical studies show that the ETM can detect field-specific keywords and provide exclusive topics, which is more meaningful for interpretation and topic detection than other models such as the latent Dirichlet allocation model.

3.1 Introduction

Digital documents are generated, disseminated and disclosed in books, research papers, newspapers, online feedback, and other contents containing large amounts of information, for which topic discovery becomes very important. Given that the

H. Lei
National University of Singapore, Block S16, Level 7, 6 Science Drive 2, Singapore 117546, Singapore
e-mail: leihao@u.nus.edu

K. Liu · Y. Chen (✉)
National University of Singapore, Block S17, Level 4, 10 Lower Kent Ridge Road, Singapore 119076, Singapore
e-mail: matcheny@nus.edu.sg

K. Liu
e-mail: liukl@u.nus.edu

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_3

goal is to find hidden semantic structures in documents, this requires probabilistic topic models that group similar phrases by learning large amounts of raw text in an unsupervised manner. Unsupervised machine learning and natural language processing algorithms are used for text data analysis, including topic modeling techniques that have become popular in recent years.

Probabilistic topic modeling aims to uncover the latent structure of corpus from large-scale words to topics, without any prior topic annotations [11]. It has been widely adopted for information retrieval in various fields and its application goes beyond just organizing textual data such as natural scene categories in computer vision [17], genomic data in bioinformatics [15, 24, 26], public opinion on urban issues in information management [27], market competition [50], service quality [23], customer preferences [44] in marketing, financial analyst reports [22], 10-K forms [5], and algorithmic trading engines [32, 42] in finance, to name just a few. Despite its popularity, there are two well-known technical challenges in topic modeling: (1) the estimated topics often overlap with commonly used words and are therefore semantically similar, and (2) the predominance of frequently occurring words in the estimated topics. The former makes interpretation uneasy, while the latter poses computational challenges in terms of estimation and memory.

To alleviate the overlapping problem, there are two popular strategies: restrict the number of topics or induce latent variables. Several approaches have been proposed for selecting the number of topics, including, e.g., cross-validation using perplexity [8–10], topic coherence [29, 31, 41], nested Chinese restaurant process as a prior for the number of topics [18]. Word embedding information has also been used to improve topic quality [16, 38, 47], where some latent variables are introduced to increase flexibility [34]. Yet, these methods still yield many topics, where the overlapping problem exists.

Previous works address the predominance of frequent words by incorporating syntactic information [14, 19], using the asymmetric priors [45], and the "garbage collector" approach [39, 40]. Both the Hidden Markov Model-Latent Dirichlet Allocation (HMM-LDA) model [19] and the syntactic topic model [14] incorporate syntactic information into the thematic topic model. The effect of four different combinations of symmetric and asymmetric priors in the Latent Dirichlet Allocation (LDA) model on the estimated topics is investigated by Wallach et al. [45].

We propose exclusive topic modeling (ETM) to identify field-specific keywords that typically occur less frequently but are important for representing certain topics and to provide well-structured exclusive terms for topics within each topic. Specifically, we impose a weighted LASSO penalty to handle corpus-specific common words, where the weights are associated with their commonality in the underlying corpus. The more common a word is, the larger the penalty and the smaller the assigned probability are. Furthermore, the pairwise Kullback–Leibler (KL) divergence penalty is able to separate the overlapped topics by maximizing a linear combination of the posterior according to the Evidence Lower BOund (ELBO) [12]. We demonstrate the effectiveness of the proposed model in three simulation studies, where the ETM outperforms LDA in each case and recovers the ground truth topics. We also apply the proposed method to the benchmark dataset of NIPS. The

NIPS benchmarks and datasets are well-known in the machine learning community. The benchmark dataset has been used in many previous works on probabilistic topic modeling [3, 4, 6, 20, 36], as well as in network studies [21, 28, 37]. Compared with the LDA, the ETM assigns lower weights to frequently occurring words and deliver exclusive topics, making the topics easier to interpret. Our results show that the topic coherence score improves by 22% and 10% on average for the ETM with weighted LASSO penalty and pairwise KL divergence penalty, respectively.

The rest of the paper is structured as follows. Section 3.2 presents the details of the proposed method and the algorithms to estimate the topics. In Sect. 3.3, we conduct three simulation studies to demonstrate the effectiveness of the proposed method in tackling the two common issues in topic modeling. In Sect. 3.4, we apply the proposed method to the public NIPS dataset. The results show that the proposed method improves topic interpretability and coherence scores. We conclude the paper in Sect. 3.5.

3.2 Method

In this section, we present the details of the proposed methods, including the setting of the ETM, the weighted LASSO penalty, and the pairwise KL divergence penalty. We discuss the corresponding effects of the two penalties on topic estimation and derivations. In the ETM framework, the objective function is no longer convex. We use an algorithm combining gradient descent and Hessian descent to find updated equations.

3.2.1 ETM

The model setup is as follows. Given a corpus C, we assume it contains K topics. Every topic η_k is a multinomial distribution on the vocabulary. Every document d contains one or more topics. The topic proportion in each document is governed by the local latent parameter document-topic θ, which has a Dirichlet prior with hyperparameter ζ. Every word in document d is generated from the contained topics as follows:

- for every document $d \in C$, its topic proportion parameter θ is generated from a Dirichlet distribution, i.e., $\theta \sim \text{Dirichlet}(\zeta)$.
- for every word in the document d,
  - a topic Z is first generated from the multinomial distribution with parameter θ, i.e., $Z \sim \text{Multinomial}(1, \theta)$.
  - a word W is then generated from the multinomial distribution with parameter η_Z, i.e., $W \sim \text{Multinomial}(1, \eta_Z)$.

Although the setting is the same as [10], the latent parameters in the ETM are estimated by maximizing the following penalized posterior:

$$\log p(\theta, Z, \zeta, \eta|W) - \sum_{i=1}^{K}\sum_{j=1}^{V} \mu_i m_{ij}|\eta_{ij}| + \sum_{i=1, l\neq i}^{K} \nu_{il} D_{KL}(\eta_i||\eta_l), \tag{3.1}$$

where $p(\theta, Z, \zeta, \eta|W)$ is

$$p(\theta, Z, \zeta, \eta|W) = \frac{p(\theta, Z, \zeta, \eta, W)}{p(W)} = \frac{p(\theta, Z, \zeta, \eta, W)}{\int p(\theta, Z, \zeta, \eta, W)\, d\theta\, dZ\, d\zeta\, d\eta}$$

and

$$\begin{aligned} p(\theta, Z, \zeta, \eta, W) &= \prod_{d=1}^{D} p(\theta_d|\zeta) \prod_{n=1}^{N_d} p(z_n|\theta_d) p(w_n|z_n, \eta) \\ &= \prod_{d=1}^{D} \frac{\Gamma(\sum_{i=1}^{K} \zeta_i)}{\prod_{i=1}^{K} \Gamma(\zeta_i)} \theta_{di}^{\zeta_i - 1} \prod_{n=1}^{N_d} \prod_{i=1}^{K} \theta_{di}^{1_{\{z_n=i\}}} \prod_{j=1}^{V} \eta_{ij}^{1_{\{w_n=v_j\}} 1_{\{z_n=i\}}}. \end{aligned}$$

The $D, K, and V$ represent the total number of documents, topics, and words, respectively, in the corpus. N_d is the total number of words in document d. $\Gamma(\cdot)$ is the gamma function. The second term in Eq. (3.1) is the weighted LASSO penalty, in which μ_i is the penalty weight for topic i, m_{ij} represents the weight for word j in topic i and is known. The third term in Eq. (3.1) is the pairwise KL divergence penalty, in which $D_{KL}(\eta_i||\eta_l) = \sum_{j=1}^{V} \eta_{ij}(\log \eta_{ij} - \log \eta_{lj})$ represents the KL divergence between topic i and l and ν_{il} is the corresponding penalty weight.

Unfortunately, the posterior is intractable to compute. Instead, a "variational EM algorithm" is used to maximize the ELBO [10, 12], denoted as $L(\zeta, \eta, \gamma, \phi) = \sum_{d=1}^{D} L_d(\zeta, \eta, \gamma_d, \phi_d)$, where L_d is the ELBO for document d

$$\begin{aligned} L_d(\zeta, \eta, \gamma_d, \phi_d) &= E_q[\log p(\theta_d, Z_d, W_d|\zeta, \eta)] - E_q[\log q(\theta_d, Z_d|\gamma_d, \phi_d)] \\ &\leq \log p(W_d|\zeta, \eta), \end{aligned}$$

where $p(\cdot)$ is the density function derived from the LDA and $q(\theta_d, Z_d|\gamma_d, \phi_d)$ is the mean-field variational distribution

$$q(\theta_d, Z_d|\gamma_d, \phi_d) = q(\theta_d|\gamma_d) \prod_{n=1}^{N_d} q(Z_{dn}|\phi_{dn}),$$

where $q(\theta_d|\gamma_d) \sim \text{Dirichlet}(\gamma_\text{d})$, and $q(Z_{dn}|\phi_{dn}) \sim \text{Multinomial}(1, \phi_\text{dn})$. E_q represents the expectation under the variational distribution. The inequality is a result by applying Jensen's inequality. The data W_d provides more evidence to our prior belief, hence it is named as the ELBO. The exact form of L_d is as follows [10]:

$$
\begin{aligned}
L_d(\zeta, \eta, \gamma, \phi) = \log \Gamma(\sum_{i=1}^{K} \zeta_i) - \sum_{i=1}^{K} \log(\zeta_i) + \sum_{i=1}^{K} (\zeta_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{i=1}^{K} \gamma_i)) \\
+ \sum_{n=1}^{N_d} \sum_{i=1}^{K} \phi_{ni} (\Psi(\gamma_i) - \Psi(\sum_{i=1}^{K} \gamma_i)) + \sum_{n=1}^{N_d} \sum_{i=1}^{K} \sum_{j=1}^{V} \phi_{ni} w_{dn}^{j} \log \eta_{ij} \\
- \log \Gamma(\sum_{i=1}^{K} \gamma_i) + \sum_{i=1}^{K} \log \Gamma(\gamma_i) \\
- \sum_{i=1}^{K} (\gamma_i - 1)(\Psi(\gamma_i) - \Psi(\sum_{i=1}^{K} \gamma_i)) - \sum_{n=1}^{N_d} \sum_{i=1}^{K} \phi_{ni} \log \phi_{ni}.
\end{aligned}
$$

Therefore, in the actual optimization, we maximize the following penalized ELBO:

$$
L(\zeta, \eta, \gamma, \phi) - \sum_{i=1}^{K} \sum_{j=1}^{V} \mu_i m_{ij} |\eta_{ij}| + \sum_{i=1, l \neq i}^{K} \nu_{il} D_{KL}(\eta_i || \eta_l). \tag{3.2}
$$

Equation (3.2) is maximized following an "EM"-like procedure. In the E-step, the ELBO is maximized w.r.t. the local variational parameter ϕ, γ for every document, conditional on the global latent parameter η, ζ. Since the penalties do not contain any local variational parameters, the updating equations will be the same as that of the LDA.

$$
\begin{aligned}
\phi_{ni} &\propto \eta_{iw_n} \exp E_q[\log(\theta_i)|\gamma], \\
\gamma_i &= \zeta_i + \sum_{n=1}^{N} \phi_{ni},
\end{aligned} \tag{3.3}
$$

where

$$
\exp E_q[\log(\theta_i)|\gamma] = \Psi(\gamma_i) - \Psi(\sum_{l=1}^{K} \gamma_l)
$$

and $\Psi(\cdot)$ is the digamma function, i.e., the logarithmic derivative of the gamma function. Note that these are local variables and we omit the subscript d in $\phi_{dni}, \theta_{di}, \gamma_{di}$.

Then conditional on all the local latent variables ϕ, γ, the penalized ELBO is maximized w.r.t. the latent global parameter η, ζ. The global parameter ζ can be estimated using Newton's method. In practice, ζ is often assumed to be a symmetric Dirichlet parameter.

3.2.2 Only Weighted LASSO Penalty: $\nu = 0$

In this section, we consider the case where we only have the weighted LASSO penalty, i.e., $\nu = 0$:

$$\max \quad L(\zeta, \eta, \gamma, \phi) - \sum_{i=1}^{K} \sum_{j=1}^{V} \mu_i m_{ij} |\eta_{ij}| \tag{3.4}$$

subject to

$$\sum_{j=1}^{V} \eta_{ij} = 1, \forall i \in \{1, \ldots, K\},$$
$$\eta_{ij} \geq 0, \forall i, j,$$
$$\sum_{i=1}^{K} \gamma_{di} = 1, \forall d,$$

where K and V represent the number of topics and vocabulary size respectively and $\mu_i \geq 0$ is the penalty weight for topic i and is selected using cross-validation, η_{ij} represents the probability of the jth word in the topic i, m_{ij} is the weight for η_{ij} and is known in advance, reflecting the prior information about the topic word distribution. One possible candidate for the weight is the document frequency, i.e., $m_{ij} = df_j, \forall i, j$, where df_j is the number of documents containing the word j. The larger the document frequency for the word j, the larger the penalty. Consider an extreme case that the word j appears in every document of the corpus. Because it co-occurs with every other word, the LDA would assign a large probability to it in every topic. As a result, the word j contains little information to distinguish one topic from another. It's barely useful in the dimension reduction process, i.e., from word space to topic space. With the document frequency penalty, it will be penalized the most and result in a low probability in the topic distributions.

We emphasize that the penalized model is not constrained to only solving the frequent words dominance issue. Any weight reflecting the prior information about the topic distribution can be used to achieve the practitioners' goal. We give an illustration here and in the simulation study; see also Sect. 3.3.2. Often practitioners found the field-important words are not assigned large probabilities in the estimated topics. One possible reason is that they appear infrequently in the underlying corpus. But these keywords contain important information about the field and are crucial to distinguish one topic from another. Practitioners might prefer they appear in the top-T words for easy topic interpretation (in practice, T is usually set as 10 or 20). In this situation, practitioners can utilize our proposed model by assigning negative weights to these keywords and zero weights to all the other words. Section 3.3.2 is devoted to the situation.

To further understand the penalization, we rewrite the optimization problem (3.4) in the following equivalent form:

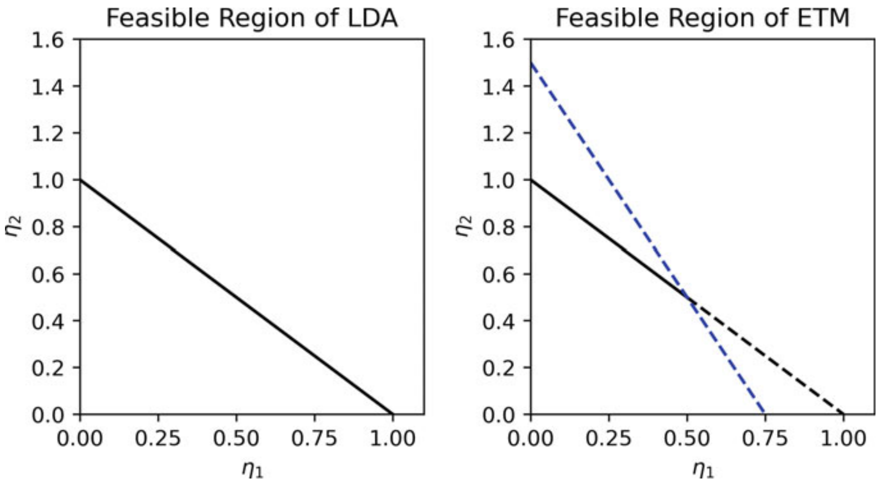


Fig. 3.1 The feasible region of topics of LDA (left) and ETM (right)

$$\max \quad L(\zeta, \eta, \gamma, \phi)$$

subject to

$$\sum_{j=1}^{V} m_{ij}|\eta_{ij}| \leq \nu_i, \forall i \in \{1, \ldots, K\},$$

$$\sum_{j=1}^{V} \eta_{ij} = 1, \forall i \in \{1, \ldots, K\},$$

$$\eta_{ij} \geq 0, \forall i, j,$$

$$\sum_{i=1}^{K} \gamma_{di} = 1, \forall d,$$

where $\nu_i > 0$ is a hyperparameter and there is a one-to-one correspondence between μ_i and ν_i. The penalty alters the feasible region.

Figure 3.1 plots the feasible regions of the topics of LDA and the df-weighted LASSO penalized LDA for a simple case of having two words w_1 and w_2. The document frequencies are 2 and 1 for the word w_1 and w_2, respectively. The black solid line in the left subplot represents the feasible region of LDA. The feasible region of the ETM is plotted in the right subplot. The blue dashed line is the penalty induced constraint line $2\eta_1 + \eta_2 = 1.5$. Due to the extra constraint, the feasible region is reduced to the upper-left black solid line. As a result, the feasible probability range η_1 is reduced to $(0, 0.5)$. A smaller weight will be assigned to the relatively more frequently appearing word w_1 in the estimated topic.

As mentioned in Sect. 3.2.1, the optimization is done using the variational inference and the local latent parameters are updated the same as the LDA, as given in Eq. (3.3). The global latent parameter η is estimated by maximizing the following equation:

$$-f(\eta) = \sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{i=1}^{K}\sum_{j=1}^{V}\phi_{dni}w_{dn}^{j}\log(\eta_{ij}) - \sum_{i=1}^{K}\sum_{j=1}^{V}\mu_i m_{ij}|\eta_{ij}|$$

subject to

$$\sum_{j=1}^{V}\eta_{ij} = 1, \forall i \in \{1, \dots, K\},$$
$$\eta_{ij} \geq 0, \forall i, j,$$

where the first term is by taking out all the terms containing η from the ELBO. The problem can be further reduced to K sub-optimization problems below. Due to $\log(\eta_{ij})$ in the target function, the constraint $\eta_{ij} \geq 0, \forall i, j$ can be ignored. The absolute value $|\eta_{ij}|$ in the target function equals η_{ij}. We rewrite the maximization as an equivalent minimization problem:

$$\min \quad f_i(\eta_i) = -\sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{j=1}^{V}\phi_{dni}w_{dn}^{j}\log(\eta_{ij}) + \sum_{j=1}^{V}\mu_i m_{ij}\eta_{ij} \tag{3.5}$$

subject to

$$\sum_{j=1}^{V}\eta_{ij} = 1.$$

We use the Newton method with equality constraints [13] to solve Eq. (3.5). The updating direction $\Delta\eta_i$ with a feasible starting point η_i^0 can be calculated using the equation

$$\begin{bmatrix} \nabla^2 f_i & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix}\begin{pmatrix} \Delta\eta_i \\ \alpha_i \end{pmatrix} = \begin{pmatrix} -\nabla f_i \\ 0 \end{pmatrix}.$$

The updating direction is

$$\Delta\eta_{ij} = \frac{-f'_{ij} - \alpha_i}{f''_{ij}} = \eta_{ij} - \frac{(\alpha_i + \mu_i m_{ij})\eta_{ij}^2}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^{j}}, \tag{3.6}$$

where f'_{ij} and f''_{ij} are the first and second partial derivatives of f_i with respect to η_j, and

Algorithm 1: Variational M-step

Result: Update the ith topic word distribution η_i
Initialize η_i with a feasible point;
Choose the stopping criteria ϵ and the line search parameter $\delta \in (0, 0.5)$, $\gamma \in (0, 1)$;
while *not reaching the maximum iteration* **do**
  Compute the feasible descent direction $\Delta\eta_i$ and Newton decrement $\lambda(\eta_i)$;
  if $\lambda(\eta_i)^2/2 \leq \epsilon$ **then**
    Stop the algorithm;
  else
    Find step size t by backtracking line search:
    Initialize the step size $t := 1$;
    while $f_i(\eta_i + t\Delta\eta_i) > f_i(\eta_i) + \delta t \nabla f_i^T \Delta\eta$ **do**
      $t := \gamma t$;
  $\eta_i := \eta_i + t\Delta\eta_i$;

$$\alpha_i = \frac{-\sum_{j=1}^{V} f'_{ij}/f''_{ij}}{\sum_{j=1}^{V} 1/f''_{ij}} = \frac{\sum_{j=1}^{V}\left(\eta_{ij} - \frac{\mu_i m_{ij}\eta_{ij}^2}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^j}\right)}{\sum_{j=1}^{V}\frac{\eta_{ij}^2}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^j}}.$$

The Newton decrement is

$$\lambda(\eta_i) = (\Delta\eta_i^T \nabla^2 f_i \Delta\eta_i)^{1/2} = \left(\sum_{j=1}^{V}\frac{(\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^j - (\alpha_i + \mu_i m_{ij})\eta_{ij})^2}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^j}\right)^{1/2}. \quad (3.7)$$

We use the backtracking line search [13] to estimate the step size. The complete step to estimate topics of the weighted LASSO penalized LDA is given in Algorithm 1.

3.2.3 Only Pairwise Kullback–Leibler Divergence Penalty: $\mu = 0$

Practitioners often find some estimated topics are "close" to each other, in the sense, they have similar semantic meaning and share several common words in their top-N words. It makes the topic interpretation and the analyzing steps following topic modeling, e.g., [25], difficult. In this section, we consider the case where we only have a pairwise KL divergence penalty, i.e., $\mu = 0$. The optimization takes the following form:

$$\max \quad L(\zeta, \eta, \gamma, \phi) + \sum_{i=1, l\neq i}^{K} \nu_{il} D_{KL}(\eta_i || \eta_l) \quad (3.8)$$

subject to

$$\sum_{j=1}^{V} \eta_{ij} = 1, \forall i \in \{1, \dots, K\},$$

$$\eta_{ij} \geq 0, \forall i, j,$$

$$\sum_{i=1}^{K} \gamma_{di} = 1, \forall d,$$

where $D_{KL}(\eta_i||\eta_l)$ is the KL divergence between topic i and l, i.e.,

$$D_{KL}(\eta_i||\eta_l) = \sum_{j=1}^{V} \eta_{ij} \log(\frac{\eta_{ij}}{\eta_{lj}}).$$

Since the penalty only involves the topic distribution η and thus only plays a role in the M-step. The E-step is the same as Eq. (3.3). For the M-step, we solve

$$\min \quad g(\eta) = -\sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{i=1}^{K}\sum_{j=1}^{V} \phi_{dni} w_{dn}^{j} \log(\eta_{ij}) - \sum_{i=1,l\neq i}^{K} \sum_{j=1}^{V} \nu_{il}\eta_{ij} \log(\frac{\eta_{ij}}{\eta_{lj}}) \tag{3.9}$$

subject to

$$\sum_{j=1}^{V} \eta_{ij} = 1, \forall i \in \{1, \dots, K\},$$

$$\eta_{ij} \geq 0, \forall i, j.$$

There are two differences between the current optimization and the optimization in Sect. 3.2.2: (1) the optimization can no longer be split into K sub-optimization problems, since the different topics are now intertwined by the penalty; (2) the objective function is no longer convex. For the first difference, we borrow the idea of coordinate descent and sequentially optimize one topic at a time while conditioning on all other topics. The advantage of this approach instead of updating all topics simultaneously is that it simplifies the constrained Newton updating equation involving the Hessian matrix. Under the conditional approach, the Hessian matrix for a particular topic $\nabla^2 g_i$ is diagonal. For the second difference, due to the non-convexity, the Hessian matrix may not be positive semi-definite. As a result, the Newton decrement could be a complex number. We use a combination of gradient descent and Hessian descent algorithm to solve the second issue [2, 30]. The Hessian descent is invoked when the Hessian is not positive semi-definite. The Hessian descent moves to a smaller value along the Newton direction.

We now show the update equations for the gradient descent step. For topic i, conditioning on all other topics, we solve

$$\min \quad g_i(\eta_i|\eta_l, l \neq i) = -\sum_{d=1}^{D}\sum_{n=1}^{N_d}\sum_{j=1}^{V}\phi_{dni} w_{dn}^j \log(\eta_{ij}) - \sum_{l\neq i}\sum_{j=1}^{V}\nu_{il}\eta_{ij}\log(\frac{\eta_{ij}}{\eta_{lj}}) - \sum_{l\neq i}\sum_{j=1}^{V}\nu_{li}\eta_{lj}\log(\frac{\eta_{lj}}{\eta_{ij}}) \tag{3.10}$$

subject to

$$\sum_{j=1}^{V}\eta_{ij} = 1.$$

The updating direction $\Delta\eta_i$ with a feasible starting point η_i^0 can be calculated using the following equation:

$$\begin{bmatrix} \nabla^2 g_i & \mathbf{1} \\ \mathbf{1}^T & 0 \end{bmatrix}\begin{pmatrix} \Delta\eta_i \\ \alpha_i \end{pmatrix} = \begin{pmatrix} -\nabla g_i \\ 0 \end{pmatrix}. \tag{3.11}$$

The updating direction is

$$\Delta\eta_{ij} = \frac{-g'_{ij} - \alpha_i}{g''_{ij}} = \frac{\eta_{ij}(\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^j - \sum_{l\neq i}\nu_{li}\eta_{lj}) + \eta_{ij}^2\left(\sum_{l\neq i}\nu_{il}(\log\eta_{ij} - \log\eta_{lj} + 1) - \alpha_i\right)}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^j - \sum_{l\neq i}\nu_{li}\eta_{lj} - \eta_{ij}\sum_{l\neq i}\nu_{il}}, \tag{3.12}$$

where g'_{ij} and g''_{ij} are the first and second partial derivatives of g_i with respect to η_j, and

$$\alpha_i = \frac{-\sum_{j=1}^{V} g'_{ij}/g''_{ij}}{\sum_{j=1}^{V} 1/g''_{ij}} = \frac{\sum_{j=1}^{V}\frac{\eta_{ij}(\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^j - \sum_{l\neq i}\nu_{li}\eta_{lj}) + \eta_{ij}^2\left(\sum_{l\neq i}\nu_{il}(\log\eta_{ij} - \log\eta_{lj}+1)\right)}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^j - \sum_{l\neq i}\nu_{li}\eta_{lj} - \eta_{ij}\sum_{l\neq i}\nu_{il}}}{\sum_{j=1}^{V}\frac{\eta_{ij}^2}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^j - \sum_{l\neq i}\nu_{li}\eta_{lj} - \eta_{ij}\sum_{l\neq i}\nu_{il}}}. \tag{3.13}$$

The Newton decrement is

$$\lambda(\eta_i) = (\Delta\eta_i^T\nabla^2 g_i\Delta\eta_i)^{1/2} = \left(\sum_{j=1}^{V}\frac{\left(\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^j - \sum_{l\neq i}\nu_{li}\eta_{lj} + \eta_{ij}\left(\sum_{l\neq i}\nu_{il}(\log\eta_{ij} - \log\eta_{lj} + 1) - \alpha_i\right)\right)^2}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^j - \sum_{l\neq i}\nu_{li}\eta_{lj} - \eta_{ij}\sum_{l\neq i}\nu_{il}}\right)^{1/2}.$$

Due to the non-convexity, $\Delta\eta_i^T\nabla^2 g_i\Delta\eta_i$ could be negative at some points. When it happens, the Hessian descent is invoked and finds a new position along $\Delta\eta_i$ with smaller values (see details in Algorithm 2).

To update all the topics, we optimize the topics sequentially and stops until an overall convergence measured by the Frobenius norm of the successive updates, as in Algorithm 2.

Algorithm 2: Variational M-step for the pairwise KL Divergence

Result: Update the topic word distribution η
Initialize the topic word distribution η^0 with a feasible point for every topic $\eta_i^0, i = 1, \dots, K$;
Choose the stopping criterion ϵ;
while $||\eta^{t+1} - \eta^t||_F > \epsilon$ **do**
  for *topic* $i, i = 1, \dots, K$ **do**
    if $\Delta\eta_i^T \nabla^2 f_i \Delta\eta_i \geq 0$ **then**
      Gradient Descent:
      update topic word distribution $\eta_i|\eta_j, j \neq i$ using Algorithm 1;
    else
      Hessian Descent:
      find step size h satisfying $\eta_i + h\Delta\eta_i > 0$ and $\eta_i - h\Delta\eta_i > 0$; **if** $g_i(\eta_i + h\Delta\eta_i) > g_i(\eta_i - h\Delta\eta_i)$ **then**
        $\eta_i := \eta_i - h\Delta\eta_i$;
      else
        $\eta_i := \eta_i + h\Delta\eta_i$;

3.2.4 Combination of Two Penalties

The combination of the two penalties in a single algorithm is straightforward. Algorithm 2 can be used for the combined penalties. Adding the weighted LASSO penalty changes the updating direction $\Delta\eta$ in Eq. (3.12) to

$$\Delta\eta_{ij} = \left(\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^{j} - \sum_{l\neq i}\nu_{li}\eta_{lj} - \eta_{ij}\sum_{l\neq i}\nu_{il}\right)^{-1}$$
$$\left(\eta_{ij}(\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^{j} - \sum_{l\neq i}\nu_{li}\eta_{lj}) + \eta_{ij}^2(\sum_{l\neq i}\nu_{il}(\log\eta_{ij} - \log\eta_{lj} + 1) - \mu_i m_{ij} - \alpha_i)\right),$$

and the Eq. (3.13) becomes

$$\alpha_i = \frac{\sum_{j=1}^{V}\frac{\eta_{ij}(\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^{j} - \sum_{l\neq i}\nu_{li}\eta_{lj}) + \eta_{ij}^2\left(\sum_{l\neq i}\nu_{il}(\log\eta_{ij} - \log\eta_{lj}+1) - \mu_i m_{ij}\right)}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^{j} - \sum_{l\neq i}\nu_{li}\eta_{lj} - \eta_{ij}\sum_{l\neq i}\nu_{il}}}{\sum_{j=1}^{V}\frac{\eta_{ij}^2}{\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^{j} - \sum_{l\neq i}\nu_{li}\eta_{lj} - \eta_{ij}\sum_{l\neq i}\nu_{il}}}$$

and the Newton decrement $\lambda(\eta)$ in the gradient descent, as it alters the first derivative by subtracting the weights.

$$\lambda(\eta_i) = \Big(\sum_{j=1}^{V}\Big(\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^{j} - \sum_{l\neq i}\nu_{li}\eta_{lj} - \eta_{ij}\sum_{l\neq i}\nu_{il}\Big)^{-1}$$
$$\Big(\sum_{d=1}^{D}\sum_{n=1}^{N_d}\phi_{dni}w_{dn}^{j} - \sum_{l\neq i}\nu_{li}\eta_{lj} + \eta_{ij}\big(\sum_{l\neq i}\nu_{il}(\log\eta_{ij} - \log\eta_{lj} + 1) - \mu_i m_{ij} - \alpha_i\big)\Big)^2\Big)^{1/2}.$$

3.2.5 *Dynamic Penalty Weight Implementation*

Both Algorithms 1 and 2 apply to the variational M-step. Recall that the ELBO is maximized by an iterative variational EM algorithm. The M-step depends on the E-step output, i.e., $\phi_{dni}w_{dn}^{j}$ in Eqs. (3.5) and (3.10). In each iteration, the E-step would possibly produce $\phi_{dni}w_{dn}^{j}$ with different magnitudes, especially at the beginning of the iterations. A penalty weight suitable for the current iteration may be too large (small) for the next iteration. Therefore, we reparametrize the penalty weight μ_i as $\nu_i * \max_j(\sum_{dn}\phi_{dni}w_{dn}^{j})$ in the final variational EM algorithm. The reparameterization makes the penalty similar scale as its ELBO part, and thus effective in every EM iteration.

3.3 Simulation

In this section, we use simulated data to demonstrate the effectiveness of the proposed the ETM. In Sect. 3.3.1, we simulate the situation where the corpus contains several frequently appearing words. Compared to the LDA, the ETM effectively avoids the frequent word dominance issue in the estimated topics. In Sect. 3.3.2, we simulate the case where field-important words are rare in the underlying corpus. The ETM is able to reveal the importance of these words and recover the true distribution by using negative weights for these words and zero weights for all other words, while the LDA cannot. In Sect. 3.3.3, we simulate a "close" estimated topic situation, by adding several frequently occurring words. The LDA topics share these frequent words in their corresponding top T words. When practitioners interpret topics based on these top T words, they might misinterpret them as the same topic. The ETM is able to separate the topics while restoring the true topics.

3.3.1 Case 1: Corpus-Specific Common Words

The setup is as follows. The number of topics K is 2 and the prior $\zeta = 0.1$. The topic word distribution η is randomly drawn from a Dirichlet prior $\eta \sim \text{Dirichlet}(0.1 * \mathbb{1})$, where $\mathbb{1}$ is a 300×1 vector of 1s. Then, we use the word generation process of the LDA to generate the words for a corpus of 500 documents. During the word generation, we set an upper bound on the maximum number of words in each document to 100. Some words do not appear in the corpus because they have been assigned low topic word probabilities. Therefore, the number of words generated is 202. We further assume that the corpus contains 3 corpus-specific common words 301, 302, and 303 and each of them randomly appears in 50% of the documents. The appearance frequency in a document is 3. These words are then added to the generated corpus. In total, we have 205 words (202 generated words + 3 manually inserted words) in our final corpus.

We apply the LDA and the ETM ($\nu = 0$) to the simulated corpus. The weights are the document frequencies of all words, scaled to a maximum of 100. The penalty weight is selected to be $\mu = 0.3$. In practice, the estimated topics are interpreted using their corresponding top T words. We list the top 10 words of the true topics, the LDA estimated topics, and the ETM topics in Table 3.1. The corpus-specific common words 301, 302, and 303 are assigned high probabilities in the LDA and they appear in the top 10 words. The probabilities of these 3 words in the ETM are about half of those in the LDA (further reduction is achievable with a larger penalty). The high probabilities assigned to these frequently appearing words not only complicate topic interpretation, but also distort the document-topic frequencies θ, thus reducing the accuracy of information retrieval.

We repeat the above procedure 1000 times and record the number of corpus-specific common words that occur in the top 10 words, as well as the ratio of the average probabilities of these three words in the ETM and those in the LDA. The summary statistics are given in Table 3.2. Row 1 is the summary statistics for the number of common words occurring in the top 10 words in the LDA topics. The minimum and maximum are 4 and 6, respectively. The mean is 5.989 and the variance

Table 3.1 Top 10 words of the true topics, LDA estimated topics, and ETM estimated topics of the simulation case 1

Topic 1	Top 10 words									
True	47	83	86	81	153	270	80	14	291	258
LDA	47	**303**	**302**	**301**	83	86	81	270	153	80
ETM	47	86	83	153	81	14	258	30	270	196
Topic 2	Top 10 words									
True	170	256	206	0	219	286	243	114	132	82
LDA	170	206	256	0	**301**	**303**	**302**	219	243	114
ETM	170	206	256	219	243	114	0	286	82	132

Table 3.2 Summary statistics of 1000 repetition

	Min	Max	Mean	Variance
LDA: number of common words	4	6	5.989	0.013
ETM: number of common words	0	6	0.585	1.410
Ratio of common words average probabilities	0.113	0.547	0.322	0.004

is 0.013, indicating that these 3 corpus-specific common words almost always appear in the top 10 words of the LDA estimated topics. Row 2 lists the summary statistics of the number of common words appearing in the top 10 words of the document frequency ETM. Their minimum and maximum are 0 and 6, respectively. The mean is 0.585 and the variance is 1.410. It shows that the majority of the simulation does not get any common words in the top 10 words of ETM estimated topics. The last row shows the summary statistics of the ratio of the average probabilities of these three corpus-specific common words in the ETM and those in the LDA. The minimum and maximum are 0.113 and 0.547. The mean is 0.322 and the variance is 0.004. This shows that the probabilities assigned to these corpus-specific common words in the ETM are on average about 1/3 of those in the LDA.

3.3.2 Case 2: Important Words Appear Rarely in Corpus

We consider another situation. In practice, certain words are important for this domain. But unfortunately, they appear rarely in the current corpus. Practitioners might find these words very important and want them to be assigned high probabilities in the topic word distribution. The LDA word generating process is the same as before. Namely, the corpus contains 2 topics and the prior $\zeta = 0.1$. The topic word distribution η is randomly drawn from a Dirichlet prior $\eta \sim \text{Dirichlet}(0.1 * \mathbb{1})$, where $\mathbb{1}$ is a 300×1 vector of 1s. We assume that the top 2 words in each topic appear rarely in the current corpus for some reason. They only occur in 10% of the total documents, i.e., we randomly select 50 documents that contain the top words and delete them from the remaining documents that contain them.

We then apply the LDA and the ETM ($\nu = 0$) to these words. Unlike the previous setup, where corpus-specific common words are not known and we use document frequencies as weights, these important words are known and we assign negative weights to them and zero weights to all other words. In the simulation, the words 275 and 22 are important for Topic 1, and the words 73 and 195 are important for Topic 2; see also Table 3.3. Their total appearance is limited to 50 documents, i.e., 10% of the corpus size. Due to their rare appearance, the LDA is not able to capture their importance and they are assigned small probabilities. They do not appear in the top 10 words of the LDA topics. By assigning these four words weight -100 and penalty

Table 3.3 Top 10 words of the true topics, LDA estimated topics, and ETM estimated topics of the simulation case 2

Topic 1	Top 10 words								
True	**275**	**22**	110	251	291	151	171	18	253	187
LDA	110	251	151	291	171	253	18	187	35	287
ETM	**275**	**22**	110	251	151	291	171	253	18	187
Topic 2	Top 10 words									
True	**73**	**195**	294	207	48	248	19	211	43	175
LDA	294	207	48	248	19	43	211	175	269	213
ETM	**73**	**195**	294	207	48	248	19	43	211	175

Table 3.4 Summary statistics of 1000 repetition

	Min	Max	Mean	Variance
LDA: number of rare important words	0	3	0.299	0.288
LASSO LDA: number of rare important words	2	4	3.993	0.009
Ratio of rare important words average probabilities	4.187	24.853	9.767	6.427

weight $\mu = 0.2$, the ETM successfully restores their position in the top 10 words of the estimated topics.

We repeat the above procedure 1000 times. The summary statistics are reported in Table 3.4. Row 1 reports the number of important words appearing in the top 10 topics estimated by LDA. The minimum and maximum are 0 and 3, respectively. The mean is 0.299 and the variance is 0.288, indicating that these important but rarely appearing words barely appear in the top 10 LDA topics, i.e., LDA is not able to recover their importance. Row 2 reports the number of important but rarely appearing words that occur in the top 10 ETM topics. The minimum and maximum are 2 and 4, respectively. The mean is 3.993 and the variance is 0.009, which shows that these words are almost always recovered by the ETM. The last row reports the statistics of the average ratio between the probabilities assigned to these important but rarely appearing words in the ETM and those in the LDA. The minimum and maximum are 4.187 and 24.857, respectively. The average is 9.767 and the variance is 6.427, which shows that the probabilities assigned to these important but rarely appearing words are on average about 10 times higher in the ETM than in the LDA.

3.3.3 Case 3: "Close" Topics

Practitioners often find that some estimated topics are "close". By "close" we mean that the estimated topics share several common words and have similar semantic meaning. The exact reason for this phenomenon is unclear. We hypothesize that it is due to the exchangeability assumption of words in the LDA. Nevertheless, we simulate the case with frequently occurring words. The basic setup is similar to case 1. Namely, the corpus contains 2 topics. The prior $\zeta = 0.1$. The topic word distribution η is randomly drawn from a Dirichlet prior $\eta \sim \text{Dirichlet}(0.1 * \mathbb{1})$, where $\mathbb{1}$ is a 300×1 vector of 1s. To simulate the estimated "close" topics in the LDA, we further add 6 common words 301, 302, 303, 304, 305, and 306 to the corpus and assume that they occur in 80% of the documents. With this setup, we apply the LDA and ETM ($\mu = 0$) with penalty weight for pairwise KL divergence $\nu = 0.5$ to the simulated corpus. The true and estimated topics are listed in Table 3.5. Because of the dominance of frequently appearing words, the LDA topics are "close" to each other according to our design. The ETM clearly separates them from each other. Although the appearance order of the ETM is slightly different from that of the true model, the number of the same words appearing in both true and ETM is 8 for topics 1 and 2, while the number of the same words for true topics and LDA is 4 and 5 for topics 1 and 2, respectively. The Jensen–Shannon divergence (JSD) of the true, LDA, and ETM topics is 0.81, 0.63, and 0.88, respectively.

We repeat the simulation 1000 times and report the summary statistics in Table 3.6. The first column shows the number of top 10 words shared between topics 0 and 1. The true topics share on average 0.34 words with a standard deviation of 0.57. The LDA estimated topics share an average of 3.10 words with a standard deviation of 1.61. The ETM shares an average of 0.08 words with a standard deviation of 0.30. This shows that the ETM is able to separate the "close" topics and ensure that they have few words in common in their top words. While the first column focuses on the top words, the second column deals with the overall topic distribution. It reports the JSD of the topic distributions. The JSD of the true topic is on average 0.80 with a standard deviation of 0.05, while the LDA is on average 0.64 with a standard

Table 3.5 Top 10 words of the true topics, LDA estimated topics, and ETM estimated topics of the simulation case 3

Topic 1	Top 10 words									
True	196	56	166	122	219	161	18	104	276	86
LDA	196	56	166	122	301	**303**	**306**	**302**	**305**	**304**
ETM	134	56	196	122	161	166	86	219	104	301
Topic 2	Top 10 words									
True	46	165	115	140	53	280	138	19	174	290
LDA	46	165	115	**304**	140	**305**	**302**	**306**	280	**303**
ETM	85	46	165	140	115	280	53	19	304	138

Table 3.6 Summary statistics of 1000 repetition

	# of shared top 10 words between topics 0 and 1	JSD	# of shared top 10 words between true topic 0 and estimated topic 0	# of shared top 10 words between true topic 1 and estimated topic 1
True	0.34	0.80		
	(0.57)	(0.05)		
LDA	3.10	0.64	5.54	5.50
	(1.61)	(0.04)	(1.24)	(1.19)
KL Div	0.08	0.79	8.23	8.20
	(0.30)	(0.06)	(1.25)	(1.28)

deviation of 0.04 and the ETM is on average 0.79 with a standard deviation of 0.06. It can be seen that the ETM separates the "close" topics. The last two columns report the number of top 10 words in common between the true topics and the estimated topics. It makes no sense if the ETM separates topics but the estimate is far from the true topics. Because of the setup, the LDA topics and true topics share 5.54 and 5.50 words on average with standard deviations of 1.24 and 1.19 for topics 0 and 1, respectively. The ETM and true topics share 8.23 and 8.20 words on average with standard deviations of 1.25 and 1.28, respectively. This means that judging from the top T words, the ETM topics are semantically close to the true model.

3.4 Real Data Application

To test the empirical performance of our proposed method, we apply the LDA, the ETM with LASSO penalty only ($\nu = 0$), and the ETM with the pairwise KL divergence only ($\mu = 0$) to the NIPS dataset, which consists of 11,463 words and 7,241 NIPS-conference posts from 1987 to 2017. The data is randomly split into two parts: Training (80%) and Testing (20%). We choose the number of topics for the LDA using cross-validation with perplexity [8–10] on the training dataset [10]. It is assumed that the chosen number is the true number of topics in the dataset NIPS. The candidates are {5, 10, 15, 20, 25, 30}. The choice $K = 10$ produces the lowest average validation perplexity. For the ETM, we use homogeneous hyperparameters in this experiment, i.e., $\mu_i = \mu, \forall i \in \{1, \dots, K\}$ for the weighted LASSO penalty and $\nu_{il} = \nu, \forall i, l \in \{1, \dots, K\}$ and $l \neq i$ for the pairwise KL divergence penalty, since we have no prior information about the subjects.

We perform cross-validation with the training data to select the penalty weight μ. In selecting μ, perplexity is no longer an appropriate measure. Perplexity is the negative probability per word. A higher probability for frequently occurring words leads to a lower perplexity. As a result, $\mu = 0$ is selected. Another commonly used metric for hyperparameter selection is the Topic Coherence Score. Researchers have proposed several calculation methods of Topic Coherence Score [1, 29, 31, 35]. It was shown in [35] that among all these proposed topic coherence scores, C_V achieves the highest correlation with all available human topic ranking data (see also [43]). Roughly speaking, C_V takes into account both generalization and localization of topics. Generalization means that C_V measures performance on the unseen test dataset. Localization means that C_V uses a rolling window to measure the co-occurrence of the word.

Here we describe the details of the C_V calculation. The top N words of each topic are selected as the representation of the topic, denoted as $W = \{w_1, \ldots, w_N\}$. Each word w_i is represented by an N-dimensional vector $v(w_i) = \{NPMI(w_i, w_j)\}_{j=1,\ldots,N}$, where the j-th entry is the Normalized Pointwise Mutual Information (NPMI) between word w_i and w_j, i.e., $NPMI(w_i, w_j) = \frac{\log P(w_i,w_j) - \log(P(w_i)P(w_j))}{-\log P(w_i,w_j)}$. Here, W represents the sum of all word vectors, $v(W) = \sum_{j=1}^{N} v(w_j)$. The calculation of the NPMI between word w_i and w_j involves the marginal and joint probabilities $p(w_i)$, $p(w_j)$, $p(w_i, w_j)$. A sliding window of size 110, which is the default in the Python package "gensim" and robust for many applications, is used to create pseudo-documents and estimate the probabilities. The purpose of the sliding window is to account for the distance between two words. For each word w_i, a pair $(v(w_i), v(W))$ is formed. A cosine similarity measure $\phi_i(v(w_i), v(W)) = \frac{v(w_i)^T v(W)}{\|v(w_i)\|\|v(W)\|}$ is then calculated for each pair. The final C_V value for the topic is the average of all ϕ_is.

We use C_V to select the penalty weight from {0, 0.5, 1, 1.5, 2, 2.5, 3}. Choosing $\mu = 0.5$ gives the highest average coherence score 0.56 for the ETM, while $\mu = 0$, i.e., the LDA, gives the coherence score 0.51. We then refit both the LDA and the ETM ($\mu = 0.5$, $\nu = 0$) to the entire training dataset and use the test dataset to compute the coherence score C_V as the final evaluation of performance on unseen data. The results are presented in Table 3.7. Overall, the C_V score of the LDA topics is 0.51 and that of the ETM is 0.62, a 22% improvement. We highlight two common words "data" and "using" in the top 20 words of both topics. They appear more frequently in the LDA topics than in the ETM topics. Both words occur in 5 out of 10 LDA topics and 1 out of 10 ETM topics. We also observe some large improvements for the topics "*Reinforcement Learning*", "*Neural Network* ", "*Computer Vision*", and some other topics. Take "*Reinforcement Learning* " as an example. If we compare the top 20 words, we find that the words *time, value, function, model, based, problem* appear in the LDA topic, but not in the ETM topic. Certainly these words are associated with reinforcement learning, but they are also associated with the "*Neural Network*", "*Bayesian*", "*Optimization*", etc. topics. We refer to these words as corpus-specific common words, i.e., for the current corpus, they contain little information to distinguish one topic from another. Their positions in the ETM topic are filled by the words *game, trajectory, robot, control*. These words are related to reinforcement learning applications and represent the topics better than the previous corpus-specific common words. We see that the C_V score has increased from 0.56 to 0.77, an increase of 38%. We observe that an LDA topic related to the NLP is missing in the ETM. One possible reason is that the algorithm converges to different points for this topic on

Table 3.7 Top 20 words of the topics estimated by the LDA and the ETM

Topics	Top 20 words	C_V
LDA		0.51
Machine Learning	Matrix **data** kernel problem algorithm sparse linear method rank methods **using** dimensional analysis vector space function norm error matrices set	0.42
Reinforcement Learning	*State learning policy action time value reward function model optimal actions states agent control reinforcement algorithm* **using** *based decision problem*	*0.56*
Neural Network	*Model time neurons figure spike neuron neural response stimulus activity visual input information cells signal fig cell noise brain synaptic*	*0.65*
Computer Vision	*Image images learning model training deep* **using** *layer neural object network networks features recognition use models dataset feature results different*	*0.57*
NLP	Model word models words features **data** set figure **using** human topic speech object language objects used recognition context based feature	0.50
Neural Network	Network networks neural input learning output training units hidden error weights time function weight layer figure number set used memory	0.52
Bayesian	Model distribution **data** models log gaussian likelihood bayesian inference parameters posterior prior **using** process distributions latent variables mean time probability	0.49
Graph models	Graph algorithm tree set clustering node nodes number cluster structure problem **data** time variables graphs edge clusters random algorithms edges	0.52
Optimization	Algorithm bound theorem log function learning let algorithms bounds problem convex loss optimization case set convergence functions optimal gradient probability	0.43
Classification	Learning **data** training classification set class test error examples function classifier **using** label feature features loss problem kernel performance svm	0.46
ETM		0.62
Machine Learning	Matrix rank sparse pca tensor LASSO subspace spectral manifold norm matrices recovery sparsity eigenvalues kernel principal eigenvectors singular entries embedding	0.55
Reinforcement Learning	*Policy action reward agent state actions reinforcement policies game agents states trajectory robot planning control trajectories rewards games exploration transition*	*0.77*
Neural Network	*Neurons network neuron spike input neural synaptic time firing activity dynamics output networks fig circuit spikes cell signal analog patterns*	*0.71*

(continued)

Table 3.7 (continued)

Topics	Top 20 words	C_V
Computer Vision	*Image images object objects segmentation scene pixel face detection video pixels vision patches visual shape recognition motion color pose patch*	*0.78*
Neural Network	Model visual stimulus brain response spatial human stimuli responses subjects motion frequency cells temporal cortex signals signal activity filter motor	0.73
Neural Network	Layer network deep networks units hidden word layers training convolutional trained neural speech recognition architecture language recurrent net input output	0.73
Bayesian	Inference latent posterior tree variational bayesian node models topic nodes variables model likelihood Markov distribution graphical Gibbs prior Dirichlet sampling	0.54
Graph models	Convex graph algorithm optimization clustering gradient convergence problem theorem solution algorithms dual descent submodular stochastic iteration graphs objective max problems	0.50
Optimization	Bound theorem regret loss bounds algorithm risk lemma log let proof online ranking bounded bandit query setting hypothesis complexity learner	0.52
Classification	Learning **data** model set **using** function algorithm number time figure given results training used based problem error models use distribution	0.37

the path of the variational algorithm. One way to avoid this is to initialize the ETM with a rough estimate from the LDA.

For the ETM with only pairwise KL divergence penalty ($\mu = 0$), we initialize the subject with an estimate of the LDA due to non-convexity in the M-step optimization. Coincidentally, $\nu = 0.5$ also yields the largest coherence score 0.53, and $\nu = 0$ obtains 0.50. As before, both models are again applied to the entire training dataset and their C_V score is calculated using the test dataset as the final evaluation. The results are presented in Table 3.8. Overall, the topic estimated by the LDA has a coherence score of 0.52, while that of the ETM is 0.57. To measure the "distance" between the estimated topics, we calculate the JSD of both topics. The JSD of the LDA topics is 0.93, while that of the ETM topics is 1.90. In terms of distance, the ETM topics are more separated from each other. For the NIPS dataset, we don't observe similar topics. Although the third and eighth topics are both interpreted as neural networks, they emphasize different aspects of the topic. The third topic is from a biological point of view. It contains the words *spike, neuron, stimulus, brain, synaptic*. The eighth topic is from a computer science point of view. It contains the words *network, learning, training, output, layer, hidden*. When separating the estimated topics, the ETM suppresses the appearance of less topic relevant words

Table 3.8 Top 20 words of the topics estimated by the LDA and the ETM

Topics	Top 20 words	C_V
LDA		0.52
Machine Learning	*Matrix data kernel sparse linear points problem rank algorithm space using dimensional method analysis matrices vector clustering error set methods*	*0.42*
Reinforcement Learning	*State learning policy action time reward value* ***function*** *algorithm optimal agent actions states reinforcement problem control* ***model*** *decision using based*	*0.56*
Neural Network	**Model** time neurons figure spike neuron neural information response activity stimulus visual cells cell input fig signal brain different synaptic	0.65
Computer Vision	*Image images* ***model*** *object training deep using learning features recognition models layer feature figure objects use visual different results vision*	*0.55*
Theory	Theorem bound algorithm let log learning **function** probability bounds loss case distribution error set proof lemma functions following sample given	0.42
Bayesian	**Model** data distribution models gaussian log likelihood parameters using posterior bayesian prior inference process latent **function** mean time distributions sampling	0.47
Graphical Models	Graph tree **model** node nodes set algorithm structure number variables models inference graphs clustering cluster edge edges figure time topic	0.48
Neural Network	Network neural networks input learning output training units layer hidden time weights error figure **function** weight used set using state	0.52
Optimization	Algorithm optimization gradient problem **function** convex algorithms method methods convergence solution learning set time objective problems linear stochastic step descent	0.54
Classification	Learning data training classification set features feature using class **model** test task classifier label used based performance examples number labels	0.55
ETM		0.57
Machine Learning	*Clustering kernel matrix norm rank kernels spectral pca matrices tensor subspace LASSO manifold eigenvalues embedding principal singular completion recovery eigenvalue*	*0.54*
Reinforcement Learning	*Action policy reward agent actions state game reinforcement regret arm planning policies exploration robot games agents player states bandit rewards*	*0.76*
Neural Network	Neural input time figure **model** neurons visual neuron fig spike response information spatial signal activity pattern cell different temporal cells	0.64

(continued)

Table 3.8 (continued)

Topics	Top 20 words	C_V
Computer Vision	*Faces image images layer object deep segmentation layers convolutional objects pixel scene pixels video architecture recognition vision networks face pose*	*0.73*
Theory	Bound algorithm theorem let **function** learning log set case probability bounds functions loss error following proof problem given optimal random	0.40
Neural Network	Gates network units networks sonn recurrent net hidden architecture layer analog feedforward backpropagation chip nets connectionist gate modules module feed	0.60[a]
Bayesian	Data **Model** gaussian distribution prior models log mean parameters likelihood noise estimation estimate density using variance bayesian mixture samples process	0.47
Graphical Models	Graph **model** models tree nodes node inference structure variables number Markov set graphs edge time topic edges probability cluster graphical	0.47
Optimization	Algorithm optimization problem algorithms gradient method methods solution **function** convergence objective problems step linear iteration stochastic max update descent learning	0.54
Classification	Learning data classification training set feature test features task class classifier label using examples performance used labels tasks based word	0.57

The 0.60[a] is computed by removing the word *sonn*. As *sonn* doesn't appear in the testing dataset, we get NaN for the C_V score

and improves the topic coherence. For example, the LDA topic "*Machine Learning*" contains the words *points, problem, using, set, methods*. The ETM suppresses the appearance of these words. Instead, it promotes words *spectral, pca, LASSO, manifold, eigenvalues, embedding, principal, singular*, which are better representatives of the topic. As a result, the C_V score improves from 0.42 to 0.54, an improvement of 29%. Similarly, the LDA topic "*Reinforcement Learning*" contains the words *learning, time, value, function, problem, model, using, based*, which are quite common and may have high probabilities in other topics, such as "*Neural Network*", "*Computer Vision*", "*Theory*", and "*Optimization*". The ETM replaces these words with more specific and related words *game, regret, planning, exploration, robot*. The C_V score increases from 0.56 to 0.76, an improvement of 36%. The same is true for the topic *Computer Vision*. Words *model, training, using, learning, use, different, results* are suppressed in the ETM. The words *faces, segmentation, convolutional, pixel, video* that are unique to the topic are promoted in the ETM topic. The C_V score increases from 0.55 to 0.73, an improvement of 33%. Although some corpus-specific common words are suppressed under both penalties, the underlying reasons are different. Under the weighted LASSO penalty, common words are penalized because they appear in too many documents. Under the pairwise KL divergence penalty, some common words are suppressed because they appear in other topics. To distance topics from each other, the pairwise KL divergence penalty suppresses their appearance in less relevant topics.

3.5 Conclusion

Motivated by the frequent word predominance in the discovered topics, we have proposed an ETM containing a weighted LASSO penalty term and a pairwise KL divergence term. The penalties destroy the closed form solution for the topic distribution as in the LDA. Instead, we have estimated the topics using the constrained Newton's method for the case where only the weighted LASSO penalty is present, and using a combination of gradient descent and Hessian descent for the case where only the pairwise KL divergence is present. The combination of gradient descent and Hessian descent algorithm is ready to be applied to the ETM with both penalties, with a small twist to the gradient of the objective function. Although the intent is to solve the problem of frequent word intrusion for the weighted LASSO penalty, it is not limited to this sole purpose. Practitioners can use the weights to incorporate their prior knowledge of the topics. We have demonstrated the effectiveness of the proposed model on three simulation studies, where the ETM outperforms the LDA in every case and recovers the true topics. We also have applied the proposed method to the publicly available NIPS dataset. Compared to the LDA, our proposed method assigns lower weights to frequently occurring words, making the topics easier to interpret. The topic coherence score C_V also shows that the topics are semantically more consistent than those estimated by the LDA.

The ETM with only one weighted LASSO penalty is related to [7]. They claimed that the Dirichlet-Laplace priors have optimal posterior concentration and lead to efficient posterior computation. The LASSO penalties can be considered as the Laplace prior in the posterior. The weights control the mixture between Dirichlet prior for topic drawing and Laplace prior. In our current setup, the topic distributions have been treated as estimated parameters. With the Dirichlet-Laplace prior, we can adopt the full Bayesian approach that the topics are generated from the Dirichlet-Laplace prior. We would achieve a very similar posterior with our current setup, except for an additional variational distribution for the topics. Instead of estimating the topic parameters, we would estimate the variational parameters for the topics. The advantage of the full Bayesian approach is that we can use the property of optimal posterior concentration [7] and theoretical properties of variational inference [33, 46, 48, 49] to show some properties of the proposed method. On the other hand, it is more challenging to derive the theoretical properties related to the pairwise KL divergence penalty because there is no ready prior distribution corresponding to it. The non-convexity means that we are only able to obtain a local minimum for the topic distributions.

References

1. ALETRAS, N. AND STEVENSON, M. (2013). Evaluating topic coherence using distributional semantics. *In: Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers* 13–22.
2. ALLEN- ZHU, Z. AND LI, Y. (2018). Neon2: Finding local minima via first-order oracles. *In: Advances in Neural Information Processing Systems* 3716–3726.
3. ALSUMAIT, L., BARBARÁ, D., GENTLE, J. AND DOMENICONI, C. (2009). Topic significance ranking of LDA generative models. *In: Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer 67–82.
4. ARUN, R., SURESH, V., MADHAVAN, C. V. AND MURTHY, M. N. (2010). On finding the natural number of topics with latent dirichlet allocation: Some observations. *In: Pacific-Asia conference on knowledge discovery and data mining*. Springer 391–402.
5. BAO, Y. AND DATTA, A. (2014). Simultaneously discovering and quantifying risk types from textual risk disclosures. *Management Science* **60** 1371–1391.
6. BATMANGHELICH, K., SAEEDI, A., NARASIMHAN, K. AND GERSHMAN, S. (2016). Non-parametric spherical topic modeling with word embeddings. *In: Proceedings of the conference. Association for Computational Linguistics. Meeting*. NIH Public Access, vol 2016p 537.
7. BHATTACHARYA, A., PATI, D., PILLAI, N. S. AND DUNSON, D. B. (2015). Dirichlet–Laplace priors for optimal shrinkage. *Journal of the American Statistical Association* **110** 1479–1490.
8. BLEI, D. AND LAFFERTY, J. (2006a). Correlated topic models. *Advances in Neural Information Processing Systems* **18** 147.
9. BLEI, D. AND LAFFERTY, J. (2006b). Dynamic topic models. *In: Proceedings of the 23rd International Conference on Machine Learning*. ACM 113–120.
10. BLEI, D., NG, A. AND JORDAN, M. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research* **3** 993–1022.
11. BLEI, D. M. (2012). Probabilistic topic models. *Communications of the ACM* **55** 77–84.
12. BLEI, D. M., KUCUKELBIR, A. AND MCAULIFFE, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association* **112** 859–877.
13. BOYD, S., BOYD, S. P. AND VANDENBERGHE, L. (2004). *Convex Optimization*. Cambridge University Press.
14. BOYD- GRABER, J. L. AND BLEI, D. M. (2009). Syntactic topic models. *In: Advances in Neural Information Processing Systems*.
15. CHEN, X., HU, X., SHEN, X. AND ROSEN, G. (2010). Probabilistic topic modeling for genomic data interpretation. *In: 2010 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*. IEEE 149–152.
16. DAS, R., ZAHEER, M. AND DYER, C. (2015). Gaussian lda for topic models with word embeddings. *In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* 795–804.
17. FEI- FEI, L. AND PERONA, P. (2005). A Bayesian hierarchical model for learning natural scene categories. *In: Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*. IEEE, vol 2 524–531.
18. GRIFFITHS, T. L., JORDAN, M. I., TENENBAUM, J. B. AND BLEI, D. M. (2004). Hierarchical topic models and the nested chinese restaurant process. *In: Advances in Neural Information Processing Systems* 17–24.
19. GRIFFITHS, T. L., STEYVERS, M., BLEI, D. M. AND TENENBAUM, J. B. (2005). Integrating topics and syntax. *In: Advances in Neural Information Processing Systems* 537–544.
20. GRUBER, A., WEISS, Y. AND ROSEN- ZVI, M. (2007). Hidden topic Markov models. *In: Artificial intelligence and statistics* 163–170.
21. HEAUKULANI, C. AND GHAHRAMANI, Z. (2013). Dynamic probabilistic models for latent feature propagation in social networks. *In: International Conference on Machine Learning* 275–283.

22. Huang, A. H., Lehavy, R., Zang, A. Y. and Zheng, R. (2018). Analyst information discovery and interpretation roles: A topic modeling approach. *Management Science* **64** 2833–2855.
23. Korfiatis, N., Stamolampros, P., KouroutHanassis, P. and Sagiadinos, V. (2019). Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. *Expert Systems with Applications* **116** 472–486.
24. La Rosa, M., Fiannaca, A., Rizzo, R. and Urso, A. (2015). Probabilistic topic modeling for the analysis and classification of genomic sequences. *BMC bioinformatics* **16** 1–9.
25. Lei, H., Chen, Y. and Chen, C. Y.- H. (2020). Investor attention and topic appearance probabilities: Evidence from treasury bond market. *Available at SSRN* **3646257**.
26. Liu, L., Tang, L., Dong, W., Yao, S. and Zhou, W. (2016). An overview of topic modeling and its current applications in bioinformatics. *SpringerPlus* **5** 1608.
27. Ma, B., Zhang, N., Liu, G., Li, L. and Yuan, H. (2016). Semantic search for public opinions on urban affairs: A probabilistic topic modeling-based approach. *Information Processing & Management* **52** 430–445.
28. Miller, K., Jordan, M. and Griffiths, T. (2009). Nonparametric latent feature models for link prediction. *Advances in neural information processing systems* **22** 1276–1284.
29. Mimno, D., Wallach, H. M., Talley, E., Leenders, M. and McCallum, A. (2011). Optimizing semantic coherence in topic models. *In: Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics 262–272.
30. Nesterov, Y. and Polyak, B. T. (2006). Cubic regularization of Newton method and its global performance. *Mathematical Programming* **108** 177–205.
31. Newman, D., Lau, J. H., Grieser, K. and Baldwin, T. (2010). Automatic evaluation of topic coherence. *In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics 100–108.
32. Nguyen, T. H. and Shirai, K. (2015). Topic modeling based sentiment analysis on social media for stock market prediction. *In: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* 1354–1364.
33. Pati, D., Bhattacharya, A. and Yang, Y. (2018). On statistical optimality of variational Bayes. *In: International Conference on Artificial Intelligence and Statistics* 1579–1588.
34. Rabinovich, M. and Blei, D. (2014). The inverse regression topic model. *In: International Conference on Machine Learning* 199–207.
35. Röder, M., Both, A. and Hinneburg, A. (2015). Exploring the space of topic coherence measures. *In: Proceedings of the 8th ACM International Conference on Web Search and Data Mining* 399–408.
36. Rosen- Zvi, M., Griffiths, T., Steyvers, M. and Smyth, P. (2012). The author-topic model for authors and documents. arXiv: 1207.4169.
37. Sarkar, P. and Moore, A. W. (2005). Dynamic social network analysis using latent space models. *ACM SIGKDD Explorations Newsletter* **7** 31–40.
38. Shi, B., Lam, W., Jameel, S., Schockaert, S. and Lai, K. P. (2017). Jointly learning word embeddings and latent topics. *In: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval* 375–384.
39. Soleimani, H. and Miller, D. J. (2014a). Parsimonious topic models with salient word discovery. *IEEE Transactions on Knowledge and Data Engineering* **27** 824–837.
40. Soleimani, H. and Miller, D. J. (2014b). Sparse topic models by parameter sharing. *In: 2014 IEEE International Workshop on Machine Learning for Signal Processing (MLSP)*. IEEE 1–6.
41. Stevens, K., Kegelmeyer, P., Andrzejewski, D. and Buttler, D. (2012). Exploring topic coherence over many models and many topics. *In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning* 952–961.

42. SUN, A., LACHANSKI, M. AND FABOZZI, F. J. (2016). Trade the tweet: Social media text mining and sparse matrix factorization for stock market prediction. *International Review of Financial Analysis* **48** 272–281.
43. SYED, S. AND SPRUIT, M. (2017). Full-text or abstract? examining topic coherence scores using latent dirichlet allocation. *In: 2017 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE 165–174.
44. VU, H. Q., LI, G. AND LAW, R. (2019). Discovering implicit activity preferences in travel itineraries by topic modeling. *Tourism Management* **75** 435–446.
45. WALLACH, H. M., MIMNO, D. M. AND MCCALLUM, A. (2009). Rethinking LDA: Why priors matter. *In: Advances in Neural Information Processing Systems* 1973–1981.
46. WANG, Y. AND BLEI, D. M. (2019). Frequentist consistency of variational Bayes. *Journal of the American Statistical Association* **114** 1147–1161.
47. XU, H., WANG, W., LIU, W. AND CARIN, L. (2018). Distilled Wasserstein learning for word embedding and topic modeling. *In: Advances in Neural Information Processing Systems* 1716–1725.
48. YANG, Y., PATI, D. AND BHATTACHARYA, A. (2020). α-variational inference with statistical guarantees. *Annals of Statistics* **48** 886–905.
49. ZHANG, F., GAO, C. (2020). Convergence rates of variational posterior distributions. *Annals of Statistics* **48** 2180–2207.
50. ZHANG, H., KIM, G. AND XING, E. P. (2015). Dynamic topic modeling for monitoring market competition from online text and image data. *In: Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 1425–1434.

Chapter 4
A Simple Isotropic Correlation Family in $\mathbb{R}^3$ with Long-Range Dependence and Flexible Smoothness

Victor De Oliveira

Abstract Most geostatistical applications use covariance functions that display short-range dependence, in part due to the wide variety and availability of these models in statistical packages, and in part due to spatial interpolation being the main goal of many analyses. But when the goal is spatial extrapolation or prediction based on sparsely located data, covariance functions that display long-range dependence may be more adequate. This paper constructs a new family of isotropic correlation functions whose members display long-range dependence and can also model different degrees of smoothness. This family is compared to a sub-family of the Matérn family commonly used in geostatistics, and two other recently proposed families of covariance functions with long-range dependence are discussed.

Keywords Fractal dimension · Geostatistics · Hurst coefficient · Mean square differentiability · Radial distribution

4.1 Introduction

Random fields are ubiquitous for the modeling of spatial data in most natural and earth sciences. When the main goal of the analysis is spatial prediction, an adequate specification of the correlation function of the random field is of utmost importance. In this paper, attention is restricted to correlation functions in $\mathbb{R}^d$ with the properties of being *isotropic*, i.e., functions of the Euclidean distance which separates two locations that decrease monotonically to zero as distance increases without bound. These features are common in many spatial phenomena. A large number of parametric families of correlation functions with these properties have been proposed in the literature and used in applications; see, for instance, [2, 3]. Most of these families display *short-term* dependence, meaning that the correlation function decays to zero fast, usually exponentially fast, so spatial association between far-away observations is negligible. The Matérn family is a commonly used example. On the other hand,

V. De Oliveira (✉)
The University of Texas at San Antonio, San Antonio, TX 78249, USA
e-mail: victor.deoliveira@utsa.edu

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_4

some spatial phenomena display *long-term* dependence, meaning that the correlation function decays to zero slowly, usually hyperbolically fast, so the spatial association between far-away observations is not negligible. An early example of this behavior was provided by [4] using data from agricultural uniformity trials, who empirically found that, for large distances r, the correlation function decays approximately as r^{-1} (a so-called power law). Similar behavior is commonly found in spatial geophysics and hydrology data; see [8] and the references therein. Fewer models have been proposed in the literature for phenomena that display this behavior.

Time series models displaying long-range dependence were discussed in [5] (discrete time) and [12] (continuous time). Spatial data models displaying long-range dependence were discussed in [10, 17] for the case when the index set is $\mathbb{Z}^d$ (usually $d = 2$). Some families of correlation functions for random fields in $\mathbb{R}^d$ that display long-range dependence were constructed by [21], and more recently [8, 13] developed new families in this class. In this paper, we construct what appears to be a new family using a correspondence between continuous isotropic correlation functions in $\mathbb{R}^3$ and probability density functions (pdfs) in $\mathbb{R}$ with support $[0, \infty)$. In addition to displaying long-range dependence, the new family of correlation functions allows different degrees of smoothness, which is important for efficient spatial interpolation under infill asymptotics [20]. After describing its main properties, this family of correlation functions is contrasted with a sub-family of the Matérn family which also provides flexibility regarding smoothness, but displays short-range dependence. This paper ends with a discussion of two other families of correlations functions that also display long-range dependence.

4.1.1 A Spectral Representation

Let $K : [0, \infty) \to \mathbb{R}$ be the correlation function of a mean square continuous and isotropic random field $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$, with $D \subset \mathbb{R}^d$ and $d \geq 1$, and let Φ_d denote the class of all such functions. A characterization of Φ_d is given in a classical result by [18], who showed that any $K \in \Phi_d$ can be written as

$$K(r) = \int_0^\infty \Lambda_d(rx)dF(x), \qquad r \geq 0,$$

where

$$\Lambda_d(t) = \left(\frac{2}{t}\right)^{\frac{d}{2}-1} \Gamma\left(\frac{d}{2}\right) J_{\frac{d}{2}-1}(t), \qquad t > 0,$$

$\Gamma(\cdot)$ is the gamma function, $J_\nu(\cdot)$ is the Bessel function of the first kind and order ν, and $F(\cdot)$ is a cumulative distribution function on $\mathbb{R}$ with support $[0, \infty)$. Such a function $K(\cdot)$ is also a radial positive definite function on $\mathbb{R}^d$, and can also be viewed as the *Hankel* transform of F of order $\frac{d}{2} - 1$. Any $K \in \Phi_d$ is continuous on $[0, \infty)$, and $[(d-1)/2]$-times continuously differentiable on $(0, \infty)$, where $[a]$ denotes the integer part of a, and $\lim_{r\to\infty} K(r) = F(0) - F(0-)$; see [6, 15, 20, 22] for further properties of such functions.

For all $d \geq 1$, $\Lambda_d(t)$ is itself a continuous isotropic correlation function in $\mathbb{R}^d$ (so $\Lambda_d(0) = 1$), which can be written in terms of elementary functions when d is an odd integer. For instance, for $d = 1, 2, 3$ it holds that

$$\Lambda_1(t) = \cos(t), \quad \Lambda_2(t) = J_0(t), \quad \Lambda_3(t) = \frac{\sin(t)}{t},$$

and $\Lambda_\infty(t) := \lim_{d\to\infty} \Lambda_d(t) = e^{-t^2}$. In particular, any isotropic correlation function in $\mathbb{R}^3$ admits the representation

$$K(r) = \int_0^\infty \frac{\sin(rx)}{rx} dF(x), \qquad r \geq 0,$$

and any such function is also an isotropic correlation function in $\mathbb{R}^2$ and $\mathbb{R}^1$, since the classes of functions Φ_d are decreasing in d. If $F(\cdot)$ is absolutely continuous, with pdf $f(\cdot)$ say, then

$$K(r) = \int_0^\infty \frac{\sin(rx)}{rx} f(x)dx, \qquad r \geq 0; \tag{4.1}$$

the functions F and f are also called, respectively, the *radial distribution* and *radial pdf* functions of the random field $Z(\cdot)$ [15]. Therefore, (4.1) establishes a bijection between the class of continuous isotropic correlation functions in $\mathbb{R}^3$ and the class of pdfs in $\mathbb{R}$ (w.r.t. Lebesgue measure) having support $[0, \infty)$. Consequently, choosing a continuous isotropic correlation function $K(\cdot)$ amounts to choosing a pdf $f(\cdot)$ with support in $[0, \infty)$.

4.2 A New Correlation Family

In this section, we use (4.1) with a particular family of radial pdfs to construct what appears to be a new family of correlation functions in $\mathbb{R}^3$ whose members display long-range dependence and various degrees of smoothness. For $\sigma > 0$ and $m \in \mathbb{N}_0$, let $f_{\sigma,m}(x)$ be the pdf of the $t_{2m+1}(0, \sigma^2)$ distribution[1] truncated to $[0, \infty)$, i.e.,

[1] The symbol $t_\nu(\mu, \sigma^2)$ denotes the t distribution with ν degrees of freedom, location parameter μ and scale parameter σ.

$$f_{\sigma,m}(x) = \frac{2\Gamma(m+1)}{\sigma\sqrt{\pi(2m+1)}\Gamma(\frac{2m+1}{2})}\left(1+\frac{1}{2m+1}\left(\frac{x}{\sigma}\right)^2\right)^{-(m+1)}\mathbf{1}_{(0,\infty)}(x)$$
$$= \frac{2\Gamma(m+1)}{\sigma\sqrt{\pi(2m+1)}\Gamma(\frac{2m+1}{2})}\big((2m+1)\sigma^2\big)^{m+1}\big((2m+1)\sigma^2+x^2\big)^{-(m+1)},$$

when $x > 0$. Then from (4.1), the correlation function in $\mathbb{R}^3$ that corresponds to this radial pdf is

$$K_{\sigma,m}(r) = \frac{2\Gamma(m+1)\big((2m+1)\sigma^2\big)^{m+1}}{\sigma\sqrt{\pi(2m+1)}\Gamma(\frac{2m+1}{2})r}\int_0^\infty \frac{\sin(rx)}{x\big((2m+1)\sigma^2+x^2\big)^{m+1}}dx$$
$$= \frac{2\Gamma(m+1)\big((2m+1)\sigma^2\big)^{m+1}}{\sigma\sqrt{\pi(2m+1)}\Gamma(\frac{2m+1}{2})r}$$
$$\times\frac{\pi}{2\big((2m+1)\sigma^2\big)^{m+1}}\left(1-\frac{e^{-\sigma\sqrt{2m+1}r}}{2^m m!}P_m\big(\sigma\sqrt{2m+1}r\big)\right)$$
$$= \frac{\sqrt{\pi}\Gamma(m+1)}{\sigma\sqrt{(2m+1)}\Gamma(\frac{2m+1}{2})r}\left(1-\frac{e^{-\sigma\sqrt{2m+1}r}}{2^m m!}P_m\big(\sigma\sqrt{2m+1}r\big)\right), \quad (4.2)$$

where the second equality follows from [9, 3.737.3], and $P_m(\cdot)$ is the polynomial of degree m obtained by the recursion

$$P_m(x) = (x+2m)P_{m-1}(x) - xP'_{m-1}(x), \qquad m \geq 1, \quad \text{with } P_0(x) = 1.$$

For instance, for $m = 1, 2, 3$ we have

$$P_1(x) = x+2, \quad P_2(x) = x^2+5x+8, \quad P_3(x) = x^3+9x^2+33x+48.$$

Reparametrizing (4.2) with $\theta := (\sigma\sqrt{2m+1})^{-1}$, we obtain the following two-parameter family of continuous isotropic correlation functions in $\mathbb{R}^3$

$$\mathcal{D} = \left\{K_{\theta,m}(r) := c_m\frac{\theta}{r}\left(1-\frac{e^{-\frac{r}{\theta}}}{2^m m!}P_m\left(\frac{r}{\theta}\right)\right) \;:\; \theta > 0, m \in \mathbb{N}_0\right\}, \quad (4.3)$$

with[2]

$$c_m = \frac{\sqrt{\pi}\Gamma(m+1)}{\Gamma(\frac{2m+1}{2})}.$$

For instance, for $r \geq 0$ and $m = 0, 1, 2$ we have

[2] The fact $K_{\theta,m}(0) = 1$ follows by continuity.

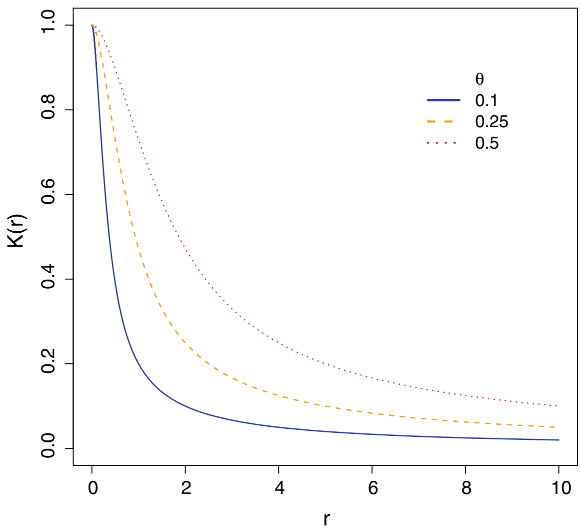


Fig. 4.1 Plots of $K_{\theta,m}(r)$ for $m = 1$ and three values of θ

$$K_{\theta,0}(r) = \frac{\theta}{r}\left(1 - e^{-\frac{r}{\theta}}\right),$$
$$K_{\theta,1}(r) = \frac{\theta}{r}\left(2 - e^{-\frac{r}{\theta}}\left(\frac{r}{\theta} + 2\right)\right),$$
$$K_{\theta,2}(r) = \frac{\theta}{3r}\left(8 - e^{-\frac{r}{\theta}}\left(\left(\frac{r}{\theta}\right)^2 + 5\frac{r}{\theta} + 8\right)\right).$$

4.3 Properties

In this section, we describe some of the properties of the new family of correlation functions. First, any of the correlation functions in (4.3) displays long-range dependence, as it decays slowly with increasing distance r. Specifically, for any $\theta > 0$ and $m \in \mathbb{N}_0$ it holds that $K_{\theta,m}(r) \to 0$ and $K_{\theta,m}(r) = O(1/r)$ as $r \to \infty$, so

$$\int_0^\infty r^{d-1} K_{\theta,m}(r)dr \quad \text{diverges} \qquad (d = 1, 2, 3). \tag{4.4}$$

For isotropic correlation functions, the above property defines long-range dependence. Second, the interpretation of the parameters is the following. The parameter θ is a *range* parameter that controls how fast the correlation function decays with distance r. This is illustrated in Fig. 4.1 where plots $K_{\theta,m}(r)$ are displayed for $m = 1$ and three values of θ. On the other hand, m is a *smoothness* parameter that controls the mean square differentiability of the random field $Z(\cdot)$, as stated by the following result.

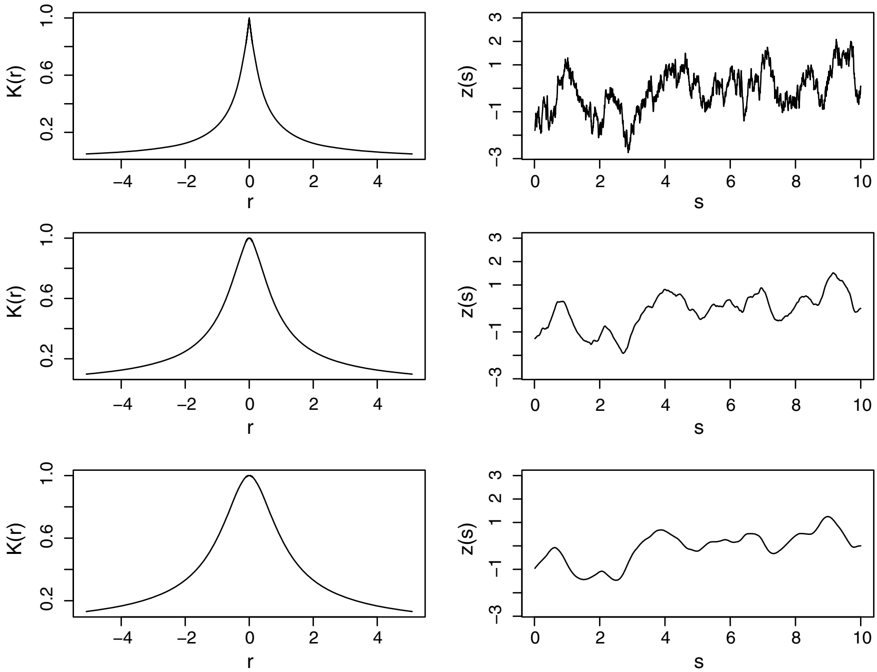


Fig. 4.2 Plots of the even extension of $K_{\theta,m}(r)$ (left) and corresponding realizations of zero-mean Gaussian random fields with these correlation functions (right). In all $\theta = 0.25$ and $m = 0$ (top), 1 (middle) and 2 (bottom)

Proposition 4.1 *Let $\{Z(\mathbf{s}) : \mathbf{s} \in D\}$, with $D \subset \mathbb{R}^d$ and $d \leq 3$, be an isotropic mean square continuous random field with correlation function $K_{\theta,m}(r)$ from the family (4.3). Then $Z(\cdot)$ is m-times mean square differentiable.*

Proof For any $k \in \mathbb{N}_0$, an isotropic random field $Z(\cdot)$ is k-times mean square differentiable if and only if its correlation function is $2k$-times differentiable at zero[3] [20]. In addition, for any radial positive definite function in $\mathbb{R}^d$, $K(r)$ say, $K^{(2k)}(0)$ exists if and only if the radial pdf f in the representation (4.1) has a finite moment of order $2k$ [6, Lemma 3]. Since the radial distribution associated with $K_{\theta,m}(r)$ is the $t_{2m+1}(0, \sigma^2)$ distribution truncated to $[0, \infty)$ and this has finite moments up to order $2m$, the above results imply that a random field with correlation function $K_{\theta,m}(r)$ is exactly m-times mean square differentiable. □

To illustrate the above result, Fig. 4.2 plots the even extension of $K_{\theta,m}(r)$ (left) and corresponding realizations of zero-mean Gaussian random fields in the real line with these correlation functions (right), where $\theta = 0.25$ and $m = 0$ (top), 1 (middle),

[3] Differentiability of $K(\cdot)$ at zero refers to differentiability of its even extension over the real line, defined as $K^e(r) := K(|r|)$, $r \in \mathbb{R}$. Also, the phrase '$Z(\cdot)$ is 0-times mean square differentiable' is used if $Z(\cdot)$ is mean square continuous.

and 2 (bottom). The plots show the smoothness of $K_{\theta,m}(r)$ at the origin increasing with m, and its corresponding effect on the smoothness of the realizations. The three realizations were obtained from the same seed. Together, Figs. 4.1 and 4.2 suggest that $\mathcal{D}$ is a flexible family of correlation functions capable of describing different degrees of spatial association and smoothness in random fields that display long-range dependence.

4.4 Comparison With a Matérn Sub-family

The Matérn family of correlation functions [15, 20] is a two-parameter family of correlation functions in $\mathbb{R}^d$ for all $d \geq 1$ that is commonly used in geostatistical applications. It is given by

$$M_{\theta,\nu}(r) = \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(\frac{r}{\theta}\right)^{\nu}\mathcal{K}_{\nu}\left(\frac{r}{\theta}\right), \tag{4.5}$$

where $\theta,\ \nu > 0$ and $\mathcal{K}_{\nu}(.)$ is the modified Bessel function of second kind and order ν; see [9, 8.40] for details on the behavior of this special function. This family contains the exponential correlation function $e^{-r/\theta}$ (obtained when $\nu = 0.5$), and the squared exponential correlation function $e^{-(r/\vartheta)^2}$ is a limit case (obtained when $\theta = \vartheta/2\sqrt{\nu}$ and $\nu \to \infty$). We consider here the following sub-family:

$$\mathcal{M} = \{M_{\theta,m+0.5}(r) : \theta > 0, m \in \mathbb{N}_0\}. \tag{4.6}$$

Like the family $\mathcal{D}$ in (4.3), θ is a *range* parameter that controls how fast the correlation function decreases with distance r, and m is a *smoothness* parameter that controls the mean square differentiability of the random field $Z(\cdot)$. It was shown by [20] that a random field $Z(\cdot)$ with correlation function $M_{\theta,m+0.5}(r)$ is exactly m-times mean square differentiable. Additionally, $M_{\theta,m+0.5}(r)$ can be written as $e^{-r/\theta}$ times a polynomial in r of degree m [9, 8.468]. For instance, for $r \geq 0$ and $m = 0, 1, 2$ we have

$$\begin{aligned}
M_{\theta,0.5}(r) &= e^{-\frac{r}{\theta}}, \\
M_{\theta,1.5}(r) &= e^{-\frac{r}{\theta}}\left(\frac{r}{\theta} + 1\right), \\
M_{\theta,2.5}(r) &= e^{-\frac{r}{\theta}}\left(\frac{1}{3}\left(\frac{r}{\theta}\right)^2 + \frac{r}{\theta} + 1\right).
\end{aligned}$$

But unlike the family $\mathcal{D}$, the correlation functions in (4.6) display short-range dependence, since for any $\theta > 0$ and $m \in \mathbb{N}_0$

$$M_{\theta,m+0.5}(r) \sim ar^{m-\frac{1}{2}}e^{-\frac{r}{\theta}}, \qquad \text{as } r \to \infty,$$

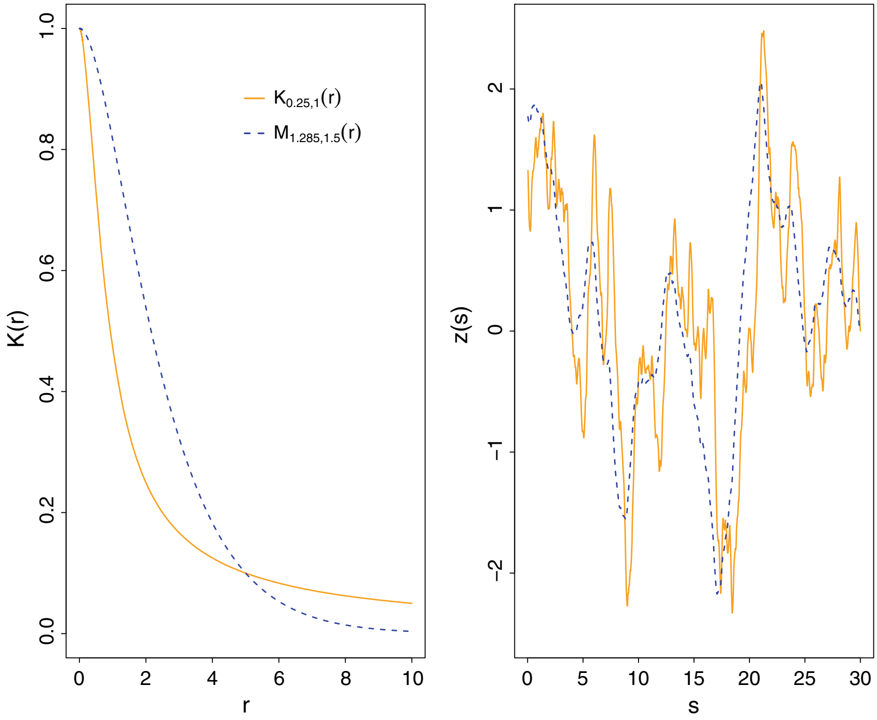


Fig. 4.3 Plots of $K_{0.25,1}(r)$ and $M_{1.285,1.5}(r)$ (left) and corresponding realizations of zero-mean Gaussian random fields with these correlation functions (right)

for some $a > 0$ [1, 9.7.2]. So $M_{\theta,m+0.5}(r)$ decreases to zero exponentially fast as $r \to \infty$, and consequently $\int_0^\infty r^{d-1} M_{\theta,m}(r)dr$ converges. To illustrate the different behaviors of correlation functions in the families $\mathcal{D}$ and $\mathcal{M}$ with the same smoothness and similar rates of decay, Fig. 4.3 plots $K_{0.25,1}(r)$ and $M_{1.285,1.5}(r)$ (left) and corresponding realizations of zero-mean Gaussian random fields in the real line with these correlations functions (right); the two realizations were obtained from the same seed. Both correlation functions correspond to random fields that are 1-time mean square differentiable, and their range parameters are such that their correlations at distance $r = 5$ is 0.1. Note that $M_{1.285,1.5}(r)$ has larger correlations than $K_{0.25,1}(r)$ for small distances, but the opposite holds for large distances. As a result, the realization of the random field with correlation function $K_{0.25,1}(r)$ displays more 'oscillatory' behavior for small distances, but process values for large distances are more 'alike' than process values of the realization of the random field with correlation function $M_{1.285,1.5}(r)$. Therefore, the families of correlations $\mathcal{D}$ and $\mathcal{M}$ appear equally flexible in terms of describing different degrees of spatial association and smoothness, but are complementary in terms of the range of dependence, as one displays long-range dependence while the other short-range dependence.

4.5 Other Correlation Families With Long-Range Dependence

4.5.1 The Generalized Cauchy Family

The generalized Cauchy family [8, 11] is a three-parameter family of isotropic correlation functions in $\mathbb{R}^d$, for all $d \geq 1$, given by

$$C_{\alpha,\beta,\theta}(r) = \left(1 + \left(\frac{r}{\theta}\right)^{\alpha}\right)^{-\beta/\alpha},$$

where $\alpha \in (0, 2]$, $\beta > 0$ and $\theta > 0$. As in the previous families, θ is a range parameter. The main virtue of this family is that it allows for independent choices of *fractal dimension* and *Hurst coefficient*, where the former is a measure of the 'roughness' of realizations of random fields with this correlation function, and the latter is a measure of 'persistence' or long-range dependence [7]. Specifically, realizations of a random field in $\mathbb{R}^d$ with correlation function $C_{\alpha,\beta,\theta}(r)$ have fractal dimension [11]

$$D = d + 1 - \alpha/2 \in [d, d + 1),$$

with $D = d$ ($D > d$) when the random field is (is not) mean square differentiable; the larger D is, the rougher the realizations. So this property is entirely controlled by the parameter α.

Additionally, $C_{\alpha,\beta,\theta}(r) \sim r^{-\beta}$ as $r \to \infty$, so it satisfies (4.4) when $\beta \in (0, d]$, and the random field has long-range dependence with Hurst coefficient [11, 16]

$$H = \frac{d + \beta}{2} \in (d/2, d];$$

the closer H is to $d/2$, the stronger the persistence. So this property is entirely controlled by the parameter $\beta \in (0, d]$. The random field has short-range dependence when $\beta > d$. Hence, D and H can vary independently of each other, and they can take any value in their respective ranges of possible values [8]. This property is in sharp contrast with that of self-affine processes often used to model long-range dependence [14] where the fractal dimension and Hurst coefficient are tied by the relation $D + H = d + 1$.

The generalized Cauchy family allows a wide range of 'roughness' and 'persistence' behaviors controlled by the parameters α and β, respectively. On the other hand, this family does not allow a wide range of smoothness behaviors, since a random field with correlation function $C_{\alpha,\beta,\theta}(r)$ is non-differentiable in mean square when $\alpha \in (0, 2)$ and infinitely differentiable when $\alpha = 2$, with no possible intermediate behaviors [19].

4.5.2 The Confluent Hypergeometric Family

Recently, [13] derived a new family of isotropic correlation functions in $\mathbb{R}^d$, for all $d \geq 1$, that display long-range dependence. The construction involves mixing the Matérn correlation functions (in a parametrization different than (4.5)) over the (new) squared range parameter, with the IG$(\alpha, \beta^2/2)$ distribution as the mixing distribution[4]. Specifically, their correlation function is given by

$$\begin{aligned} H_{\alpha,\beta,\nu}(r) &= \int_0^{\infty} M_{\phi/\sqrt{2\nu},\nu}(r) \frac{\beta^{2\alpha}}{2^{\alpha}\Gamma(\alpha)} \phi^{-2(\alpha+1)} e^{-\frac{\beta^2}{2\phi^2}} d\phi^2 \\ &= \frac{\beta^{2\alpha}\Gamma(\nu+\alpha)}{\Gamma(\nu)\Gamma(\alpha)} \int_0^{\infty} x^{\nu-1}(x+\beta^2)^{-(\nu+\alpha)} e^{-\frac{\nu r^2}{x}} dx \\ &= \frac{\Gamma(\nu+\alpha)}{\Gamma(\nu)} U\Big(\alpha, 1-\nu, \nu\Big(\frac{r}{\beta}\Big)^2\Big), \end{aligned}$$

where $\alpha, \beta, \nu > 0$ and $U(a, b, c)$ is the confluent hypergeometric function of the second kind ([1, 13.2]), so this family was named the Confluent Hypergeometric family; see [13] for details.

It was shown by [13] that the Confluent Hypergeometric and Matérn covariance functions with the same parameter ν have the same asymptotic behavior as $r \to 0$. Hence, like the Matérn family, a random field with correlation function $H_{\alpha,\beta,\nu}(r)$ is $[\nu]$-times mean square differentiable [20], and the fractal dimension of realizations from this random field is [7, 16]

$$D = \begin{cases} d+1-\nu & \text{if } \nu \in (0,1) \\ d & \text{if } \nu \geq 1 \end{cases} \quad (D \in [d, d+1)).$$

So this model allows any degree of smoothness and roughness which is controlled by the parameter ν.

Additionally, it was shown by [13] that

$$H_{\alpha,\beta,\nu}(r) \sim a r^{-2\alpha} L(r^2), \qquad \text{as } r \to \infty,$$

for some $a > 0$, where $L(x) := \big(x/(x+\beta^2/(2\nu))\big)^{\nu+\alpha}$ is a slowly varying function at infinity. Then, for any $d \in \mathbb{N}$, $H_{\alpha,\beta,\nu}(r)$ satisfies (4.4) when $\alpha \in (0, d/2]$. Hence, unlike the Matérn family, the Confluent Hypergeometric correlation functions display long-range dependence when $\alpha \in (0, d/2]$. In this case, they decay hyperbolically fast with increasing distance, with the rate of decay controlled by the parameter α, and the Hurst coefficient [16] is

$$H = d/2 + \alpha \ \in \ (d/2, d].$$

[4] The symbol IG$(\alpha, \beta^2/2)$ denotes the inverse gamma distribution with shape parameter α and scale parameter $\beta^2/2$.

The random field has short-term dependence when $\alpha > d/2$. Like the generalized Cauchy family, D and H in the Confluent Hypergeometric family can vary independently of each other and they can take any value in their respective ranges of possible values. But in contrast to the former, the latter family allows a wide range of smoothness behaviors.

4.6 Discussion

Correlation functions displaying short-range dependence are the most often used in geostatistical applications. This practice is mainly due to the following:

(I) Correlation families displaying long-range dependence are fewer and less known in geostatistics than short-range correlation families.
(II) The detection of long-range dependence requires abundant data collected over large regions, which is often not available.
(III) The main goal in many geostatistical applications is spatial interpolation based on densely collected data, in which case the behavior of the correlation function at short distances is much more important than the behavior at large distances.

Nevertheless, recent developments in theory and applications have shown that correlation functions displaying long-range dependence have a role to play in geostatistics.

When the goal is spatial extrapolation or interpolation with sparsely located data, short-range correlation models may provide less satisfactory predictive inferences. In this case, the effect of the correlation function on the optimal linear predictor is negligible, as this predictor is essentially the estimated mean function. This is an unwanted outcome because it is rarely the case in applications that the modeler has strong confidence in the proposed mean function, even when this is constant. In an analysis of carbon dioxide measured in the United States by satellite, [13] found that, for spatial extrapolation, predictive inference based on the Confluent Hypergeometric family was better than that based on the Matérn family (when ν was fixed at the same value in both families). On the other hand, for spatial interpolation, predictive inference based on both families was about the same. These behaviors are explained by the fact that both families are equally flexible in modeling smoothness of the random field, while only the confluent hypergeometric family can model both short- and long-range dependence.

Acknowledgements This work was partially supported by the U.S. National Science Foundation grant DMS–2113375.

References

1. ABRAMOWITZ, M. AND STEGUN, I. (1970). *Handbook of Mathematical Functions*. Dover, New York.
2. CHILÈS, J.-P. AND DELFINER, P. (2012). *Geostatistics: Modeling Spatial Uncertainty*, 2nd ed. Wiley, Hoboken.
3. CRESSIE, N.A.C. (1993). *Statistics for Spatial Data*, rev. ed. Wiley, New York.
4. FAIRFIELD SMITH, H. (1938). An empirical law describing heterogeneity in the yields of agricultural crops. *The Journal of Agricultural Science* **28** 1-23.
5. GNEITING, T. (2000). Power-law correlations, related models for long-range dependence and their simulation. *Journal of Applied Probability* **37** 1104-1110.
6. GNEITING, T. (1999). On the derivatives of radial positive definite functions. *Journal of Mathematical Analysis and Applications* **236** 86-93.
7. GNEITING, T., ŠEVCÍKOVÁ, H. AND PERCIVAL, D.B. (2012). Estimators of fractal dimension: Assessing the roughness of time series and spatial data. *Statistical Science* **27** 247-277.
8. GNEITING, T. AND SCHLATHER, M. (2004). Stochastic models that separate fractal dimension and the Hurst effect. *SIAM Review* **46** 269-282.
9. GRADSHTEYN, I.S. AND RYZHIK, I.M. (2000). *Table of Integrals, Series and Products*, 6th ed., Academic Press, San Diego.
10. LAVANCIER, F. (2006). Long memory random fields. In: *Dependence in Probability and Statistics*, Bertail, P., P. Doukhan and P. Soulier (eds.), Lecture Notes in Statistics **187** pp 195-220, Springer-Verlag, New York.
11. LIM, S.C. AND TEO, L.P. (2009). Gaussian fields and Gaussian sheets with generalized Cauchy covariance structure. *Stochastic Processes and Their Applications* **119** 1325-1356.
12. MA, C. (2003). Long-memory continuous-time correlation models. *Journal of Applied Probability* **40** 1133-1146.
13. MA, P. AND BHADRA, A. (2022). Beyond Matérn: On a class of interpretable confluent Hypergeometric covariance functions. *Journal of the American Statistical Association*, to appear.
14. MANDELBROT, B.B. (1982). *The Fractal Geometry of Nature*. W.H. Freeman, San Francisco.
15. MATÉRN, B. (1986). *Spatial Variation*, 2nd ed., Lecture Notes in Statistics **36**. Springer-Verlag, Berlin.
16. PORCU, E. AND STEIN, M.L. (2012). On some local, global and regularity behaviour of some classes of covariance functions. In: *Advances and Challenges in Space-time Modelling of Natural Events*, Porcu, E., J.M. Montero and M. Schlather (eds.), Lecture Notes in Statistics **207** pp 221-238, Springer-Verlag, Berlin.
17. ROBINSON, P.M. (2020). Spatial long memory. *Japanese Journal of Statistics and Data Science* **3** 243-256.
18. SCHOENBERG, I.J. (1938). Metric spaces and completely monotone functions. *Annals of Mathematics* **39** 811-841.
19. STEIN, M.L. (2005). Nonstationary spatial covariance functions. Technical report, The University of Chicago.
20. STEIN, M.L. (1999). *Interpolation for Spatial Data*. Springer-Verlag, New York.
21. WHITTLE, P. (1962). Topographic correlation, power-law covariance functions, and diffusion. *Biometrika* **49** 305-314.
22. YAGLOM, A.M. (1987). *Correlation Theory of Stationary and Related Random Functions I. Basic Results*. Springer-Verlag, New York.

Chapter 5
Portmanteau Tests for Semiparametric Nonlinear Conditionally Heteroscedastic Time Series Models

Christian Francq, Thomas Verdebout, and Jean-Michel Zakoian

Abstract A class of multivariate time series models is considered, with general parametric specifications for the conditional mean and variance. In this general framework, the usual Box–Pierce portmanteau test statistic, based on the sum of the squares of the first residual autocorrelations, cannot be accurately approximated by a parameter-free distribution. A first solution is to estimate from the data the complicated asymptotic distribution of the Box–Pierce statistic. The solution proposed by Li [23] consists of changing the test statistic by using a quadratic form of the residual autocorrelations which follows asymptotically a chi-square distribution. Katayama [21] proposed a distribution-free statistic based on a projection of the autocorrelation vector. The first aim of this paper is to show that the three methods, initially introduced for specific time series models, can be applied in our general framework. The second aim is to compare the three approaches. The comparison is made on (i) the mathematical assumptions required by the different methods and (ii) the computations of the Bahadur slopes (in some cases via Monte Carlo simulations).

5.1 Introduction

A crucial step in the Box–Jenkins methodology for ARMA model fitting is the model diagnostic checking, see Brockwell and Davis [10]. This step aims at answering the following question: is a given ARMA(p, q) model adequate for the time series at hand? If the ARMA(p, q) model suitably describes the linear dependence structure

C. Francq (✉)
CREST-ENSAE and University of Lille, BP 60149, 59653 Villeneuve d'Ascq cedex, France
e-mail: christian.francq@univ-lille3.fr

T. Verdebout
ECARES and Département de Mathématique, Université libre de Bruxelles (ULB), Boulevard du Triomphe, CP 210, 1050 Brussels, Belgium
e-mail: thomas.verdebout@ulb.be

J.-M. Zakoian
University of Lille and CREST-ENSAE, 5 Avenue Henri Le Chatelier, 91120 Palaiseau, France
e-mail: zakoian@ensae.fr

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_5

of the time series, in particular, if the orders p and q are appropriately chosen, the residuals should be close to the linear innovations, and thus approximately uncorrelated. The so-called portmanteau tests of Box and Pierce [9] and Ljung and Box [27]—arguably the most widely used adequacy tests in time series—reject the candidate ARMA(p, q) model if a weighted sum of the squares of the first m residual autocorrelations exceeds some critical value. Under the assumption that the ARMA innovations are independent and identically distributed (iid), the critical value is approximated by a quantile of a chi-square distribution. The iid assumption for the linear innovations is, however, extremely restrictive, since it implies that the optimal predictions are linear, and it precludes any form of conditional heteroscedasticity, a typical stylized fact of the financial series, see, e.g., Tsay [36] and the references therein.

Several authors have relaxed the iid assumption by studying ARMA models under the more realistic assumption that the innovations are uncorrelated but could exhibit conditional heteroskedasticity or another nonlinear dynamics of unknown form (see in particular Romano and Thombs [33], Francq and Zakoian [16], Shao [34], Zhu [40], Boubacar Maïnassara and Saussereau [8], and Wang and Sun [39], among others). An ARMA equation with iid innovations is generally called *strong* ARMA model, whereas it is called *weak* ARMA when its innovations are uncorrelated but possibly dependent. Goodness-of-fit tests for weak ARMA models (see Francq et al. [15], and Zhu and Li [41]) aim at answering the following question: is a given ARMA(p, q) model adequate to take into account the *linear* dynamics of the time series at hand?

In the present paper, we do not investigate exactly the same question. We follow a more parametric approach, by considering a general d-dimensional time series model with a parametric conditional mean and a parametric conditional variance, and with iid $(\mathbf{0}, \mathbf{Id}_d)$ rescaled innovations, where $\mathbf{Id}_d$ denotes the $d \times d$ identity matrix. Our approach is, however, not fully parametric because we do not assume a particular distribution for the iid noise. The ultimate goal of the diagnostic checking tests studied in the present paper is to answer the following question: is a given nonlinear model adequate to represent the level and volatility of a given time series? However, we mainly focus on checking the adequacy of the conditional mean part.

The rest of the paper is organized as follows. In Sect. 5.2, we introduce a general nonlinear multivariate parametric model driven by a sequence of iid errors. We derive the asymptotic distributions of the Quasi Maximum Likelihood (QML) estimator and of a vector of empirical correlations of the residuals. In Sect. 5.3, we first consider portmanteau test statistics based on the residual autocorrelations. We compare the usual Box–Piece [9] statistic whose asymptotic distribution is chi-bar-square, the Li [23] statistic based on a quadratic form of the residual autocorrelations whose asymptotic distribution is chi-square, and Katayama [21] distribution-free statistic based on a projection of the autocorrelation vector. These test statistics are used to check the adequacy of the conditional mean. We also consider portmanteau test statistics based on autocorrelations of nonlinear transformation of the residuals, in order to check the adequacy of the volatility component. The properties of the tests under fixed alternatives are compared via the Bahadur approach. Numerical illustrations, including numerical evaluations of the Bahadur slopes and Monte Carlo experiments, are displayed in Sect. 5.4. Section 5.5 concludes the paper, while Sect. 5.6 contains the proofs and complementary results.

The following notations will be used throughout this paper. For a matrix $\boldsymbol{A}$ of generic term $a(i, j)$ we use the norm $\|\boldsymbol{A}\| = \sum |a(i, j)|$. The spectral radius of a square matrix $\boldsymbol{A}$ is denoted by $\rho(\boldsymbol{A})$, its trace is denoted by $\mathrm{Tr}(\boldsymbol{A})$, its transpose by $\boldsymbol{A}'$ while the diagonal matrix whose diagonal is that of $\boldsymbol{A}$ is denoted by diag $\boldsymbol{A}$. We denote by $\boldsymbol{A} \otimes \boldsymbol{B}$ the Kronecker product of two matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, vec$\boldsymbol{A}$ denotes the vector obtained by stacking the columns of $\boldsymbol{A}$, and $\boldsymbol{A}^{\otimes 2}$ stands for $\boldsymbol{A} \otimes \boldsymbol{A}$ (see, e.g., Harville [18] for more details about these matrix operators). The symbol $\stackrel{\mathcal{L}}{\to}$ denotes the convergence in distribution.

5.2 Model and Preliminaries

Let $\boldsymbol{X} = (\boldsymbol{X}_t)$ be a d-multivariate stationary process satisfying

$$\boldsymbol{X}_t = \boldsymbol{M}_{\boldsymbol{\theta}_0}(\boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \dots) + \boldsymbol{S}_{\boldsymbol{\theta}_0}(\boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \dots)\boldsymbol{\eta}_t, \tag{5.1}$$

where $\boldsymbol{\theta}_0$ is a s-dimensional vector of unknown parameters belonging to $\boldsymbol{\Theta} \subset \mathbb{R}^s$, and $\boldsymbol{S}_{\boldsymbol{\theta}_0}(\cdot)$ is almost surely (a.s.) symmetric and positive definite. It is assumed that the random variable $\boldsymbol{\eta}_t$ is independent of $\{\boldsymbol{X}_u, u < t\}$, and that $(\boldsymbol{\eta}_t)$ is an iid sequence of random vectors with mean zero and variance $\mathbf{Id}_d$, which is denoted by

$$(\boldsymbol{\eta}_t) \sim \mathrm{IID}(\mathbf{0}, \mathbf{Id}_d). \tag{5.2}$$

The first and second conditional moments of (5.1) are given by

$$\begin{aligned} \boldsymbol{M}_t(\boldsymbol{\theta}_0) &:= \boldsymbol{M}_{\boldsymbol{\theta}_0}(\boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \dots) = E(\boldsymbol{X}_t \mid \boldsymbol{X}_u, u < t), \\ \boldsymbol{\Sigma}_t(\boldsymbol{\theta}_0) &:= \boldsymbol{S}^2_{\boldsymbol{\theta}_0}(\boldsymbol{X}_{t-1}, \boldsymbol{X}_{t-2}, \dots) = \mathrm{Var}(\boldsymbol{X}_t \mid \boldsymbol{X}_u, u < t). \end{aligned}$$

Clearly, (5.2) is restrictive, but Model (5.1) is, however, quite general. Since it includes almost all the existing time series models, the practitioner is faced to the difficult problem of the choice of an appropriate specification for the conditional moments $\boldsymbol{M}_t$ and $\boldsymbol{\Sigma}_t$. In particular, there exists a huge number of generalized autoregressive conditional heteroscedastic (GARCH) models that can be employed to specify the conditional variance $\boldsymbol{\Sigma}_t$ (see Bollerslev [7] for an impressive list of more than one hundred GARCH-type models).

When the distribution of $\boldsymbol{\eta}_1$ is not specified, the unknown parameter $\boldsymbol{\theta}_0$ is generally estimated by QML. We first collect useful results concerning this estimator.

5.2.1 Asymptotic Distribution of the QML Estimator

Given observations $\boldsymbol{X}_1, \ldots, \boldsymbol{X}_n$, and initial values $\boldsymbol{X}_0 = \boldsymbol{x}_0, \boldsymbol{X}_{-1} = \boldsymbol{x}_{-1}, \ldots$, at any $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_s)' \in \boldsymbol{\Theta}$, the conditional moments $\boldsymbol{M}_t(\boldsymbol{\theta})$ and $\boldsymbol{\Sigma}_t(\boldsymbol{\theta}) =: \boldsymbol{S}_t^2(\boldsymbol{\theta})$ can be approximated by the measurable functions

$$\widetilde{\boldsymbol{M}}_t(\boldsymbol{\theta}) = \boldsymbol{M}_{\boldsymbol{\theta}}(\boldsymbol{X}_{t-1}, \ldots, \boldsymbol{X}_1, \boldsymbol{x}_0, \ldots)$$

and

$$\widetilde{\boldsymbol{\Sigma}}_t(\boldsymbol{\theta}) := \boldsymbol{S}_{\boldsymbol{\theta}}^2(\boldsymbol{X}_{t-1}, \ldots, \boldsymbol{X}_1, \boldsymbol{x}_0, \ldots) =: \widetilde{\boldsymbol{S}}_t^2(\boldsymbol{\theta}),$$

where the matrices $\boldsymbol{S}_t(\boldsymbol{\theta})$ and $\widetilde{\boldsymbol{S}}_t(\boldsymbol{\theta})$ are a.s. symmetric and positive definite. The QML estimator of $\boldsymbol{\theta}_0$ is defined as any measurable solution $\widehat{\boldsymbol{\theta}}$ of

$$\widehat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \widetilde{Q}_n(\boldsymbol{\theta}),$$

where, omitting the term "$(\boldsymbol{\theta})$" when there is no ambiguity,

$$\widetilde{Q}_n(\boldsymbol{\theta}) = \frac{1}{n} \sum_{t=1}^{n} \widetilde{\ell}_t \quad \text{and} \quad \widetilde{\ell}_t = \widetilde{\ell}_t(\boldsymbol{\theta}) = \widetilde{\boldsymbol{\epsilon}}_t' \widetilde{\boldsymbol{\Sigma}}_t^{-1} \widetilde{\boldsymbol{\epsilon}}_t + \log \det \widetilde{\boldsymbol{\Sigma}}_t,$$

with $\widetilde{\boldsymbol{\epsilon}}_t = \widetilde{\boldsymbol{\epsilon}}_t(\boldsymbol{\theta}) = \boldsymbol{X}_t - \widetilde{\boldsymbol{M}}_t(\boldsymbol{\theta})$. Let also $Q_n(\boldsymbol{\theta}) = n^{-1} \sum_{t=1}^{n} \ell_t(\boldsymbol{\theta})$, where $\ell_t(\boldsymbol{\theta}) = \boldsymbol{\epsilon}_t' \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta}) \boldsymbol{\epsilon}_t + \log \det \boldsymbol{\Sigma}_t(\boldsymbol{\theta})$, with $\boldsymbol{\epsilon}_t = \boldsymbol{\epsilon}_t(\boldsymbol{\theta}) = \boldsymbol{X}_t - \boldsymbol{M}_t(\boldsymbol{\theta})$. In the sequel, ρ denotes a generic constant belonging to [0, 1), and $K > 0$ denotes a positive constant or a positive random variable measurable with respect to the σ-field generated by $\{\boldsymbol{X}_u, u \leq 0\}$. It can be shown that the QML estimator is consistent and asymptotically normal under the following assumption:

Assumption 5.1

(i) $\boldsymbol{\Theta}$ is a compact set and the functions $\boldsymbol{\theta} \to \boldsymbol{M}_t(\boldsymbol{\theta})$ and $\boldsymbol{\theta} \to \boldsymbol{\Sigma}_t(\boldsymbol{\theta}) > 0$ are continuous;

(ii) $\{\boldsymbol{X}_t\}$ is a nonanticipative strictly stationary and ergodic solution of (5.1);

(iii) $\det \boldsymbol{\Sigma}_t(\boldsymbol{\theta}) > 0$ a.s. and $E \log^- \det \boldsymbol{\Sigma}_t(\boldsymbol{\theta}) < \infty$ for all $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, and $E \log^+ \det \boldsymbol{\Sigma}_t(\boldsymbol{\theta}_0) < \infty$ ($\log^+$ and $\log^-$ are, respectively, the positive and negative parts of the log);

(iv) $E\|\boldsymbol{X}_t\|^r < \infty$ and $E \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\boldsymbol{M}_t(\boldsymbol{\theta})\|^r < \infty$ for some $r > 0$. Furthermore, $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\widetilde{\boldsymbol{S}}_t^{-1}(\boldsymbol{\theta})\| \leq K$, $\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \|\boldsymbol{S}_t^{-1}(\boldsymbol{\theta})\| \leq K$ and for some $\rho \in (0, 1)$, $\rho^{-t} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left|\boldsymbol{M}_t(\boldsymbol{\theta}) - \widetilde{\boldsymbol{M}}_t(\boldsymbol{\theta})\right| \to 0$ and $\rho^{-t} \sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left|\boldsymbol{S}_t(\boldsymbol{\theta}) - \widetilde{\boldsymbol{S}}_t(\boldsymbol{\theta})\right| \to 0$ a.s.;

(v) if $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$ then $\boldsymbol{M}_t(\boldsymbol{\theta}) \neq \boldsymbol{M}_t(\boldsymbol{\theta}_0)$ or $\boldsymbol{\Sigma}_t(\boldsymbol{\theta}) \neq \boldsymbol{\Sigma}_t(\boldsymbol{\theta}_0)$ with non zero probability;

(vi) $\boldsymbol{\theta}_0$ belongs to the interior $\overset{\circ}{\boldsymbol{\Theta}}$ of $\boldsymbol{\Theta}$;

(vii) $\boldsymbol{\theta} \to \boldsymbol{M}_t(\boldsymbol{\theta})$ and $\boldsymbol{\theta} \to \boldsymbol{\Sigma}_t(\boldsymbol{\theta})$ admit continuous second order derivatives;

(viii) for some neighborhood $\mathcal{V}(\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}_0$, we have

$$\sup_{\boldsymbol{\theta}\in\mathcal{V}(\boldsymbol{\theta}_0)} \sqrt{n} \left\| \frac{\partial Q_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} - \frac{\partial \widetilde{Q}_n(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \right\| = o_P(1) \tag{5.3}$$

and, for all $i, j \in \{1, \ldots, m\}$ and some $r_k > 0$, $k \in \{1, \ldots, 5\}$, such that $2r_2^{-1} + 2r_3^{-1} \leq 1, r_1^{-1} + 2r_3^{-1} \leq 1, r_2^{-1} + r_3^{-1} + r_5^{-1} \leq 1, r_3^{-1} + r_4^{-1} \leq 1$, and $r_5 \geq 2$, we have

$$E \sup_{\boldsymbol{\theta}\in\mathcal{V}(\boldsymbol{\theta}_0)} \left\| \boldsymbol{S}_t^{-1}(\boldsymbol{\theta}) \frac{\partial^2 \boldsymbol{\Sigma}_t(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} \boldsymbol{S}_t^{-1}(\boldsymbol{\theta}) \right\|^{r_1} < \infty, \tag{5.4}$$

$$E \sup_{\boldsymbol{\theta}\in\mathcal{V}(\boldsymbol{\theta}_0)} \left\| \boldsymbol{S}_t^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{\Sigma}_t(\boldsymbol{\theta})}{\partial\theta_i} \boldsymbol{S}_t^{-1}(\boldsymbol{\theta}) \right\|^{r_2} < \infty, \tag{5.5}$$

$$E \sup_{\boldsymbol{\theta}\in\mathcal{V}(\boldsymbol{\theta}_0)} \left\| \boldsymbol{S}_t^{-1}(\boldsymbol{\theta}) \boldsymbol{S}_t(\boldsymbol{\theta}_0) \right\|^{r_3} < \infty, \tag{5.6}$$

$$E \sup_{\boldsymbol{\theta}\in\mathcal{V}(\boldsymbol{\theta}_0)} \left\| \boldsymbol{S}_t^{-1}(\boldsymbol{\theta}) \frac{\partial^2 \boldsymbol{M}_t(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j} \right\|^{r_4} < \infty, \tag{5.7}$$

$$E \sup_{\boldsymbol{\theta}\in\mathcal{V}(\boldsymbol{\theta}_0)} \left\| \boldsymbol{S}_t^{-1}(\boldsymbol{\theta}) \frac{\partial \boldsymbol{M}_t(\boldsymbol{\theta})}{\partial\theta_i} \right\|^{r_5} < \infty, \tag{5.8}$$

where θ_i denotes the i-th component of the vector $\boldsymbol{\theta}$;

(ix) $E\|\boldsymbol{\eta}_0\|^4 < \infty$, and the information matrices

$$\boldsymbol{I} = E\left[\{\partial\ell_t(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}\}\left\{\partial\ell_t(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}'\right\}\right] \text{ and } \boldsymbol{J} = E\left\{\partial^2\ell_t(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'\right\}$$

are non singular.

To illustrate Assumption 5.1, let us consider the following simple example.

Example 5.1 Consider the univariate ARMA(1,1)-GARCH(1,1) model defined by

$$\begin{cases} X_t - a_0 X_{t-1} = c_0 + \epsilon_t + b_0\epsilon_{t-1}, \\ \epsilon_t = \sigma_t\eta_t, \quad \sigma_t^2 = \omega_0 + \alpha_0\epsilon_{t-1}^2 + \beta_0\sigma_{t-1}^2. \end{cases}$$

Since $d = 1$, the random variables are not written in boldface. When the unknown parameter $\boldsymbol{\theta}_0 = (c_0, a_0, b_0, \omega_0, \alpha_0, \beta_0)'$ belongs to the compact set

$$\boldsymbol{\Theta} \subset (-\infty, +\infty) \times (-1, 1)^2 \times (0, +\infty)^2 \times [0, 1),$$

with $E\log(\alpha_0\eta_1^2 + \beta_0) < 0$ and $a_0 \neq b_0$, and when the support of the law of η_1 contains more than 2 points, Assumption 5.1(i)–(v) required for the consistency are satisfied. If, in addition, $\boldsymbol{\theta}_0$ belongs to the interior of $\boldsymbol{\Theta}$ and

$$2\alpha_0\beta_0 + \beta_0^2 + \alpha_0^2 E\eta_1^4 < 1, \tag{5.9}$$

it can be shown, using Francq and Zakoïan [17], that Assumption 5.1 is entirely satisfied. In particular, (5.4)–(5.6) hold true for arbitrary large values of r_1, r_2, and r_3 (see Francq and Zakoïan ([17]; (4.25) and (4.29))) and for $r_3 = r_4 = 4$. Note that (5.9) is the necessary and sufficient condition for $E\epsilon_1^4 < \infty$, and that moments assumptions are indeed required for the existence of $\boldsymbol{I}$ and $\boldsymbol{J}$ in this framework (see Francq and Zakoïan ([17]; Remark 3.5)).

Write $a \stackrel{c}{=} b$ when $a = b + c$. The following lemma states the asymptotic behavior of the QMLE. Similar results can be found elsewhere in the literature under slightly different assumptions (see, e.g., Pötscher and Prucha [32]). For the sake of self-containedness, the proof is however given in Sect. 5.6.

Lemma 5.1 *Assume (5.1)–(5.2). Under Assumption 5.1(i)–(v), $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}_0$ a.s. while under Assumption 5.1,*

$$\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \stackrel{o_P(1)}{=} -\boldsymbol{J}^{-1} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \boldsymbol{Z}_t \stackrel{\mathcal{L}}{\to} \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}_{\boldsymbol{\theta}}) \tag{5.10}$$

as $n \to \infty$, where $\boldsymbol{\Sigma}_{\boldsymbol{\theta}} := \boldsymbol{J}^{-1}\boldsymbol{I}\boldsymbol{J}^{-1}$ and $\boldsymbol{Z}_t = \partial \ell_t(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta}$. Letting $\boldsymbol{Z}_t = (Z_{1t}, \ldots, Z_{st})'$, we have

$$Z_{it} = -2 \frac{\partial \boldsymbol{M}_t'(\boldsymbol{\theta}_0)}{\partial \theta_i} \boldsymbol{S}_t^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\eta}_t + \mathrm{Tr}\left\{ \boldsymbol{S}_t^{-1}(\boldsymbol{\theta}_0) \frac{\partial \boldsymbol{\Sigma}_t(\boldsymbol{\theta}_0)}{\partial \theta_i} \boldsymbol{S}_t^{-1}(\boldsymbol{\theta}_0) \left(\boldsymbol{Id}_d - \boldsymbol{\eta}_t \boldsymbol{\eta}_t' \right) \right\}.$$

5.2.2 *Asymptotic Distribution of the Residuals Empirical Autocorrelations*

Define the standardized residuals

$$\widehat{\boldsymbol{\eta}}_t = \widetilde{\boldsymbol{S}}_t^{-1}(\widehat{\boldsymbol{\theta}})\widetilde{\boldsymbol{\epsilon}}_t(\widehat{\boldsymbol{\theta}}), \quad t = 1, \ldots, n.$$

For some $m \geq 1$, define the vectors of sample autocovariances and autocorrelations

$$\widehat{\boldsymbol{\gamma}}_m = \left(\left\{ \mathrm{vec}\widehat{\boldsymbol{\Gamma}}(1) \right\}', \ldots, \left\{ \mathrm{vec}\widehat{\boldsymbol{\Gamma}}(m) \right\}' \right)',$$
$$\widehat{\boldsymbol{\rho}}_m = \left(\left\{ \mathrm{vec}\widehat{\boldsymbol{R}}(1) \right\}', \ldots, \left\{ \mathrm{vec}\widehat{\boldsymbol{R}}(m) \right\}' \right)',$$

where, for $0 \leq \ell < n$,

$$\widehat{\boldsymbol{\Gamma}}(\ell) = \frac{1}{n} \sum_{t=\ell+1}^{n} \widehat{\boldsymbol{\eta}}_t \widehat{\boldsymbol{\eta}}_{t-\ell}' \quad \text{and} \quad \widehat{\boldsymbol{R}}(\ell) = \left\{ \mathrm{diag}\, \widehat{\boldsymbol{\Gamma}}(0) \right\}^{-1/2} \widehat{\boldsymbol{\Gamma}}(\ell) \left\{ \mathrm{diag}\, \widehat{\boldsymbol{\Gamma}}(0) \right\}^{-1/2}.$$

Let $\boldsymbol{\Upsilon}_t = \boldsymbol{\eta}_{t-1:t-m} \otimes \boldsymbol{\eta}_t$, where $\boldsymbol{\eta}_{t-1:t-m} = (\boldsymbol{\eta}'_{t-1}, \ldots, \boldsymbol{\eta}'_{t-m})'$ and define the $s \times md^2$ and $d^2 \times s$ matrices

$$\boldsymbol{\Sigma}_{\boldsymbol{\theta}\boldsymbol{\Upsilon}} = -\boldsymbol{J}^{-1} E \boldsymbol{Z}_t \boldsymbol{\Upsilon}'_t \quad \text{and} \quad \boldsymbol{c}_\ell = -E\boldsymbol{\eta}_{t-\ell} \otimes \boldsymbol{S}_t^{-1}(\boldsymbol{\theta}_0) \frac{\partial \boldsymbol{M}_t(\boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}'}. \tag{5.11}$$

Define also the $md^2 \times s$ matrix $\boldsymbol{C}_m = (\boldsymbol{c}'_1 \boldsymbol{c}'_2 \cdots \boldsymbol{c}'_m)'$.

The portmanteau tests studied in the next section are quadratic forms of $\widehat{\boldsymbol{\gamma}}_m$ and $\widehat{\boldsymbol{\rho}}_m$. The following proposition gives the asymptotic distribution of these vectors. Its proof is given in Sect. 5.6.

Proposition 5.1 *Under the assumptions of Lemma 5.1, we have*

$$\sqrt{n}\widehat{\boldsymbol{\gamma}}_m \xrightarrow{\mathcal{L}} \mathcal{N}\left\{\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_m}\right\} \quad \text{and} \quad \sqrt{n}\widehat{\boldsymbol{\rho}}_m \xrightarrow{\mathcal{L}} \mathcal{N}\left\{\mathbf{0}, \boldsymbol{\Sigma}_{\boldsymbol{\rho}_m}\right\} \tag{5.12}$$

as $n \to \infty$, where

$$\boldsymbol{\Sigma}_{\boldsymbol{\rho}_m} = \boldsymbol{\Sigma}_{\boldsymbol{\gamma}_m} = \boldsymbol{Id}_{md^2} + \boldsymbol{C}_m \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{C}'_m + \boldsymbol{C}_m \boldsymbol{\Sigma}_{\boldsymbol{\theta}\boldsymbol{\Upsilon}} + \boldsymbol{\Sigma}'_{\boldsymbol{\theta}\boldsymbol{\Upsilon}} \boldsymbol{C}'_m. \tag{5.13}$$

As usual, the information matrices $\boldsymbol{I}$ and $\boldsymbol{J}$ are consistently estimated by their empirical counterparts

$$\widehat{\boldsymbol{I}} = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial \widetilde{\ell}_t(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \frac{\partial \widetilde{\ell}_t(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}'} \quad \text{and} \quad \widehat{\boldsymbol{J}} = \frac{1}{n}\sum_{t=1}^{n} \frac{\partial^2 \widetilde{\ell}_t(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}'}.$$

A strongly consistent estimator $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\rho}_m}$ is obtained by using similar empirical estimators for the other matrices involved in $\boldsymbol{\Sigma}_{\boldsymbol{\rho}_m}$. More precisely, letting

$$\widehat{\boldsymbol{C}}_m = (\widehat{\boldsymbol{c}}'_1 \widehat{\boldsymbol{c}}'_2 \cdots \widehat{\boldsymbol{c}}'_m)', \qquad \widehat{\boldsymbol{c}}_\ell = -\frac{1}{n}\sum_{t=\ell+1}^{n} \widehat{\boldsymbol{\eta}}_{t-\ell} \otimes \widetilde{\boldsymbol{S}}_t^{-1}(\widehat{\boldsymbol{\theta}}) \frac{\partial \widetilde{\boldsymbol{M}}_t(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}'},$$

one can use the estimator

$$\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\rho}_m} = \left(\widehat{\boldsymbol{C}}_m \mathbf{Id}_{md^2}\right) \frac{1}{n}\sum_{t=1}^{n} \begin{pmatrix} -\widehat{\boldsymbol{J}}^{-1} \frac{\partial \widetilde{\ell}_t(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \\ \widehat{\boldsymbol{\Upsilon}}_t \end{pmatrix} \left(-\frac{\partial \widetilde{\ell}_t(\widehat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}'} \widehat{\boldsymbol{J}}^{-1} \widehat{\boldsymbol{\Upsilon}}'_t\right) \begin{pmatrix} \widehat{\boldsymbol{C}}'_m \\ \mathbf{Id}_{md^2} \end{pmatrix}, \tag{5.14}$$

where $\widehat{\boldsymbol{\Upsilon}}_t = \widehat{\boldsymbol{\eta}}_{t-1:t-m} \otimes \widehat{\boldsymbol{\eta}}_t$ with $\widehat{\boldsymbol{\eta}}_{t-1:t-m} = (\widehat{\boldsymbol{\eta}}'_{t-1}, \ldots, \widehat{\boldsymbol{\eta}}'_{t-m})'$ and $\widehat{\boldsymbol{\eta}}_t = \mathbf{0}$ for $t \leq 0$.

By construction, the estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\rho}_m}$ defined in (5.14) is a semi-positive definite matrix for finite n. On the other hand, the plug-in estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\rho}_m}$ obtained by using consistent estimators of the matrices involved in the right-hand side of (5.13) may result in a matrix that is not semi-positive definite in finite sample, and even asymptotically when the model is misspecified (see Sect. 5.4.1).

5.3 Different Portmanteau Goodness-of-Fit Tests

In order to determine whether or not the parametric specifications chosen for $\boldsymbol{M}_t$ and $\boldsymbol{\Sigma}_t$ are appropriate, a common practice in time series is to look at the residuals. Visual inspection of the residuals autocorrelograms, as well as more formal portmanteau tests based on quadratic forms of the residual autocorrelations, have been found useful for checking the adequacy of $\boldsymbol{M}_t$. We first consider such tests, and then turn to portmanteau tests based on autocorrelations of nonlinear transformations of the residuals for checking the adequacy of the volatility component $\boldsymbol{\Sigma}_t$.

5.3.1 Portmanteau Test Statistics

For goodness-of-fit checking of univariate ARMA models, Box and Pierce [9] introduced the so-called portmanteau test statistics

$$Q_m^{BP} = n\widehat{\boldsymbol{\rho}}_m'\widehat{\boldsymbol{\rho}}_m. \tag{5.15}$$

McLeod [29] showed that, for an ARMA(p, q) model with iid errors, the distribution of Q_m^{BP} can be approximated by a $\chi^2_{m-(p+q)}$, for m sufficiently large. Ljung and Box [27] proposed a modified portmanteau test which has the same asymptotic distribution as Q_m^{BP}. Multivariate versions of these portmanteau statistics have been introduced by Chitturi [11] and Hosking [19].

For general nonlinear models, the (asymptotic) distribution of Q_m^{BP} is often badly approximated by a χ^2 (see, e.g., Duchesne and Francq [12]). To solve this problem, Li [23, 24] considered, in a univariate framework, the test statistic

$$Q_m^{Li} = n\widehat{\boldsymbol{\rho}}_m'\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\rho}_m}^{-1}\widehat{\boldsymbol{\rho}}_m, \tag{5.16}$$

which follows asymptotically a $\chi^2_{md^2}$ distribution under the assumption of Proposition 5.1 when $\boldsymbol{\Sigma}_{\boldsymbol{\rho}_m}$ is invertible.

In the strong ARMA framework, Katayama [12] proposed a test statistic based on the fact that the projection of $\widehat{\boldsymbol{\gamma}}_m$ on the orthogonal of the column space of $\boldsymbol{C}_m$ does not depend on $\widehat{\boldsymbol{\theta}}$, and thus has a simple asymptotic distribution. To state this result more precisely, we need to introduce some notation. Assume that $\boldsymbol{C}_m$ has full rank s. Let $\boldsymbol{D}_m$ be another $md^2 \times s$ matrix with rank s, such that $\boldsymbol{D}_m'\boldsymbol{C}_m$ be invertible. The projection on the column space of $\boldsymbol{C}_m$ orthogonally to the column space of $\boldsymbol{D}_m$ is defined by $\boldsymbol{P}_{\boldsymbol{C}\perp\boldsymbol{D}} = \boldsymbol{C}_m(\boldsymbol{D}_m'\boldsymbol{C}_m)^{-1}\boldsymbol{D}_m'$. Define the projection $\boldsymbol{M}_{\boldsymbol{C}\perp\boldsymbol{D}} = \mathbf{Id}_{md^2} - \boldsymbol{P}_{\boldsymbol{C}\perp\boldsymbol{D}}$ and let $\widehat{\boldsymbol{M}}_{\boldsymbol{C}\perp\boldsymbol{D}}$ be the empirical estimator of $\boldsymbol{M}_{\boldsymbol{C}\perp\boldsymbol{D}}$ (or any weakly consistent estimator of this matrix). The following result shows that Katayama's test statistic can by used for the general nonlinear model (5.1)–(5.2).

Proposition 5.2 *Under the assumptions of Lemma 5.1, if $\boldsymbol{C}_m$ and $\boldsymbol{D}_m$ have full rank s, the test statistic*

$$Q_m^K = n\widehat{\boldsymbol{\rho}}_m' \widehat{\boldsymbol{M}}_{\boldsymbol{C}\perp\boldsymbol{D}} \widehat{\boldsymbol{\rho}}_m$$

converges weakly to a $\chi^2_{md^2-s}$ *random variable as* $n \to \infty$.

Katayama [12] employed orthogonal projections (i.e., $\boldsymbol{D}_m = \boldsymbol{C}_m$), but it will be shown in Sect. 5.3.2 that the use of oblique projections can lead to efficiency gains. The assumption that $\boldsymbol{C}_m$ has full rank s is quite restrictive. In particular, it precludes the use of the test statistic Q_m^K for $md^2 < s$. Another frequent situation where the rank of $\boldsymbol{C}_m$ is strictly smaller than s is when, as for an ARMA-GARCH model, $\boldsymbol{M}_t$ and $\boldsymbol{\Sigma}_t$ depend, respectively, of the first s_0 and last s_1 components of $\boldsymbol{\theta}_0$. In such a case, the rank of $\boldsymbol{C}_m$ is not larger than s_0. To solve this problem, let us introduce the projection

$$\widehat{\boldsymbol{M}}_+ = \mathbf{Id}_{md^2} - \widehat{\boldsymbol{P}}_+, \qquad \lim_{n\to\infty} \widehat{\boldsymbol{P}}_+ = \boldsymbol{P}_+ := \boldsymbol{C}_m(\boldsymbol{C}_m'\boldsymbol{C}_m)^+\boldsymbol{C}_m' \quad \text{a.s.},$$

where $\boldsymbol{A}^+$ denotes the Moore–Penrose pseudoinverse of a matrix $\boldsymbol{A}$. Note that the estimator $\widehat{\boldsymbol{P}}_+$ cannot be defined by directly substituting $\widehat{\boldsymbol{C}}_m$ for $\boldsymbol{C}_m$ in $\boldsymbol{P}_+$. Indeed, the convergence of $\boldsymbol{A}_n$ to $\boldsymbol{A}$ does not entail the convergence of $\boldsymbol{A}_n^+$ to $\boldsymbol{A}^+$ (see Andrews [3]). For a $m \times m$ positive semidefinite symmetric matrix $\boldsymbol{A}$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_m$ and spectral decomposition $\boldsymbol{A} = \boldsymbol{P}\boldsymbol{\Lambda}\boldsymbol{P}'$, where $\boldsymbol{\Lambda} = \text{diag}(\lambda_1, \ldots, \lambda_m)$ and $\boldsymbol{P}\boldsymbol{P}' = \mathbf{Id}_m$, and for any $i \leq s_0 = \text{rank}(\boldsymbol{A})$, define the so-called {2}-inverse $\boldsymbol{A}^{-_i} = \boldsymbol{P}\text{diag}(\lambda_1^{-1}, \ldots, \lambda_i^{-1}, \mathbf{0}_{m-i}')\boldsymbol{P}'$. The name "{2}-inverse" comes from the property $\boldsymbol{A}^{-_i}\boldsymbol{A}\boldsymbol{A}^{-_i} = \boldsymbol{A}^{-_i}$, which is the second of the four requirements of the definition of the Moore–Penrose pseudoinverse. Using the fact that $\lim_{n\to\infty} \boldsymbol{A}_n = \boldsymbol{A}$ entails $\lim_{n\to\infty} \boldsymbol{A}_n^{-_{s_0}} = \boldsymbol{A}^{-_{s_0}} = \boldsymbol{A}^+$, one can take the estimator

$$\widehat{\boldsymbol{P}}_+ = \widehat{\boldsymbol{C}}_m(\widehat{\boldsymbol{C}}_m'\widehat{\boldsymbol{C}}_m)^{-_{s_0}}\widehat{\boldsymbol{C}}_m'.$$

Proposition 5.3 *Under the assumptions of Lemma 5.1, if* $\boldsymbol{C}_m$ *has the rank* s_0*, the test statistic*

$$Q_m^{K+} = n\widehat{\boldsymbol{\rho}}_m' \widehat{\boldsymbol{M}}_+ \widehat{\boldsymbol{\rho}}_m \tag{5.17}$$

converges weakly to a $\chi^2_{md^2-s_0}$ *random variable as* $n \to \infty$.

5.3.2 Bahadur Asymptotic Relative Efficiency

We now compare the asymptotic behaviors of the previous portmanteau tests by using Bahadur's approach. This approach considers fixed alternatives, and compares the rates at which the p-values converge to zero (see, e.g., van der Vaart [37] for details).

The approximate p-values of two tests Q_1 and Q_2 are defined by $\Lambda_i(Q_i)$, where $\Lambda_i(t) = \lim_{t\to\infty} P_{H_0}(Q_i > t)$, for $i = 1, 2$. In the sense of Bahadur [4], the (approximate) slope of the test Q_i is defined by $c_i := \lim_{n\to\infty} -2\log \Lambda_i(Q_i)/n$, under alternatives such that this limit exits in probability. The asymptotic relative efficiency (ARE) of the test Q_1 with respect to Q_2 is then defined as the ratio of the slopes: $\mathrm{ARE}(Q_1 \mid Q_2) = c_1/c_2$. We say that Q_1 is asymptotically more efficient than Q_2 when $\mathrm{ARE}(Q_1 \mid Q_2) > 1$.

In view of Lemma 5.1, the residual autocorrelations of the estimated model (5.1) generally satisfy, as $n \to \infty$

$$\widehat{\boldsymbol{\rho}}_m \to \boldsymbol{\rho}_m^* \qquad \text{and} \qquad \widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\rho}_m} \to \boldsymbol{\Sigma}_{\boldsymbol{\rho}_m^*} \neq \mathbf{0} \quad \text{a.s.} \tag{5.18}$$

for some $\boldsymbol{\rho}_m^*$ such that $\boldsymbol{\rho}_m^* = \mathbf{0}$ when $\boldsymbol{X}$ does satisfy (5.1). In general, (5.18) still holds true when the model is misspecified, but possibly with $\boldsymbol{\rho}_m^* \neq \mathbf{0}$ (and $\boldsymbol{\Sigma}_{\boldsymbol{\rho}_m^*} \neq \boldsymbol{\Sigma}_{\boldsymbol{\rho}_m}$). We now give a very simple example of such a situation.

Example 5.2 Assume that an AR(1) model $X_t = aX_{t-1} + \sigma\eta_t$ is fitted to a non-degenerated stationary and ergodic sequence (X_t) that admits finite moments of order two, but is not necessarily an AR(1) process. By the ergodic theorem, the QMLE of the AR(1) coefficient a converges to $\rho_X(1)$, where $\rho_X(\cdot)$ denotes the autocorrelation function of (X_t), and the QMLE of σ^2 converges to

$$\sigma_0^2 := E(X_t - \rho_X(1)X_{t-1})^2 > 0.$$

For $t \geq 2$, the QML residuals thus satisfy $\hat{\eta}_t = \sigma_0^{-1}\{X_t - \rho_X(1)X_{t-1}\} + o_p(1)$ as $n \to \infty$. The first convergence in (5.18) thus holds with $\boldsymbol{\rho}_m^* = (\rho_\eta(1), \ldots, \rho_\eta(m))'$ and

$$\rho_\eta(h) = \frac{\gamma_X(h) - \rho_X(1)\gamma_X(h-1) + \rho_X^2(1)\gamma_X(h) - \rho_X(1)\gamma_X(h+1)}{\gamma_X(0) - 2\rho_X(1)\gamma_X(1) + \rho_X^2(1)\gamma_X(0)}.$$

Obviously, when X_t follows a weak AR(1) model, we retrieve $\boldsymbol{\rho}_m^* = \mathbf{0}$. If, for instance, X_t follows a MA(1) model of the form $X_t = \epsilon_t + b\epsilon_{t-1}$, then we have

$$\boldsymbol{\rho}_m^* = \left(\frac{b^3}{(1+b^2)(1+b^2+b^4)}, \frac{-b^2}{1+b^2+b^4}, 0, \ldots, 0\right)'.$$

We now give a simple example that is not related to ARMA processes.

Example 5.3 Assume that $X_t = \prod_{\ell=0}^{k} |\eta_{t-\ell}|^j$, where k and j are positive integers and (η_t) is an independent sequence of $\mathcal{N}(0,1)$ distributed random variables. We then have, for $h = 0, \ldots, k$,

$$\gamma_X(h) = \mu_j^{2h}\left(\mu_{2j}^{k+1-h} - \mu_j^{2(k+1-h)}\right), \qquad \mu_j = \frac{2^{j/2}\Gamma(\frac{j+1}{2})}{\sqrt{\pi}},$$

and $\gamma_X(h) = 0$ for $h > k$, from which $\boldsymbol{\rho}_m^*$ is obtained as a function of j, k, and m.

The previous discussion leads us to consider the testing problem

$$H_0: \quad \boldsymbol{\rho}_m^* = \mathbf{0} \qquad \text{against} \qquad H_1: \quad \boldsymbol{\rho}_m^* \neq \mathbf{0}. \tag{5.19}$$

Let $N_1, \ldots, N_{md^2}$ be independent $\mathcal{N}(0, 1)$ random variables, and let $\widehat{\lambda}_1 \geq \cdots \geq \widehat{\lambda}_{md^2}$ be the eigenvalues of $\widehat{\boldsymbol{\Sigma}}_{\boldsymbol{\rho}_m}$. Denote by $k_\alpha = k_\alpha\left(\widehat{\lambda}_1, \ldots, \widehat{\lambda}_{md^2}\right)$ the $(1-\alpha)$-quantile of $\sum_{i=1}^{md^2} \widehat{\lambda}_i N_i^2$, and by $\chi^2_{\nu,\alpha}$ the α-quantile of the χ^2_ν distribution. The algorithm developed by Imhof [20] can be used to compute k_α (see Duchesne and Lafaye de Micheaux [13]). The following proposition specifies the critical regions of the portmanteau tests and gives their Bahadur slopes.

Proposition 5.4 *Under the assumptions of Lemma 5.1, the following statements hold:*

(i) Under H_0 in (5.19), (5.18) holds and the tests with critical regions

$$\left\{Q_m^{BP} > k_\alpha\right\} \quad \textit{and} \quad \left\{Q_m^{K+} > \chi^2_{md^2-s_0,\alpha}\right\}$$

have asymptotic level $\alpha \in (0, 1)$. Under the additional assumption that $\boldsymbol{C}_m$ and $\boldsymbol{D}_m$ have full rank s, the test

$$\left\{Q_m^{K} > \chi^2_{md^2-s,\alpha}\right\},$$

and under the assumption that $\boldsymbol{\Sigma}_{\boldsymbol{\rho}_m^*}$ is invertible, the test

$$\left\{Q_m^{Li} > \chi^2_{md^2,\alpha}\right\}$$

also have the asymptotic level α;

(ii) Under H_1 and assuming that (5.18) holds, the BP and Li tests are consistent and have the respective Bahadur slopes

$$c_{BP} = \frac{\boldsymbol{\rho}_m^{*\prime}\boldsymbol{\rho}_m^*}{\lambda_1} \quad \textit{and} \quad c_{Li} = \boldsymbol{\rho}_m^{*\prime}\boldsymbol{\Sigma}_{\boldsymbol{\rho}_m^*}^{-1}\boldsymbol{\rho}_m^*,$$

where $\lambda_1 \neq 0$ is the largest eigenvalue of $\boldsymbol{\Sigma}_{\boldsymbol{\rho}_m^*}$. The test based on Q_m^{Li} is asymptotically more efficient than that based on Q_m^{BP} in the sense that $ARE(Q_{Li} \mid Q_{BP}) \geq 1$ for all $\boldsymbol{\rho}_m^* \neq 0$, with equality if and only if $\boldsymbol{\rho}_m^*$ belongs to the first eigenvector space of $\boldsymbol{\Sigma}_{\boldsymbol{\rho}_m^*}$. On the other hand, assume that a.s. $\widehat{\boldsymbol{M}}_{\boldsymbol{C}\perp\boldsymbol{D}} = \boldsymbol{Id}_{md^2} - \widehat{\boldsymbol{C}}_m(\widehat{\boldsymbol{D}}_m'\widehat{\boldsymbol{C}}_m)^{-1}\widehat{\boldsymbol{D}}_m'$, with $\widehat{\boldsymbol{D}}_m \to \boldsymbol{D}_m^*$ and $\widehat{\boldsymbol{C}}_m \to \boldsymbol{C}_m^*$. Then if $\boldsymbol{D}_m^*$ and $\boldsymbol{C}_m^*$ have full rank s, for alternatives $\boldsymbol{\rho}_m^*$ that do not belong to the column space of $\boldsymbol{C}_m^*$, the Q_m^K-test is consistent and has Bahadur slope

$$c_K = \boldsymbol{\rho}_m^{*\prime}\boldsymbol{M}_{\boldsymbol{C}^*\perp\boldsymbol{D}^*}\boldsymbol{\rho}_m^*, \quad \boldsymbol{M}_{\boldsymbol{C}^*\perp\boldsymbol{D}^*} = \boldsymbol{Id}_{md^2} - \boldsymbol{C}_m^*(\boldsymbol{D}_m^{*\prime}\boldsymbol{C}_m^*)^{-1}\boldsymbol{D}_m^{*\prime}.$$

The optimal slope $c_{K,opt} = \boldsymbol{\rho}_m^{*\prime}\boldsymbol{\rho}_m^*$ is obtained when $\boldsymbol{D}_m^*$ belongs to the orthogonal of the linear space generated by $\boldsymbol{\rho}_m^*$. For alternatives, $\boldsymbol{\rho}_m^*$ that do not belong to the column space of $\boldsymbol{C}_m^*(\boldsymbol{C}_m^{*\prime}\boldsymbol{C}_m^*)^+$, the Q_m^{K+}-test is consistent and has Bahadur *slope*

$$c_{K+} = \boldsymbol{\rho}_m^{*\prime} \boldsymbol{M}_+^* \boldsymbol{\rho}_m^*, \quad \boldsymbol{M}_+^* = \boldsymbol{Id}_{md^2} - \boldsymbol{C}_m^* (\boldsymbol{C}_m^{*'} \boldsymbol{C}_m^*)^+ \boldsymbol{C}_m^{*'}.$$

5.3.3 Nonlinear Transformations of the Residuals

The portmanteau tests based on the residual autocovariances are irrelevant for detecting certain nonlinearities in the conditional mean or for testing the adequacy of the volatility model $\boldsymbol{\Sigma}_t(\boldsymbol{\theta})$. Indeed, assuming for simplicity that there is no conditional mean in the model (i.e., $\widetilde{\boldsymbol{M}}_t(\boldsymbol{\theta}) \equiv \boldsymbol{0}$) and that $\boldsymbol{X}_t = \boldsymbol{S}_{0t}\boldsymbol{\eta}_t$, the process $\widehat{\boldsymbol{\eta}}_t = \widetilde{\boldsymbol{S}}_t^{-1}(\widehat{\boldsymbol{\theta}})\boldsymbol{X}_t$ is uncorrelated whatever the volatility model $\boldsymbol{S}_t(\cdot)$, since $\widetilde{\boldsymbol{S}}_t^{-1}(\widehat{\boldsymbol{\theta}})\boldsymbol{S}_{0t}$ is always independent of $\boldsymbol{\eta}_t$. For detecting nonlinearities in the conditional mean, McLeod and Li [30] proposed portmanteau tests based on the autocorrelations of squared ARMA residuals. For checking the adequacy of ARCH-type models, Li and Mak (1994) and Ling and Li [18] proposed portmanteau tests based on the autocovariances of the squared rescaled residuals. Other authors suggested using portmanteau tests based on transformations of the residuals, such as the absolute values or the log of the squared (see Pérez and Ruiz [31] and Fisher and Gallagher [14]). The idea behind these tests is that for any function $\boldsymbol{g} : \mathbb{R}^d \to \mathbb{R}^{d_0}$ with $d_0 \geq 1$, the autocorrelations of $\widehat{\boldsymbol{g}}_t = \boldsymbol{g}\left(\widehat{\boldsymbol{\eta}}_t\right)$ should be close to zero when the model is well specified. We thus define the $d_0 \times d_0$ matrices $\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{g}}(\ell) = \frac{1}{n}\sum_{t=\ell+1}^n \left(\widehat{\boldsymbol{g}}_t - \overline{\boldsymbol{g}}_n\right)\left(\widehat{\boldsymbol{g}}_{t-\ell} - \overline{\boldsymbol{g}}_n\right)'$ and $\widehat{\boldsymbol{R}}_{\boldsymbol{g}}(\ell) = \left\{\mathrm{diag}\,\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{g}}(0)\right\}^{-1/2} \widehat{\boldsymbol{\Gamma}}_{\boldsymbol{g}}(\ell) \left\{\mathrm{diag}\,\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{g}}(0)\right\}^{-1/2}$ for $0 \leq \ell < n$, where $\overline{\boldsymbol{g}}_n = \frac{1}{n}\sum_{t=1}^n \widehat{\boldsymbol{g}}_t$. Let $\widehat{\boldsymbol{\gamma}}_m^{\boldsymbol{g}} = \left(\left\{\mathrm{vec}\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{g}}(1)\right\}', \ldots, \left\{\mathrm{vec}\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{g}}(m)\right\}'\right)'$, $\widehat{\boldsymbol{\rho}}_m^{\boldsymbol{g}} = \left(\left\{\mathrm{vec}\widehat{\boldsymbol{R}}_{\boldsymbol{g}}(1)\right\}', \ldots, \left\{\mathrm{vec}\widehat{\boldsymbol{R}}_{\boldsymbol{g}}(m)\right\}'\right)'$ and $\boldsymbol{\Upsilon}_t^{\boldsymbol{g}} = \boldsymbol{\eta}_{t-1:t-m}^{\boldsymbol{g}} \otimes \boldsymbol{\eta}_t^{\boldsymbol{g}}$, with $\boldsymbol{\eta}_t^{\boldsymbol{g}} = \boldsymbol{g}(\boldsymbol{\eta}_t) - E\boldsymbol{g}(\boldsymbol{\eta}_1)$. Define the $md_0^2 \times md_0^2$, $s \times md_0^2$ and $d_0^2 \times s$ matrices

$$\boldsymbol{\Omega}_{\boldsymbol{g}} = E\boldsymbol{\eta}_t^{\boldsymbol{g}}\boldsymbol{\eta}_t^{\boldsymbol{g}'}, \quad \boldsymbol{\Sigma}_{\boldsymbol{\theta}\boldsymbol{\Upsilon}^{\boldsymbol{g}}} = -\boldsymbol{J}^{-1}E\boldsymbol{Z}_t\boldsymbol{\Upsilon}_t^{\boldsymbol{g}'} \quad \text{and} \quad \boldsymbol{c}_\ell^{\boldsymbol{g}} = E\boldsymbol{\eta}_{t-\ell}^{\boldsymbol{g}} \otimes \frac{\partial \boldsymbol{\eta}_t^{\boldsymbol{g}}}{\partial \boldsymbol{\theta}'}. \tag{5.20}$$

Define also the $md_0^2 \times s$ matrix $\boldsymbol{C}_m^{\boldsymbol{g}} = (\boldsymbol{c}_1^{\boldsymbol{g}'}\ \boldsymbol{c}_2^{\boldsymbol{g}'}\ \cdots\ \boldsymbol{c}_m^{\boldsymbol{g}'})'$. The following proposition gives the asymptotic distributions of $\widehat{\boldsymbol{\gamma}}_m^{\boldsymbol{g}}$ and $\widehat{\boldsymbol{\rho}}_m^{\boldsymbol{g}}$.

Proposition 5.5 *Under the assumptions of Lemma 5.1, assume that* **g** *is continuously differentiable,* $E\|\boldsymbol{\eta}_t^{\boldsymbol{g}}\|^2 < \infty$ *and* $E\|\frac{\partial \boldsymbol{\eta}_t^{\boldsymbol{g}}}{\partial \boldsymbol{\theta}}\|^2 < \infty$*, and there exists a neighborhood* $\mathcal{V}(\boldsymbol{\theta}_0)$ *of* $\boldsymbol{\theta}_0$*, such that*

$$E \sup_{\boldsymbol{\theta}\in\mathcal{V}(\boldsymbol{\theta}_0)} \left|\frac{\partial \boldsymbol{\eta}_t^{\boldsymbol{g}}(\boldsymbol{\theta})}{\partial \theta_i}\right| < \infty, \quad E \sup_{\boldsymbol{\theta}\in\mathcal{V}(\boldsymbol{\theta}_0)} \left|\frac{\partial^2 \boldsymbol{\eta}_t^{\boldsymbol{g}}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right| < \infty, \quad E \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left|\frac{\partial^3 \boldsymbol{\eta}_t^{\boldsymbol{g}}(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j \partial \theta_k}\right| < \infty.$$

Then

$$\sqrt{n}\widehat{\boldsymbol{\gamma}}_m^{\boldsymbol{g}} \xrightarrow{\mathcal{L}} \mathcal{N}\left\{\boldsymbol{0}, \boldsymbol{\Sigma}_{\gamma_m}^{\boldsymbol{g}}\right\} \quad \text{and} \quad \sqrt{n}\widehat{\boldsymbol{\rho}}_m^{\boldsymbol{g}} \xrightarrow{\mathcal{L}} \mathcal{N}\left\{\boldsymbol{0}, \boldsymbol{\Sigma}_{\rho_m}^{\boldsymbol{g}}\right\} \tag{5.21}$$

as $n \to \infty$, *where*

$$\boldsymbol{\Sigma}^{\boldsymbol{g}}_{\boldsymbol{\gamma}_m} = \boldsymbol{I}_m \otimes \boldsymbol{\Omega}^{\otimes 2}_{\boldsymbol{g}} + \boldsymbol{C}^{\boldsymbol{g}}_m \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{C}^{\boldsymbol{g}'}_m + \boldsymbol{C}^{\boldsymbol{g}}_m \boldsymbol{\Sigma}_{\boldsymbol{\theta}\boldsymbol{\Upsilon}^g} + \boldsymbol{\Sigma}'_{\boldsymbol{\theta}\boldsymbol{\Upsilon}^g} \boldsymbol{C}^{\boldsymbol{g}'}_m,$$
$$\boldsymbol{\Sigma}_{\boldsymbol{\rho}_m} = \left\{\boldsymbol{I}_m \otimes \{\mathrm{diag}(\boldsymbol{\Omega}^{\otimes 2}_{\boldsymbol{g}})\}^{-1/2}\right\} \boldsymbol{\Sigma}^{\boldsymbol{g}}_{\boldsymbol{\gamma}_m} \left\{\boldsymbol{I}_m \otimes \{\mathrm{diag}(\boldsymbol{\Omega}^{\otimes 2}_{\boldsymbol{g}})\}^{-1/2}\right\}.$$

5.4 Illustrations

Proposition 5.4 shows that the Bahadur slope of the Li test is always greater than that of the BP test. It was not possible to obtain a general comparison of these two slopes with that of Katayama's test. We, therefore, begin with a very simple example for which the computation of the Bahadur slopes is tractable. We then give some numerical evaluations of the slopes of the 3 tests for an ARMA-GARCH model. We then study whether the finite sample empirical behaviors of the three tests reflect the theoretical ranking provided by the Bahadur slopes.

5.4.1 Computation of Bahadur's Slopes in a Particular Example

As in Example 5.2, assume that an AR(1) model $X_t = aX_{t-1} + \sigma\eta_t$ is fitted to a MA(1) model $X_t = \epsilon_t + b\epsilon_{t-1}$. Assume in addition that (ϵ_t) is a Gaussian noise of variance σ_ϵ^2. For $\theta = (a, \sigma^2)$, we have

$$\frac{\partial \tilde{\ell}_t(\theta)}{\partial \theta} = \begin{pmatrix} -2X_{t-1}\frac{X_t - aX_{t-1}}{\sigma^2} \\ -\frac{(X_t - aX_{t-1})^2}{\sigma^4} + \frac{1}{\sigma^2} \end{pmatrix}$$

and

$$\frac{\partial^2 \tilde{\ell}_t(\theta)}{\partial \theta \partial \theta'} = \begin{pmatrix} 2\frac{X_{t-1}^2}{\sigma^2} & 2\frac{X_{t-1}(X_t - aX_{t-1})}{\sigma^4} \\ 2\frac{X_{t-1}(X_t - aX_{t-1})}{\sigma^4} & 2\frac{(X_t - aX_{t-1})^2}{\sigma^6} - \frac{1}{\sigma^4} \end{pmatrix}.$$

By the ergodic theorem and tedious computations (see Sect. 5.6), it can be shown that, as $n \to \infty$, $\hat{\boldsymbol{I}}$ and $\hat{\boldsymbol{J}}$ jointly converge a.s. to

$$\boldsymbol{I} = \begin{pmatrix} 4\frac{(1+b^2)^2}{1+b^4+b^2} & 0 \\ 0 & 2\frac{(1+b^2)^2}{\sigma_\epsilon^4(1+b^4+b^2)^2} \end{pmatrix}, \quad \boldsymbol{J} = \boldsymbol{I}/2.$$

Similarly, $\widehat{\boldsymbol{c}}_\ell$ converges to $\boldsymbol{c}^*_\ell = (c^*_\ell, 0)$, with

$$c_1^* = -1, \qquad c_2^* = \frac{-b(1+b^2)}{1+b^2+b^4}$$

and $c^*_\ell = 0$ for $\ell \geq 3$. If follows that $\boldsymbol{C}_m$ and $\boldsymbol{P}^+$ have rank 1, with

$$\boldsymbol{P}^+ = \frac{1}{k}\begin{pmatrix} c_1^* \\ c_2^* \\ \mathbf{0}'_{m-2} \end{pmatrix}\begin{pmatrix} c_1^* & c_2^* & \mathbf{0}_{m-2} \end{pmatrix}, \qquad k = c_1^{*2} + c_2^{*2}.$$

Using the form of $\boldsymbol{\rho}^*_m = \left(\rho_1^*, \rho_2^*, 0, \dots, 0\right)'$ obtained in Example 5.2, it follows that

$$c_{K+} = \frac{(\rho_1^* c_1^* + \rho_2^* c_2^*)^2}{k} = \frac{b^{10}}{(1+b^2)^2\,(1+b^2+b^4)^2(1+3b^2+5b^4+3b^6+b^8)}.$$

We also have

$$\boldsymbol{I}_m + \boldsymbol{C}_m \boldsymbol{\Sigma}_{\boldsymbol{\theta}\boldsymbol{\Upsilon}} + \boldsymbol{\Sigma}'_{\boldsymbol{\theta}\boldsymbol{\Upsilon}} \boldsymbol{C}'_m + \boldsymbol{C}_m \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{C}'_m = \begin{pmatrix} \frac{b^2}{(1+b^2)^2} & \frac{-b}{1+b^2} & 0 \dots 0 \\ \frac{-b}{1+b^2} & \frac{1+b^4}{1+b^2+b^4} & 0 \dots 0 \\ 0 & 0 & 1 \dots 0 \\ \vdots & \vdots & \ddots \\ 0 & \dots & 0 \dots 1 \end{pmatrix}.$$

It is worth noting that this matrix is not positive definite. However, this is not surprising since, in this misspecified framework, the latter matrix cannot be interpreted as an asymptotic covariance matrix. Therefore, the plug-in estimator of $\boldsymbol{\Sigma}_{\boldsymbol{\rho}_m}$ obtained by using consistent estimators of the matrices involved in the r.h.s. of (5.13) converges to a matrix that is not semi-positive definite.

Let us now consider the consistent estimator in (5.14). Tedious computations (see Sect. 5.6) show that the limiting covariance matrix is

$$\begin{aligned} \boldsymbol{\Sigma}_{\boldsymbol{\rho}^*_m} &= E(\boldsymbol{\Upsilon}_t \boldsymbol{\Upsilon}'_t) + \boldsymbol{C}_m \boldsymbol{\Sigma}_{\boldsymbol{\theta}\boldsymbol{\Upsilon}} + \boldsymbol{\Sigma}'_{\boldsymbol{\theta}\boldsymbol{\Upsilon}} \boldsymbol{C}'_m + \boldsymbol{C}_m \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{C}'_m \\ &= \begin{pmatrix} \frac{b^2(1+2b^2+5b^4+2b^6+b^8)}{(1+b^2)^2(1+b^2+b^4)^2} & \frac{-b(1+b^2+4b^4+b^6+b^8)}{(1+b^2)(1+b^2+b^4)^2} & 0 \dots 0 \\ \frac{-b(1+b^2+4b^4+b^6+b^8)}{(1+b^2)(1+b^2+b^4)^2} & \frac{1+b^2+5b^4+b^6+b^8}{(1+b^2+b^4)^2} & 0 \dots 0 \\ 0 & 0 & 1 \dots 0 \\ \vdots & \vdots & \ddots \\ 0 & \dots & 0 \dots 1 \end{pmatrix}. \end{aligned}$$

It follows that the Bahadur slope of the Li test is

$$c_{Li} = \boldsymbol{\rho}^{*\prime}_m \boldsymbol{\Sigma}^{-1}_{\boldsymbol{\rho}^*_m} \boldsymbol{\rho}^*_m = \frac{b^4(1+b^2)^2}{1+3b^2+8b^4+11b^6+8b^8+3b^{10}+b^{12}}.$$

Moreover, the Bahadur slope of the BP test is

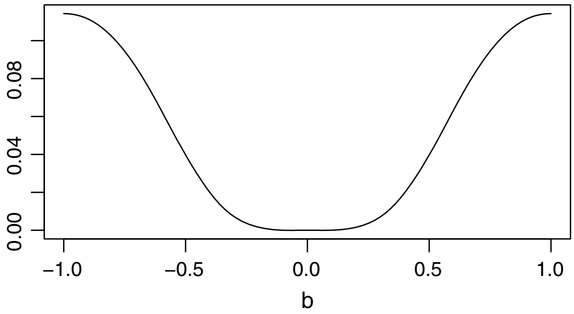


Fig. 5.1 Difference between Bahadur slopes ($c_{Li} - c_K$) as a function of the MA parameter

$$c_{BP} = \frac{\boldsymbol{\rho}_m^{*\prime}\boldsymbol{\rho}_m^*}{\lambda_1} = \frac{b^4(1+3b^2+b^4)}{\lambda_1(1+b^2)^2(1+b^2+b^4)^2}.$$

Numerical calculations (not reported here) show that, in this example, the difference between c_{Li} and c_{BP} is close to zero, though positive as shown in the general case. The difference in asymptotic efficiencies between the Li and Katayama tests is illustrated in Fig. 5.1. It can be seen that the difference is all the more important as the alternative is far from the null assumption (i.e., as $|b|$ approaches 1).

The conclusion on this example is that the ranking in terms of Bahadur slope is $Li \succ BP \succ K$. Numerical illustrations on more complex models displayed in Sect. 5.4 confirm this ranking.

5.4.2 Numerical Evaluation of Bahadur's Slopes

Consider the ARMA(1,1)-GARCH(1,1) model defined by

$$\begin{cases} x_t = ax_{t-1} + \epsilon_t - b\epsilon_{t-1} + c \\ \epsilon_t = \sigma_t\eta_t, \quad \sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \end{cases} \tag{5.22}$$

and first assume that the data generating process is the ARMA(2,1)-GARCH(1,1) model defined by

$$\begin{cases} x_t = a_1x_{t-1} + a_2x_{t-2} + \epsilon_t - b\epsilon_{t-1} + c \\ \epsilon_t = \sigma_t\eta_t, \quad \sigma_t^2 = \omega + \alpha\epsilon_{t-1}^2 + \beta\sigma_{t-1}^2, \end{cases} \tag{5.23}$$

with $\boldsymbol{\theta}_0' = (a_1, a_2, b, c, \omega, \alpha, \beta) = (1, -0.16, -0.4, 1, 1, 0.05, 0.93)$ and η_t following a standardized Student t-distribution with $\nu = 10$ degrees of freedom. Consider the Bahadur slopes of the portmanteau tests based on $m = 6$ residuals autocorrelations. For Katayama's portmanteau test, we used the version defined by (5.17). The matrices $\boldsymbol{I}$, $\boldsymbol{J}$, $\boldsymbol{C}_m$ and $\boldsymbol{\Sigma}_{\rho_m}$ and the vector $\boldsymbol{\rho}_m^*$ cannot be computed exactly. In the very simple model of Sect. 5.4.1, these quantities are obtained explicitly, but such

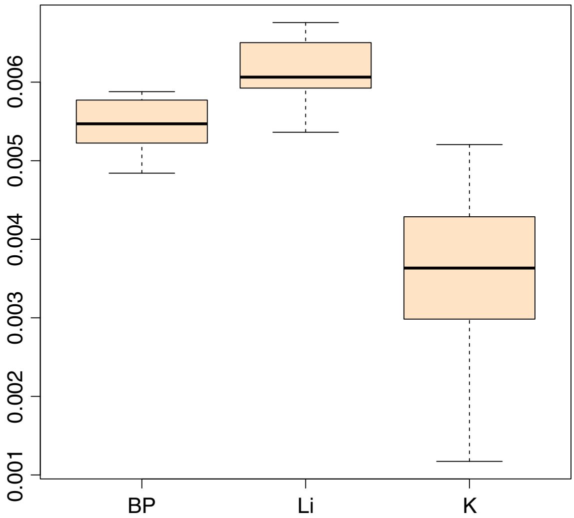


Fig. 5.2 Estimated Bahadur slopes of the three tests when the DGP is (5.23) and the model is (5.22)

computations are impossible in the presence of a GARCH component. We, therefore, evaluated the unknown mathematical expectations by corresponding means over very long simulations of the Model (5.22); we took 10 simulations of length $n = 200,000$. Numerical approximations of the slopes of Bahadur are deduced therefrom. Figure 5.2 displays the estimated Bahadur slopes. Note first that, even if the sample size $n = 200,000$ is quite large, the slopes are not estimated with great precision. Indeed, the estimates vary a lot from one replication to another. The ranking is, however, clear, i.e., $Li \succ BP \succ K$ in terms of the Bahadur efficiency. Similar results (we omit here) have been obtained for $\eta_t \sim \mathcal{N}(0,1)$ and for different values of m.

5.4.3 Monte Carlo Experiments

In a first set of Monte Carlo experiments, we simulated $N = 1,000$ independent replications of simulations of size $n = 1{,}000$ and $n = 4{,}000$ of the ARMA(1,1)-GARCH(1,1) process (5.22) where $\boldsymbol{\theta}_0' = (a, b, c, \omega, \alpha, \beta) = (0.8, -0.4, 1, 1, 0.05, 0.93)$ and η_t follows a Gaussian $\mathcal{N}(0,1)$ or a standardized Student t-distribution with $\nu = 10$ degrees of freedom. The upper part of Table 5.1 gives the empirical relative frequencies of rejection of Model (5.22) by the Box–Pierce, Li, and Katayama portmanteau tests, for several nominal levels, when η_t is Gaussian ($H_{0-\mathcal{N}}$) or when it has a Student t-distribution (H_{0-t}). For all the portmanteau tests, we took $m = 6$ residuals autocorrelations. Type I error is generally well controlled, for both distributions and sample sizes, with a slight disadvantage for the Li test which seems a bit too liberal when $n = 1000$ and $\eta_t \sim \mathcal{N}(0,1)$.

We then simulated alternative models to (5.22). We first considered an ARMA(2,1)-GARCH(1,1) process with the same volatility as (5.22) (with the two noise distributions), but with the level equation satisfying $x_t = a_1 x_{t-1} + a_2 x_{t-2} + \epsilon_t - b\epsilon_{t-1} + c$ with $(a_1, a_2, b, c) = (1, -0.16, -0.4, 1)$. The lines corresponding to $H_{1-\mathcal{N}}$ and H_{1-t} show that, for this alternative model, we have $Li \succ BP \succ K$ in terms of empirical power. Note that the ranking coincides with that of the Bahadur slopes, in particular, $Li \succ BP$ agrees with the corresponding theoretical result of Proposition 5.4.

We then considered the alternative nonlinear model $X_t = \prod_{\ell=0}^{k} |\eta_{t-\ell}|^j$ of Example 5.3: with $k = 4$ and $j = 3$ positive integers and $\eta_t \sim \mathcal{N}(0, 1)$. The lines of Table 5.1 corresponding to H_{1-NL} show that the ranking of the methods remains the same, i.e., $Li \succ BP \succ K$, with a particularly poor performance of the Katayama test.

Table 5.2 shows the same outputs as Table 5.1, but $m = 12$ instead of $m = 6$ residual autocorrelations. The results of the two tables are very similar and lead to the same empirical ranking of the tests.

5.5 Conclusion

Table 5.3 summarizes our comparison of the 3 methods for checking validity of the general nonlinear model (5.1)–(5.2). The BP test is ranked last in terms of simplicity since the computation of the critical value requires the use of an algorithm for computing the distribution of a quadratic form of normal distributions. This is not a big issue because the R package `CompQuadForm` developed by Lafaye de Micheaux (2017) is a handy tool that provides such algorithms. The Li test is more demanding in terms of assumptions since it requires the invertibility of the matrix $\boldsymbol{\Sigma}_{\rho_m}$. In contrast, this test has the highest Bahadur slope and is also empirically the most powerful in Monte Carlo experiments, but rejects a bit too often in finite samples. Overall, the BP test seems to be the best compromise since it controls well the first kind error and displays good power, while it is not complicated to implement thanks to the R package `CompQuadForm`.

5.6 Technical Details

We first state elementary derivative rules, which can be found in Lütkepohl ([28]; Appendix A.13).

Lemma 5.2 *If $f(\boldsymbol{A})$ is a scalar function of a matrix $\boldsymbol{A}$ whose elements a_{ij} are function of a variable x, then*

Table 5.1 Empirical relative frequency of rejection of the null of correct specification using the Box–Pierce (BP), Li, and Katayama (K) tests, for nominal levels varying from 0.1% to 20%. The number of residual autocorrelations is $m = 6$

DGP	n		0.1%	1%	2%	3%	4%	5%	6%	7%	10%	20%
$H_{0-\mathcal{N}}$	1000	BP	0.4	1.6	2.9	4.0	4.7	6.1	6.8	7.7	11.3	21.2
		Li	0.6	2.6	3.4	4.5	5.2	6.9	7.7	9.6	12.8	24.5
		K	0.3	1.9	2.6	3.2	4.3	5.2	6.6	7.8	10.6	19.9
	4000	BP	0.0	0.5	1.5	2.9	3.9	5.0	6.0	7.4	10.5	20.2
		Li	0.1	0.8	1.7	3.1	4.7	5.5	6.6	7.6	11.4	22.3
		K	0.0	0.4	0.9	2.2	3.5	4.5	5.9	7.0	9.7	20.5
H_{0-t}	1000	BP	0.0	0.5	1.5	2.7	4.6	5.3	6.9	7.7	10.6	20.8
		Li	0.6	1.3	2.3	3.6	4.4	5.2	6.3	7.1	10.8	22.4
		K	0.0	0.7	1.5	2.6	3.2	4.7	5.9	6.7	8.7	20.0
	4000	BP	0.1	0.8	1.8	2.6	3.5	4.8	6.2	7.6	10.2	21.1
		Li	0.2	0.6	1.5	2.5	3.7	5.2	6.4	7.1	10.0	20.1
		K	0.1	0.9	2.1	2.7	4.2	5.6	6.8	8.2	10.5	19.5
$H_{1-\mathcal{N}}$	1000	BP	60.5	77.1	81.5	84.7	86.5	88.5	89.5	90.6	92.8	97.2
		Li	77.0	86.4	90.0	91.1	91.9	92.5	93.7	94.2	95.6	98.2
		K	21.9	42.7	50.5	56.4	59.8	63.4	66.2	69.1	72.8	83.1
	4000	BP	99.4	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		Li	99.7	99.8	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		K	91.7	94.1	95.1	95.8	95.8	96.0	96.4	96.6	97.0	98.2
H_{1-t}	1000	BP	62.1	82.0	86.9	89.2	90.8	92.2	92.9	93.6	95.8	97.7
		Li	64.0	82.8	87.4	89.9	91.6	93.0	93.8	94.4	95.9	97.4
		K	41.3	64.7	72.4	76.5	79.2	81.2	83.4	85.3	88.9	93.1
	4000	BP	99.7	99.8	99.9	99.9	99.9	99.9	99.9	99.9	99.9	99.9
		Li	99.9	99.9	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
		K	74.6	89.7	92.4	94.1	94.8	96.2	96.7	97.3	98.2	99.7
H_{1-NL}	1000	BP	9.0	21.2	26.8	30.3	32.9	35.2	37.0	39.4	44.3	56.8
		Li	26.4	39.4	44.5	48.3	50.8	53.6	55.6	56.9	62.2	71.6
		K	4.0	6.5	7.9	8.9	9.5	10.4	11.3	12.2	13.6	17.9
	4000	BP	32.2	45.2	51.2	54.1	56.7	58.6	60.3	62.2	66.2	75.0
		Li	66.0	76.3	80.3	82.3	84.0	84.9	85.7	86.2	88.1	92.7
		K	9.9	14.1	15.5	16.9	17.9	19.0	19.9	21.0	22.8	28.1

$$\frac{\partial f(\boldsymbol{A})}{\partial x} = \sum_{i,j} \frac{\partial f(\boldsymbol{A})}{\partial a_{ij}} \frac{\partial a_{ij}}{\partial x} = \mathrm{Tr}\left\{ \frac{\partial f(\boldsymbol{A})}{\partial \boldsymbol{A}'} \frac{\partial \boldsymbol{A}}{\partial x} \right\}.$$

When $\boldsymbol{A}$ *is invertible, then*

$$\frac{\partial \log|\det(\boldsymbol{A})|}{\partial \boldsymbol{A}'} = \boldsymbol{A}^{-1}, \quad \frac{\partial \mathrm{Tr}(\boldsymbol{C}\boldsymbol{A}^{-1}\boldsymbol{B})}{\partial \boldsymbol{A}'} = -\boldsymbol{A}^{-1}\boldsymbol{B}\boldsymbol{C}\boldsymbol{A}^{-1}, \quad \frac{\partial \mathrm{Tr}(\boldsymbol{C}\boldsymbol{A}\boldsymbol{B})}{\partial \boldsymbol{A}'} = \boldsymbol{B}\boldsymbol{C}.$$

Table 5.2 Empirical relative frequency of rejection of the null of correct specification using the Box–Pierce (BP), Li, and Katayama (K) tests, for nominal levels varying from 0.1% to 20%. The number of residual autocorrelations is $m = 12$

DGP	n		0.1%	1%	2%	3%	4%	5%	6%	7%	10%	20%
H_{0-N}	1000	BP	0.2	1.4	2.8	4.5	5.3	6.2	7.4	8.6	12.5	21.0
		Li	0.5	2.3	3.0	4.9	5.8	7.3	8.7	10.0	13.5	23.9
		K	0.2	1.8	2.7	4.1	4.9	5.7	6.8	7.8	10.7	20.6
	4000	BP	0.2	0.6	1.6	2.9	3.9	5.1	5.7	6.7	10.1	22.2
		Li	0.1	1.1	2.3	3.1	3.7	4.9	6.4	7.2	11.5	23.2
		K	0.1	0.7	1.9	2.7	3.4	4.7	6.0	6.7	9.7	20.9
H_{0-t}	1000	BP	0.0	0.6	1.3	2.4	4.0	5.2	6.1	7.4	11.0	23.3
		Li	0.6	1.7	2.5	3.7	5.2	5.8	7.1	7.8	11.3	23.5
		K	0.0	0.4	1.2	2.7	3.6	4.5	6.1	7.3	10.2	21.4
	4000	BP	0.0	0.7	2.1	3.4	4.4	5.3	6.1	6.6	10.3	20.7
		Li	0.0	1.4	2.3	3.2	4.4	5.2	6.1	7.0	9.9	20.6
		K	0.0	1.1	2.0	3.0	3.9	4.5	6.0	7.0	11.2	20.9
H_{1-N}	1000	BP	5.4	18.2	25.6	31.3	35.3	37.6	39.9	43.7	51.7	68.2
		Li	4.4	14.2	22.9	29.5	33.1	36.4	39.9	43.1	49.2	65.9
		K	3.8	15.8	21.9	24.9	27.9	31.0	34.4	37.7	43.6	59.4
	4000	BP	70.3	89.2	93.3	95.3	96.1	96.9	97.2	97.7	98.4	99.3
		Li	64.4	87.1	91.2	93.3	94.5	95.1	96.7	97.2	98.0	99.1
		K	57.4	79.6	84.5	86.6	88.3	89.3	90.4	91.6	93.6	97.4
H_{1-t}	1000	BP	4.4	18.5	25.4	30.1	33.9	36.5	39.8	42.2	48.4	64.0
		Li	3.4	15.1	22.5	27.1	30.6	35.0	38.5	41.6	48.0	63.6
		K	3.9	13.9	20.0	24.3	26.9	29.9	32.7	35.4	41.4	56.2
	4000	BP	69.6	87.3	91.9	93.8	94.9	95.5	96.2	96.5	97.7	98.8
		Li	64.1	84.7	89.9	91.9	93.5	94.2	95.3	95.8	97.1	98.9
		K	54.2	76.2	82.3	85.5	87.5	89.2	90.1	91.8	94.0	96.7
H_{1-NL}	1000	BP	7.8	18.5	24.2	28.0	30.5	34.0	35.7	37.7	43.2	57.4
		Li	27.6	37.4	41.3	44.9	47.0	49.3	51.0	52.3	54.7	64.5
		K	4.4	6.8	7.7	8.4	9.0	9.3	10.0	10.6	12.3	15.6
	4000	BP	7.8	13.1	17.5	19.9	22.1	24.3	25.5	26.9	31.9	44.5
		Li	39.7	47.5	50.1	51.6	53.3	54.6	56.8	57.4	59.7	67.5
		K	7.1	9.6	10.2	11.1	11.9	12.4	12.8	13.0	14.1	16.4

Table 5.3 Ranking of the three methods in terms of simplicity of implementation, mathematical assumptions, asymptotic Bahadur's efficiency, and finite sample empirical efficiency

	Simplicity	Assumptions	Bahadur's slope	Empirical efficiency
BP	3	1	2	1
Li	1	3	1	2
K	2	2	3	3

***Proof of Lemma* 5.1** We have $\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left|\ell_t - \widetilde{\ell}_t\right| \le \sum_{i=1}^4 a_{ti}$, with

$$a_{t1} = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left|\left\{\boldsymbol{M}_t - \widetilde{\boldsymbol{M}}_t\right\}' \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\epsilon}_t\right|, \qquad a_{t2} = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left|\widetilde{\boldsymbol{\epsilon}}_t' \left(\boldsymbol{\Sigma}_t^{-1} - \widetilde{\boldsymbol{\Sigma}}_t^{-1}\right)\boldsymbol{\epsilon}_t\right|,$$
$$a_{t3} = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left|\widetilde{\boldsymbol{\epsilon}}_t'\widetilde{\boldsymbol{\Sigma}}_t^{-1}\left\{\widetilde{\boldsymbol{M}}_t - \boldsymbol{M}_t\right\}\right|, \qquad a_{t4} = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left|\log|\boldsymbol{\Sigma}_t| - \log|\widetilde{\boldsymbol{\Sigma}}_t|\right|.$$

By Assumption 5.1(iii) and (iv), we have

$$\begin{aligned} a_{t2} &= \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left|\widetilde{\boldsymbol{\epsilon}}_t'\boldsymbol{\Sigma}_t^{-1}\left(\widetilde{\boldsymbol{\Sigma}}_t - \boldsymbol{\Sigma}_t\right)\widetilde{\boldsymbol{\Sigma}}_t^{-1}\boldsymbol{\epsilon}_t\right| = \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left|\operatorname{Tr}\left\{\boldsymbol{\Sigma}_t^{-1}\left(\widetilde{\boldsymbol{\Sigma}}_t - \boldsymbol{\Sigma}_t\right)\widetilde{\boldsymbol{\Sigma}}_t^{-1}\boldsymbol{\epsilon}_t\widetilde{\boldsymbol{\epsilon}}_t'\right\}\right| \\ &\le K \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\|\boldsymbol{\Sigma}_t^{-1}\right\| \left\|\widetilde{\boldsymbol{\Sigma}}_t - \boldsymbol{\Sigma}_t\right\| \left\|\widetilde{\boldsymbol{\Sigma}}_t^{-1}\right\| \left\|\boldsymbol{\epsilon}_t\widetilde{\boldsymbol{\epsilon}}_t'\right\| \le K\rho^t \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\|\boldsymbol{\epsilon}_t\widetilde{\boldsymbol{\epsilon}}_t'\right\| \text{ a.s.} \end{aligned}$$

as $n \to \infty$. Note that $\sum_{t=1}^\infty a_{t2}$ is finite a.s. since, for some $r < 1$,

$$\begin{aligned} &E\left(\sum_{t=1}^\infty \rho^t \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\|\boldsymbol{\epsilon}_t\widetilde{\boldsymbol{\epsilon}}_t'\right\|\right)^r \\ &\le K\sum_{t=1}^\infty \rho^{rt}\left\{E\left\|\boldsymbol{X}_t\right\|^{2r} + E\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\|\boldsymbol{M}_t\right\|^{2r} + \rho^{rt}\left(E\left\|\boldsymbol{X}_t\right\|^r + E\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left\|\boldsymbol{M}_t\right\|^r\right)\right\} < \infty \end{aligned}$$

by Assumption 5.1(iv). Similarly, it can be shown that $\sum_{t=1}^\infty a_{ti} < \infty$ a.s. for all $i \in \{1, \ldots, 4\}$. It follows that

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left|\widetilde{Q}_n(\boldsymbol{\theta}) - Q_n(\boldsymbol{\theta})\right| \to 0 \quad \text{a.s. as } n \to \infty. \tag{5.24}$$

By Assumption 5.1(iii), $E\ell_1^-(\boldsymbol{\theta}) \le \log^- \det \boldsymbol{\Sigma}_1(\boldsymbol{\theta}) < \infty$. Now, Assumption 5.1(i) and (ii) and the ergodic theorem (see Billingsley [6], pp. 284 and 495) show that $Q_n(\boldsymbol{\theta}) \to E\ell_1(\boldsymbol{\theta}) \in \mathbb{R} \cup \{+\infty\}$ a.s. Note also that $\ell_1(\boldsymbol{\theta}_0) = \boldsymbol{\eta}_1'\boldsymbol{\eta}_1 + \log\det\boldsymbol{\Sigma}_1(\boldsymbol{\theta}_0)$ admits a finite expectation, using $E\boldsymbol{\eta}_1'\boldsymbol{\eta}_1 = d$ and Assumption 5.1(iii). We thus have shown that the criterion $\widetilde{Q}_n(\boldsymbol{\theta})$ tends a.s. to the limit criterion $E\ell_1(\boldsymbol{\theta})$ which is defined a priori in $\mathbb{R} \cup \{+\infty\}$, and is valued in $\mathbb{R}$ at $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. The limit criterion is minimized at the true value since

$$\begin{aligned} E\ell_1(\boldsymbol{\theta}) - E\ell_1(\boldsymbol{\theta}_0) &= E\boldsymbol{\epsilon}_t'(\boldsymbol{\theta})\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\boldsymbol{\epsilon}_t(\boldsymbol{\theta}) + E\log\frac{\det\boldsymbol{\Sigma}_t(\boldsymbol{\theta})}{\det\boldsymbol{\Sigma}_t(\boldsymbol{\theta}_0)} - d \\ &= E\left\{\boldsymbol{M}_t(\boldsymbol{\theta}) - \boldsymbol{M}_t(\boldsymbol{\theta}_0)\right\}'\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\left\{\boldsymbol{M}_t(\boldsymbol{\theta}) - \boldsymbol{M}_t(\boldsymbol{\theta}_0)\right\} \\ &+ E\boldsymbol{\eta}_t'\boldsymbol{S}_t(\boldsymbol{\theta}_0)\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\boldsymbol{S}_t(\boldsymbol{\theta}_0)\boldsymbol{\eta}_t + E\log\frac{\det\boldsymbol{\Sigma}_t(\boldsymbol{\theta})}{\det\boldsymbol{\Sigma}_t(\boldsymbol{\theta}_0)} - d \\ &= E\left\{\boldsymbol{M}_t(\boldsymbol{\theta}) - \boldsymbol{M}_t(\boldsymbol{\theta}_0)\right\}'\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\left\{\boldsymbol{M}_t(\boldsymbol{\theta}) - \boldsymbol{M}_t(\boldsymbol{\theta}_0)\right\} \\ &+ E\operatorname{Tr}\left(\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\boldsymbol{\Sigma}_t(\boldsymbol{\theta}_0) - \mathbf{Id}_d\right) + E\log\det\boldsymbol{\Sigma}_t(\boldsymbol{\theta})\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta}_0) \ge 0 \end{aligned}$$

with equality if and only if $\boldsymbol{M}_t(\boldsymbol{\theta}) = \boldsymbol{M}_t(\boldsymbol{\theta}_0)$ and $\boldsymbol{\Sigma}_t(\boldsymbol{\theta}) = \boldsymbol{\Sigma}_t(\boldsymbol{\theta}_0)$. For the last inequality, we used the elementary inequality $\mathrm{Tr}(\mathbf{A}^{-1}\mathbf{B}) - \log\det(\mathbf{A}^{-1}\mathbf{B}) \geq \mathrm{Tr}(\mathbf{A}^{-1}\mathbf{A}) - \log\det(\mathbf{A}^{-1}\mathbf{A}) = d$ for all symmetric positive definite matrices of order $d \times d$. In view of Assumption 5.1(v), it follows that $E\ell_1(\boldsymbol{\theta}) > E\ell_1(\boldsymbol{\theta}_0)$ when $\boldsymbol{\theta} \neq \boldsymbol{\theta}_0$. To show the consistency, the convergence of the criterion needs to be established uniformly in some neighborhood of any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ (see, e.g., Amendola and Francq ([2]; Section 5.1).

Let $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_0$ and let $\boldsymbol{V}_d(\boldsymbol{\theta}_1)$ be the open sphere with center $\boldsymbol{\theta}_1$ and radius $1/d$. The process $\left\{\inf_{\boldsymbol{\theta}\in\boldsymbol{V}_d(\boldsymbol{\theta}_1)\cap\boldsymbol{\Theta}} \ell_t(\boldsymbol{\theta})\right\}_t$ is stationary and ergodic. We thus have

$$\inf_{\boldsymbol{\theta}\in\boldsymbol{V}_d(\boldsymbol{\theta}_1)\cap\boldsymbol{\Theta}} Q_n(\boldsymbol{\theta}) \geq \frac{1}{n}\sum_{i=1}^{n} \inf_{\boldsymbol{\theta}\in\boldsymbol{V}_d(\boldsymbol{\theta}_1)\cap\boldsymbol{\Theta}} \ell_t(\boldsymbol{\theta}) \to E \inf_{\boldsymbol{\theta}\in\boldsymbol{V}_d(\boldsymbol{\theta}_1)\cap\boldsymbol{\Theta}} \ell_t(\boldsymbol{\theta}) \text{ a.s.}$$

In view of the continuity of $\ell_t(\cdot)$, the sequence $\inf_{\boldsymbol{\theta}\in\boldsymbol{V}_d(\boldsymbol{\theta}_1)\cap\boldsymbol{\Theta}} \ell_t(\boldsymbol{\theta})$ increases to $\ell_t(\boldsymbol{\theta}_1)$ when $d \to \infty$. By the Beppo–Levi theorem

$$\lim_{d\to\infty} \uparrow E \inf_{\boldsymbol{\theta}\in\boldsymbol{V}_d(\boldsymbol{\theta}_1)\cap\boldsymbol{\Theta}} \ell_t(\boldsymbol{\theta}) = E \lim_{d\to\infty} \uparrow \inf_{\boldsymbol{\theta}\in\boldsymbol{V}_d(\boldsymbol{\theta}_1)\cap\boldsymbol{\Theta}} \ell_t(\boldsymbol{\theta}) = E\ell_t(\boldsymbol{\theta}_1) > Q_\infty(\boldsymbol{\theta}_0).$$

In view of (5.24), it follows that, for all $\boldsymbol{\theta}_i \neq \boldsymbol{\theta}_0$, there exists a neighborhood $\mathcal{V}(\boldsymbol{\theta}_i)$, such that

$$\liminf_{n\to\infty} \inf_{\boldsymbol{\theta}\in\mathcal{V}(\boldsymbol{\theta}_i)\cap\boldsymbol{\Theta}} \widetilde{Q}_n(\boldsymbol{\theta}) > \lim_{n\to\infty} \widetilde{Q}_n(\boldsymbol{\theta}_0). \tag{5.25}$$

The rest of the proof of the consistency is standard and relies on the arguments of Wald [38]. The compact set $\boldsymbol{\Theta}$ is covered by a finite number of open sets $\mathcal{V}(\boldsymbol{\theta}_1), \ldots, \mathcal{V}(\boldsymbol{\theta}_m)$, and $\mathcal{V}(\boldsymbol{\theta}_0)$, where $\mathcal{V}(\boldsymbol{\theta}_0)$ is any neighborhood of $\boldsymbol{\theta}_0$, and the neighborhoods $\mathcal{V}(\boldsymbol{\theta}_i)$ $i = 1, \ldots, m$ satisfy (5.25). Then, with probability 1, we have

$$\inf_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \widetilde{Q}_n(\boldsymbol{\theta}) = \min_{i=0,1,\ldots,m} \inf_{\boldsymbol{\theta}\in\mathcal{V}(\boldsymbol{\theta}_i)} \widetilde{Q}_n(\boldsymbol{\theta}) = \inf_{\boldsymbol{\theta}\in\mathcal{V}(\boldsymbol{\theta}_0)} \widetilde{Q}_n(\boldsymbol{\theta})$$

for n large enough. Since $\mathcal{V}(\boldsymbol{\theta}_0)$ can be an arbitrarily small neighborhood of $\boldsymbol{\theta}_0$, the consistency follows.

We now turn to the proof of the asymptotic normality. The strong consistency and Assumption 5.1(vi) entail that we have a.s. $\partial\widetilde{Q}_n(\widehat{\boldsymbol{\theta}})/\partial\boldsymbol{\theta} = \mathbf{0}$ for sufficiently large n. In view of (5.3), we thus have

$$\mathbf{0} \overset{o(1)}{=} \sqrt{n}\frac{\partial Q_n(\widehat{\boldsymbol{\theta}})}{\partial\boldsymbol{\theta}} \quad \text{a.s.}$$

Now, Taylor expansions of the functions $\partial Q_n(\cdot)/\partial\boldsymbol{\theta}_i$, for $i = 1, \ldots, s$, give

$$\mathbf{0} \overset{o(1)}{=} \sqrt{n}\frac{\partial Q_n(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}} + \left[\frac{\partial^2 Q_n(\boldsymbol{\theta}_i^*)}{\partial\theta_i\partial\theta_j}\right]\sqrt{n}\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \text{ a.s.} \tag{5.26}$$

for some $\boldsymbol{\theta}_i^*$'s between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$.

By Lemma 5.2, we have

$$\frac{\partial}{\partial\theta_i}\ell_t(\boldsymbol{\theta}) = \mathrm{Tr}\left\{\left(\boldsymbol{\Sigma}_t^{-1} - \boldsymbol{\Sigma}_t^{-1}\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t'\boldsymbol{\Sigma}_t^{-1}\right)\frac{\partial\boldsymbol{\Sigma}_t}{\partial\theta_i}\right\} - 2\frac{\partial\boldsymbol{M}_t'}{\partial\theta_i}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\epsilon}_t.$$

Thus

$$\begin{aligned} Z_{it} &:= \frac{\partial}{\partial\theta_i}\ell_t(\boldsymbol{\theta}_0) \\ &= \mathrm{Tr}\left\{(\boldsymbol{I}_m - \boldsymbol{\eta}_t\boldsymbol{\eta}_t')\boldsymbol{S}_t^{-1}(\boldsymbol{\theta}_0)\frac{\partial\boldsymbol{\Sigma}_t(\boldsymbol{\theta}_0)}{\partial\theta_i}\boldsymbol{S}_t^{-1}(\boldsymbol{\theta}_0)\right\} - 2\frac{\partial\boldsymbol{M}_t'}{\partial\theta_i}\boldsymbol{S}_t^{-1}(\boldsymbol{\theta}_0)\boldsymbol{\eta}_t. \end{aligned}$$

We also have

$$\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell_t(\boldsymbol{\theta}) = \sum_{i=1}^{8} c_i,$$

with

$$\begin{aligned} c_1 &= \boldsymbol{\epsilon}_t'\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\frac{\partial\boldsymbol{\Sigma}_t(\boldsymbol{\theta})}{\partial\theta_i}\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\frac{\partial\boldsymbol{\Sigma}_t(\boldsymbol{\theta})}{\partial\theta_j}\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\boldsymbol{\epsilon}_t, \quad c_3 = -\boldsymbol{\epsilon}_t'\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\frac{\partial^2\boldsymbol{\Sigma}_t(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\boldsymbol{\epsilon}_t, \\ c_2 &= \boldsymbol{\epsilon}_t'\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\frac{\partial\boldsymbol{\Sigma}_t(\boldsymbol{\theta})}{\partial\theta_j}\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\frac{\partial\boldsymbol{\Sigma}_t(\boldsymbol{\theta})}{\partial\theta_i}\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\boldsymbol{\epsilon}_t, \quad c_5 = \mathrm{Tr}\left(\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\frac{\partial^2\boldsymbol{\Sigma}_t(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right), \\ c_4 &= -\,\mathrm{Tr}\left(\frac{\partial\boldsymbol{\Sigma}_t(\boldsymbol{\theta})}{\partial\theta_i}\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\frac{\partial\boldsymbol{\Sigma}_t(\boldsymbol{\theta})}{\partial\theta_j}\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\right), \quad c_6 = 2\boldsymbol{\epsilon}_t'\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\frac{\partial\boldsymbol{\Sigma}_t(\boldsymbol{\theta})}{\partial\theta_i}\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\frac{\partial\boldsymbol{M}_t(\boldsymbol{\theta})}{\partial\theta_j}, \\ c_7 &= 2\frac{\partial\boldsymbol{M}_t'}{\partial\theta_i}\boldsymbol{\Sigma}_t^{-1}\frac{\partial\boldsymbol{M}_t}{\partial\theta_j}, \quad c_8 = -2\frac{\partial^2\boldsymbol{M}_t'}{\partial\theta_i\partial\theta_j}\boldsymbol{\Sigma}_t^{-1}\boldsymbol{\epsilon}_t. \end{aligned}$$

The generic element of $\boldsymbol{J}$, whose existence is shown by (5.27) below, is thus

$$\begin{aligned} &E\left\{\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell_t(\boldsymbol{\theta}_0)\right\} \\ &= \mathrm{Tr}\left\{\frac{\partial\boldsymbol{\Sigma}_t(\boldsymbol{\theta}_0)}{\partial\theta_i}\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta}_0)\frac{\partial\boldsymbol{\Sigma}_t(\boldsymbol{\theta}_0)}{\partial\theta_j}\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta}_0)\right\} + 2\frac{\partial\boldsymbol{M}_t'}{\partial\theta_i}\boldsymbol{\Sigma}_t^{-1}\frac{\partial\boldsymbol{M}_t}{\partial\theta_j} \\ &= \mathrm{vec}\left(\frac{\partial\boldsymbol{\Sigma}_t(\boldsymbol{\theta})}{\partial\theta_i}\right)'\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\otimes\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\theta})\mathrm{vec}\left(\frac{\partial\boldsymbol{\Sigma}_t(\boldsymbol{\theta})}{\partial\theta_i}\right) + 2\frac{\partial\boldsymbol{M}_t'}{\partial\theta_i}\boldsymbol{\Sigma}_t^{-1}\frac{\partial\boldsymbol{M}_t}{\partial\theta_j}. \end{aligned}$$

To show that the matrix into brackets in (5.26) converges a.s. to $\boldsymbol{J}$ it suffices to use the ergodic theorem, the continuity of the derivatives, and to show that

$$E\sup_{\boldsymbol{\theta}\in\mathcal{V}(\boldsymbol{\theta}_0)}\left|\frac{\partial^2\ell_t(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right| < \infty \tag{5.27}$$

for some neighborhood $\mathcal{V}(\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}_0$. Note that

$$E \sup_{\theta \in \mathcal{V}(\theta_0)} |c_1| \leq E \sup_{\theta \in \mathcal{V}(\theta_0)} \left\| S_t^{-1}(\theta) S_t(\theta_0) \right\|^2 \sup_{\theta \in \mathcal{V}(\theta_0)} \max_{i \in \{1,\ldots,s\}} \left\| S_t^{-1}(\theta) \frac{\partial \Sigma_t(\theta)}{\partial \theta_i} S_t^{-1}(\theta) \right\|^2,$$

which is finite by (5.5), (5.6) and Hölder's inequality. Silmilary, it can be shown that $E \sup_{\theta \in \mathcal{V}(\theta_0)} |c_i| < \infty$ for all $i \in \{1, \ldots, 8\}$, which entails (5.27). By (5.2) and Assumption 5.1(viii), the sequence (Z_t) is a square integrable stationary martingale difference. By the central limit theorem of Billingsley [5]

$$\sqrt{n} \frac{\partial Q_n(\theta_0)}{\partial \theta} = \frac{1}{\sqrt{n}} \sum_{t=1}^{n} Z_t \xrightarrow{\mathcal{L}} \mathcal{N}(\mathbf{0}, I) \quad \text{as } n \to \infty.$$

The conclusion follows. □

***Proof of Proposition* 5.1** Note that $\widehat{\eta}_t = \widetilde{\eta}_t(\widehat{\theta})$ and $\eta_t = \eta_t(\theta_0)$, where

$$\widetilde{\eta}_t(\theta) = \widetilde{S}_t^{-1}(\theta) \widetilde{\epsilon}_t(\theta) \quad \text{and} \quad \eta_t(\theta) = S_t^{-1}(\theta) \epsilon_t(\theta).$$

Denote by $\widetilde{\eta}_{it}(\theta)$ (resp. $\eta_{it}(\theta)$) the i-th component of $\widetilde{\eta}_t(\theta)$ (resp. $\eta_t(\theta)$).

Lemma 5.3 *Under the assumptions of Lemma 5.1, for all* $i, j \in \{1, \ldots, s\}$ *and* $\ell > 0$,

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{t=\ell+1}^{n} \widetilde{\eta}_{it}(\theta) \widetilde{\eta}_{j,t-\ell}(\theta) - \frac{1}{n} \sum_{t=\ell+1}^{n} \eta_{it}(\theta) \eta_{j,t-\ell}(\theta) \right| = O\left(\frac{1}{n}\right) \quad a.s.$$

Proof We have

$$\begin{aligned} &\sup_{\theta \in \Theta} \left| \widetilde{\eta}_{it}(\theta) \widetilde{\eta}_{j,t-\ell}(\theta) - \eta_{it}(\theta) \eta_{j,t-\ell}(\theta) \right| \\ &\leq \sup_{\theta \in \Theta} |\widetilde{\eta}_{it}(\theta) - \eta_{it}(\theta)| \sup_{\theta \in \Theta} \left| \widetilde{\eta}_{j,t-\ell}(\theta) \right| + \sup_{\theta \in \Theta} \left| \widetilde{\eta}_{j,t-\ell}(\theta) - \eta_{j,t-\ell}(\theta) \right| \sup_{\theta \in \Theta} |\eta_{it}(\theta)| . \end{aligned}$$

Moreover, since a.s.,

$$\begin{aligned} &\sup_{\theta \in \Theta} \left\| \widetilde{\eta}_t(\theta) - \eta_t(\theta) \right\| \\ &\leq \sup_{\theta \in \Theta} \left\| \widetilde{S}_t^{-1}(\theta) - S_t^{-1}(\theta) \right\| \sup_{\theta \in \Theta} \|\widetilde{\epsilon}_t(\theta)\| + \sup_{\theta \in \Theta} \left\| \widetilde{M}_t(\theta) - M_t(\theta) \right\| \sup_{\theta \in \Theta} \left\| \widetilde{S}_t^{-1}(\theta) \right\| \\ &\leq \sup_{\theta \in \Theta} \left\| \widetilde{S}_t(\theta) - S_t(\theta) \right\| \sup_{\theta \in \Theta} \left\| \widetilde{S}_t^{-1}(\theta) \right\| \sup_{\theta \in \Theta} \left\| S_t^{-1}(\theta) \right\| \sup_{\theta \in \Theta} \|\widetilde{\epsilon}_t(\theta)\| + K\rho^t \\ &\leq K\rho^t (\sup_{\theta \in \Theta} \|M_t(\theta)\| + \|X_t\| + 1), \end{aligned}$$

we also have, a.s.

$$\sup_{\theta \in \Theta} \left\| \eta_t(\theta) \right\| \leq \sup_{\theta \in \Theta} \left\| S_t^{-1}(\theta) \right\| \sup_{\theta \in \Theta} \|\epsilon_t(\theta)\| \leq K (\sup_{\theta \in \Theta} \|M_t(\theta)\| + \|X_t\|).$$

It follows from Assumption 5.1(iv) that

$$E\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\widetilde{\eta}_{it}(\boldsymbol{\theta})\widetilde{\eta}_{j,t-\ell}(\boldsymbol{\theta})-\eta_{it}(\boldsymbol{\theta})\eta_{j,t-\ell}(\boldsymbol{\theta})\right|\right)^{s}<K\rho^{t}$$

for some $s>0$. Thus, for $s\in(0,1)$,

$$E\left(\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\sum_{t=1}^{\infty}\left|\widetilde{\eta}_{it}(\boldsymbol{\theta})\widetilde{\eta}_{j,t-\ell}(\boldsymbol{\theta})-\eta_{it}(\boldsymbol{\theta})\eta_{j,t-\ell}(\boldsymbol{\theta})\right|\right)^{s}<\infty,$$

which entails that the supremum inside the parentheses is a.s. finite. The conclusion follows. □

Lemma 5.4 *Under the assumptions of Lemma 5.1, for all* $i,j,k\in\{1,\dots,s\}$*, there exists a neighborhood* $\mathcal{V}(\boldsymbol{\theta}_0)$ *of* $\boldsymbol{\theta}_0$ *such that*

$$E\sup_{\boldsymbol{\theta}\in\mathcal{V}(\boldsymbol{\theta}_0)}\left|\frac{\partial\boldsymbol{\eta}_t(\boldsymbol{\theta})}{\partial\theta_i}\right|<\infty,\quad E\sup_{\boldsymbol{\theta}\in\mathcal{V}(\boldsymbol{\theta}_0)}\left|\frac{\partial^2\boldsymbol{\eta}_t(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j}\right|<\infty,\quad E\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}}\left|\frac{\partial^3\boldsymbol{\eta}_t(\boldsymbol{\theta})}{\partial\theta_i\partial\theta_j\partial\theta_k}\right|<\infty.$$

Proof The derivation rules of Lemma 5.2 give

$$\begin{aligned}\frac{\partial\boldsymbol{\eta}_t}{\partial\theta_i}&=-\boldsymbol{S}_t^{-1}\frac{\partial\boldsymbol{S}_t}{\partial\theta_i}\boldsymbol{S}_t^{-1}\boldsymbol{\epsilon}_t-\boldsymbol{S}_t^{-1}\frac{\partial\boldsymbol{M}_t}{\partial\theta_i},\\ \frac{\partial^2\boldsymbol{\eta}_t}{\partial\theta_i\partial\theta_j}&=\boldsymbol{S}_t^{-1}\left\{\frac{\partial\boldsymbol{S}_t}{\partial\theta_j}\boldsymbol{S}_t^{-1}\frac{\partial\boldsymbol{S}_t}{\partial\theta_i}+\frac{\partial\boldsymbol{S}_t}{\partial\theta_i}\boldsymbol{S}_t^{-1}\frac{\partial\boldsymbol{S}_t}{\partial\theta_j}\right\}\boldsymbol{S}_t^{-1}\boldsymbol{\epsilon}_t\\ &\quad-\boldsymbol{S}_t^{-1}\frac{\partial^2\boldsymbol{S}_t}{\partial\theta_i\partial\theta_j}\boldsymbol{S}_t^{-1}\boldsymbol{\epsilon}_t+\boldsymbol{S}_t^{-1}\frac{\partial\boldsymbol{S}_t}{\partial\theta_i}\boldsymbol{S}_t^{-1}\frac{\partial\boldsymbol{M}_t}{\partial\theta_j}\\ &\quad+\boldsymbol{S}_t^{-1}\frac{\partial\boldsymbol{S}_t}{\partial\theta_j}\boldsymbol{S}_t^{-1}\frac{\partial\boldsymbol{M}_t}{\partial\theta_i}-\boldsymbol{S}_t^{-1}\frac{\partial^2\boldsymbol{M}_t}{\partial\theta_i\partial\theta_j}.\end{aligned}\tag{5.28}$$

Noting in particular that

$$\boldsymbol{S}_t^{-1}(\boldsymbol{\theta})\frac{\partial\boldsymbol{\Sigma}_t(\boldsymbol{\theta})}{\partial\theta_i}\boldsymbol{S}_t^{-1}(\boldsymbol{\theta})=2\frac{\partial\boldsymbol{S}_t}{\partial\theta_i}\boldsymbol{S}_t^{-1},$$

the conclusion follows from (5.4)–(5.8). □

Note that

$$\widehat{\boldsymbol{\gamma}}_m=\frac{1}{n}\sum_{t=1}^{n}\widehat{\boldsymbol{\Upsilon}}_t.$$

Let $\boldsymbol{\gamma}_m=n^{-1}\sum_{t=1}^{n}\boldsymbol{\Upsilon}_t$. We also need to introduce the notation

$$\boldsymbol{\eta}_t(\boldsymbol{\theta})=(\eta_{1t}(\boldsymbol{\theta}),\dots,\eta_{st}(\boldsymbol{\theta}))'=\boldsymbol{S}_t^{-1}(\boldsymbol{\theta})\boldsymbol{\epsilon}_t(\boldsymbol{\theta}).$$

Lemma 5.5 *Under the assumptions of Lemma 5.1, as* $n \to \infty$

$$\widehat{\boldsymbol{\rho}}_m \overset{o_P(n^{-1/2})}{=} \widehat{\boldsymbol{\gamma}}_m \overset{o_P(n^{-1/2})}{=} \boldsymbol{\gamma}_m + \boldsymbol{C}_m(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \tag{5.29}$$

with

$$\sqrt{n}\begin{pmatrix} \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0 \\ \boldsymbol{\gamma}_m \end{pmatrix} \xrightarrow{\mathcal{L}} \mathcal{N}\left\{\mathbf{0}, \begin{pmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\theta}} & \boldsymbol{\Sigma}_{\boldsymbol{\theta}\boldsymbol{\Upsilon}} \\ \boldsymbol{\Sigma}'_{\boldsymbol{\theta}\boldsymbol{\Upsilon}} & \boldsymbol{Id}_{md^2} \end{pmatrix}\right\}. \tag{5.30}$$

Proof In view of (5.10) and (5.11), the convergence (5.30) is a direct consequence of the central limit theorem applied to the martingale difference

$$\left\{(\boldsymbol{Z}'_t, \boldsymbol{\Upsilon}'_t)', \sigma(\boldsymbol{\eta}_u, u \le t)\right\}.$$

The first equality of (5.29) comes from the consistency of $\widehat{\boldsymbol{\Gamma}}(0)$ to $E\boldsymbol{\eta}_1\boldsymbol{\eta}'_1 = \mathbf{Id}_d$. Now note that, in view of (5.28) we have

$$E\frac{\partial \boldsymbol{\eta}_{t-\ell} \otimes \boldsymbol{\eta}_t}{\partial \boldsymbol{\theta}'}(\boldsymbol{\theta}_0) = E\boldsymbol{\eta}_{t-\ell} \otimes \frac{\partial \boldsymbol{\eta}_t}{\partial \boldsymbol{\theta}'}(\boldsymbol{\theta}_0) = \boldsymbol{c}_\ell, \tag{5.31}$$

as defined in (5.11). Letting $\boldsymbol{\Gamma}(\ell) = n^{-1}\sum_{t=\ell+1}^n \boldsymbol{\eta}_t\boldsymbol{\eta}'_{t-\ell}$, Lemma 5.3 and a Taylor expansion of the function $\boldsymbol{\theta} \mapsto n^{-1}\sum_{t=\ell+1}^n \boldsymbol{\eta}_{t-\ell}(\boldsymbol{\theta}) \otimes \boldsymbol{\eta}_t(\boldsymbol{\theta})$ around $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ entail

$$\mathrm{vec}\widehat{\boldsymbol{\Gamma}}(\ell) \overset{o_P(1)}{=} \mathrm{vec}\boldsymbol{\Gamma}(\ell) + \widehat{\boldsymbol{c}}_{n,\ell}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \tag{5.32}$$

where $\widehat{\boldsymbol{c}}_{n,\ell}$ is a $d^2 \times s$ matrix whose $is + j$-th row, $i, j \in \{1, \ldots, s\}$, is of the form

$$\frac{1}{n}\sum_{t=\ell+1}^n \frac{\partial \eta_{i\,t-\ell}\eta_{jt}}{\partial \boldsymbol{\theta}'}(\boldsymbol{\theta}^*_{i\,j}),$$

and $\boldsymbol{\theta}^*_{i\,j}$ lies between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. Using again a Taylor expansion, the k-th element of the previous vector is equal to

$$\frac{1}{n}\sum_{t=\ell+1}^n \frac{\partial \eta_{i\,t-\ell}\eta_{jt}}{\partial \theta_k}(\boldsymbol{\theta}_0) + \frac{1}{n}\sum_{t=\ell+1}^n \frac{\partial^2 \eta_{i\,t-\ell}\eta_{jt}}{\partial \boldsymbol{\theta}'\partial \theta_k}(\boldsymbol{\theta}^*_{i\,j,k})(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \tag{5.33}$$

for some $\boldsymbol{\theta}^*_{i\,j,k}$ between $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$. Lemma 5.4, the ergodic theorem and the strong consistency of $\widehat{\boldsymbol{\theta}}$ show that the second term of (5.33) tends a.s. to zero. By (5.31), it follows that $\widehat{\boldsymbol{c}}_{n,\ell} \to \boldsymbol{c}_\ell$ a.s. as $n \to \infty$, and the second equality of (5.29) follows from (5.32). □

***Proof of Proposition* 5.1** This is a direct consequence of Lemma 5.5. □

***Proof of Proposition* 5.2** In view of (5.29)–(5.30), we have

$$\widehat{\boldsymbol{M}}_{\boldsymbol{C}\perp\boldsymbol{D}}\sqrt{n}\widehat{\boldsymbol{\rho}}_m \overset{o_P(1)}{=} \boldsymbol{M}_{\boldsymbol{C}\perp\boldsymbol{D}}\sqrt{n}\boldsymbol{\gamma}_m \xrightarrow{\mathcal{L}} \mathcal{N}\left\{\boldsymbol{0}, \boldsymbol{M}_{\boldsymbol{C}\perp\boldsymbol{D}}\right\}.$$

Using Slutsky's lemma and a well-known result on quadratic forms of Gaussian vectors (see, e.g., van der Vaart ([37], Lemma 17.1)), and noting that $\boldsymbol{M}_{\boldsymbol{C}\perp\boldsymbol{D}}$ has s eigenvalues equal to 0 and $md^2 - s$ eigenvalues equal to 1, the conclusion follows. □

Proof of Proposition 5.3 It comes from the fact that $\boldsymbol{P}_+ = \boldsymbol{C}_m(\boldsymbol{C}_m'\boldsymbol{C}_m)^+\boldsymbol{C}_m' = \boldsymbol{B}_m(\boldsymbol{B}_m'\boldsymbol{B}_m)^{-1}\boldsymbol{B}_m'$, where the s_0 columns of $\boldsymbol{B}_m$ form a basis of the range of $\boldsymbol{C}_m$. □

Proof of Proposition 5.4 That α is the asymptotic level of these tests is a direct consequence of (5.12) and of well-known results on the quadratic forms of Gaussian vectors.

By a standard large deviation result (see Zolotarev [42]), we have

$$\log P\left(\sum_{i=1}^{r}\lambda_i N_i^2 > x\right) \sim \frac{-x}{2\lambda_1} \quad \text{as} \quad x \to \infty,$$

when $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r > 0$. The Bahadur slope c_{BP} follows from this result and the fact that, under H_1, $Q_m^{BP}/n \to \boldsymbol{\rho}_m^{*\prime}\boldsymbol{\rho}_m^*$. The two other Bahadur slopes are obtained similarly. For two symmetric and invertible matrices $\boldsymbol{A}$ and $\boldsymbol{B}$, it is well known that $\boldsymbol{A} - \boldsymbol{B}$ is positive semidefinite iff $\boldsymbol{B}^{-1} - \boldsymbol{A}^{-1}$ is positive semidefinite. We thus have $c_{Li} \geq c_{BP}$ because $\lambda_1\boldsymbol{I}_{md^2} - \boldsymbol{\Sigma}_{\boldsymbol{\rho}_m^*}$ is positive semidefinite, by definition of λ_1.

Since the projections are contractions, we have $c_K \leq \boldsymbol{\rho}_m^{*\prime}\boldsymbol{\rho}_m^*$ with equality if and only if $\boldsymbol{D}_m'\boldsymbol{\rho}_m^* = \boldsymbol{0}$. Note that the condition that $\boldsymbol{D}_m'\boldsymbol{C}_m$ is invertible is equivalent to $\mathrm{Col}^\perp(\boldsymbol{D}_m) \cap \mathrm{Col}(\boldsymbol{C}_m) = \emptyset$ ($\mathrm{Col}(\boldsymbol{A})$ denotes here the span of the columns of $\boldsymbol{A}$). This condition and the additional condition that $\boldsymbol{\rho}_m^* \notin \mathrm{Col}(\boldsymbol{C}_m)$, required for the consistency of the Q_m^K-test, are compatible with the condition $\boldsymbol{\rho}_m^* \in \mathrm{Col}^\perp(\boldsymbol{D}_m)$ required for the optimality of the test. The Bahadur slope c_{K+} is obtained similarly. □

Proof of Proposition 5.5 Let $\boldsymbol{m}_g = \boldsymbol{m}_g(\boldsymbol{\theta}_0) = E\boldsymbol{g}(\boldsymbol{\eta}_1)$ and let $\check{\boldsymbol{\eta}}_t^{\boldsymbol{g}} = \boldsymbol{\eta}_t^{\boldsymbol{g}}(\widehat{\boldsymbol{\theta}})$ with $\boldsymbol{\eta}_t^{\boldsymbol{g}}(\boldsymbol{\theta}) = \boldsymbol{g}\left(\boldsymbol{\eta}_t(\boldsymbol{\theta})\right) - E\boldsymbol{g}(\boldsymbol{\eta}_1)$. Using Taylor expansions

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\check{\boldsymbol{\eta}}_t^{\boldsymbol{g}}\check{\boldsymbol{\eta}}_{t-\ell}^{\boldsymbol{g}} \overset{o_P(1)}{=} \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\boldsymbol{\eta}_t^{\boldsymbol{g}}\boldsymbol{\eta}_{t-\ell}^{\boldsymbol{g}} + \left[\frac{\partial}{\partial\boldsymbol{\theta}'}\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\eta}_t^{\boldsymbol{g}}(\boldsymbol{\theta})\boldsymbol{\eta}_{t-\ell}^{\boldsymbol{g}}(\boldsymbol{\theta})\right]_{\boldsymbol{\theta}_0}\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

$$\sqrt{n}\{E\boldsymbol{g}(\boldsymbol{\eta}_1) - \overline{\boldsymbol{g}}_n\} \overset{o_P(1)}{=} \sqrt{n}\left(E\boldsymbol{g}(\boldsymbol{\eta}_1) - \frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\eta}_t^{\boldsymbol{g}}\right) + \left[\frac{\partial}{\partial\boldsymbol{\theta}'}\frac{1}{n}\sum_{t=1}^{n}\boldsymbol{\eta}_t^{\boldsymbol{g}}(\boldsymbol{\theta})\right]_{\boldsymbol{\theta}_0}\sqrt{n}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

we deduce

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\check{\boldsymbol{\eta}}_t^{\boldsymbol{g}}\check{\boldsymbol{\eta}}_{t-\ell}^{\boldsymbol{g}} = O_P(1), \qquad \boldsymbol{x}_n := \sqrt{n}\{E\boldsymbol{g}(\boldsymbol{\eta}_1) - \overline{\boldsymbol{g}}_n\} = O_P(1).$$

It follows that

$$\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{g}}(\ell) = \frac{1}{n}\sum_{t=1}^{n} \check{\boldsymbol{\eta}}_t^{\boldsymbol{g}} \check{\boldsymbol{\eta}}_{t-\ell}^{\boldsymbol{g}'} + \frac{1}{n}\sum_{t=1}^{n} \check{\boldsymbol{\eta}}_t^{\boldsymbol{g}} \boldsymbol{x}_n' + \boldsymbol{x}_n \frac{1}{n}\sum_{t=1}^{n} \check{\boldsymbol{\eta}}_t^{\boldsymbol{g}'} + \boldsymbol{x}_n \boldsymbol{x}_n'$$
$$\stackrel{o_P(n^{-1/2})}{=} \frac{1}{n}\sum_{t=1}^{n} \check{\boldsymbol{\eta}}_t^{\boldsymbol{g}} \check{\boldsymbol{\eta}}_{t-\ell}^{\boldsymbol{g}'}.$$

Letting $\boldsymbol{\Gamma}_{\boldsymbol{g}}(\ell) = n^{-1}\sum_{t=\ell+1}^{n} \boldsymbol{\eta}_t^{\boldsymbol{g}} \boldsymbol{\eta}_{t-\ell}^{\boldsymbol{g}}$, a Taylor expansion of $\boldsymbol{\theta} \mapsto n^{-1}\sum_{t=\ell+1}^{n} \boldsymbol{\eta}_{t-\ell}^{\boldsymbol{g}}(\boldsymbol{\theta}) \otimes \boldsymbol{\eta}_t^{\boldsymbol{g}}(\boldsymbol{\theta})$ around $\widehat{\boldsymbol{\theta}}$ and $\boldsymbol{\theta}_0$ gives

$$\text{vec}\widehat{\boldsymbol{\Gamma}}_{\boldsymbol{g}}(\ell) \stackrel{o_P(1)}{=} \text{vec}\boldsymbol{\Gamma}_{\boldsymbol{g}}(\ell) + \widehat{\boldsymbol{c}}_{n,\ell}^{\boldsymbol{g}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), \tag{5.34}$$

where $\widehat{\boldsymbol{c}}_{n,\ell}^{\boldsymbol{g}}$ is a $d_0^2 \times s$ matrix. The conclusion follows by arguments given in the proof of Lemma 5.5. □

Tedious Computations for Sect. 5.4.1
For the MA(1) model $X_t = \epsilon_t + b\epsilon_{t-1}$, we have

$$\gamma_X(0) = \sigma_\epsilon^2(1+b^2), \qquad \gamma_X(1) = \sigma_\epsilon^2 b, \qquad \rho_X(1) = \frac{b}{1+b^2} = a_0,$$
$$\sigma_0^2 = \gamma_X(0) + \rho_X^2(1)\gamma_X(0) - 2\rho_X(1)\gamma_X(1) = \sigma_\epsilon^2 \frac{1+b^2+b^4}{1+b^2}.$$

Letting $\eta_t = (X_t - a_0 X_{t-1})/\sigma_0$, we have

$$\begin{aligned}
\gamma_\eta(1) &= \frac{E(X_t - a_0X_{t-1})(X_{t-1} - a_0X_{t-2})}{\sigma_0^2}, \\
&= \frac{\gamma_X(1) - a_0\gamma_X(0) - a_0\gamma_X(2) + a_0^2\gamma_X(1)}{\sigma_0^2} = \frac{b^3}{(1+b^2)(1+b^2+b^4)}, \\
\gamma_\eta(2) &= \frac{E(X_t - a_0X_{t-1})(X_{t-2} - a_0X_{t-3})}{\sigma_0^2} = \frac{-a_0\gamma_X(1)}{\sigma_0^2} = \frac{-b^2}{1+b^2+b^4}, \\
EX_t^2X_{t-1}^2 &= E(\epsilon_2^2 + b^2\epsilon_1^2 + 2b\epsilon_2\epsilon_1)(\epsilon_1^2 + b^2\epsilon_0^2 + 2b\epsilon_1\epsilon_0) = \sigma_\epsilon^4(1+4b^2+b^4), \\
EX_t^4 &= E(\epsilon_1^4 + 4b\epsilon_1^3\epsilon_0 + 6b^2\epsilon_1^2\epsilon_0^2 + 4b^3\epsilon_1\epsilon_0^3 + b^4\epsilon_0^4) = 3\sigma_\epsilon^4(1+b^2)^2, \\
EX_t^3X_{t-1} &= E(\epsilon_2^3 + 3b\epsilon_2^2\epsilon_1 + 3b^2\epsilon_2\epsilon_1^2 + b^3\epsilon_1^3)(\epsilon_1 + b\epsilon_0) = 3\sigma_\epsilon^4 b(1+b^2), \\
EX_tX_{t-1}X_{t-2}^2 &= E(\epsilon_3 + b\epsilon_2)(\epsilon_2 + b\epsilon_1)(\epsilon_1^2 + b^2\epsilon_0^2 + 2b\epsilon_1\epsilon_0) = \sigma_\epsilon^4 b(1+b^2), \\
EX_tX_{t-1}^2X_{t-2} &= E(\epsilon_3 + b\epsilon_2)(\epsilon_2^2 + b^2\epsilon_1^2 + 2b\epsilon_2\epsilon_1)(\epsilon_1 + b\epsilon_0) = 2\sigma_\epsilon^4 b^2, \\
EX_tX_{t-1}^3 &= E(\epsilon_2 + b\epsilon_1)(\epsilon_1^3 + 3b\epsilon_1^2\epsilon_0 + 3b^2\epsilon_1\epsilon_0^2 + b^3\epsilon_0^3) = 3\sigma_\epsilon^4 b(1+b^2), \\
EX_t^2X_{t-2}^2 &= E(\epsilon_3^2 + b^2\epsilon_2^2 + 2b\epsilon_3\epsilon_2)(\epsilon_1^2 + b^2\epsilon_0^2 + 2b\epsilon_1\epsilon_0) = \sigma_\epsilon^4(1+b^2)^2, \\
EX_t^2X_{t-1}X_{t-2} &= E(\epsilon_3^2 + b^2\epsilon_2^2 + 2b\epsilon_3\epsilon_2)(\epsilon_2 + b\epsilon_1)(\epsilon_1 + b\epsilon_0) = \sigma_\epsilon^4 b(1+b^2), \\
E\eta_tX_{t-1} &= EX_1\eta_2^3 = 0, \\
E\eta_t^2X_{t-1}^2 &= \sigma_\epsilon^2 \frac{1+b^2+b^6+b^8}{(1+b^2)(1+b^2+b^4)},
\end{aligned}$$

$$
\begin{aligned}
E\eta_t^2 X_{t-1}X_{t-2} &= E\eta_t^2\eta_{t-1}X_{t-2} = \frac{\sigma_\epsilon^4}{\sigma_0^2}\frac{b(1+b^2+b^4)}{1+b^2} = b\sigma_\epsilon^2,\\
EX_t^2 X_{t-1}\eta_{t-1} &= \sigma_\epsilon^4(1+3b^2+b^4)/\sigma_0,\\
E\eta_t^2 X_{t-1}\eta_{t-1} &= \sigma_0, \qquad E\eta_t^2 X_{t-1}\eta_{t-2} = \sigma_0 b\frac{1+b^2}{1+b^2+b^4}\\
E\eta_t(1-\eta_t^2)\eta_{t-1} &= \frac{-2b^3}{(1+b^2)(1+b^2+b^4)}, \qquad E\eta_t(1-\eta_t^2)\eta_{t-2} = \frac{2b^2}{1+b^2+b^4},\\
E\eta_t^2\eta_{t-1}\eta_{t-2} &= \frac{b^3(1-b^2+b^4)}{(1+b^2)(1+b^2+b^4)^2}.
\end{aligned}
$$

Moreover, $E\eta_2^4 = 3$ noting that η_2 is Gaussian. The formulas for matrices $\boldsymbol{I}$ and $\boldsymbol{J}$ follow. With $\boldsymbol{\Upsilon}_t = (\eta_{t-1}\eta_t, \eta_{t-2}\eta_t, \ldots, \eta_{t-m}\eta_t)'$ and $\boldsymbol{Z}_t = \partial\ell_t(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta} = \begin{pmatrix} -2X_{t-1}\eta_t/\sigma_0 \\ (1-\eta_t^2)/\sigma_0^2 \end{pmatrix}$ we have

$$
\begin{aligned}
\boldsymbol{\Sigma}_{\boldsymbol{\theta}\boldsymbol{\Upsilon}} &= -\boldsymbol{J}^{-1}E\boldsymbol{Z}_t\boldsymbol{\Upsilon}_t'\\
&= -\boldsymbol{J}^{-1}\begin{pmatrix} -\frac{2}{\sigma_0}E{\eta_t}^2X_{t-1}\eta_{t-1} & -\frac{2}{\sigma_0}E{\eta_t}^2X_{t-1}\eta_{t-2} & 0\cdots 0\\ \frac{1}{\sigma_0^2}E\eta_t(1-\eta_t^2)\eta_{t-1} & \frac{1}{\sigma_0^2}E\eta_t(1-\eta_t^2)\eta_{t-2} & 0\cdots 0 \end{pmatrix}\\
&= -\begin{pmatrix} \frac{1}{2}\frac{1+b^2+b^4}{(1+b^2)^2} & 0\\ 0 & \frac{\sigma_\epsilon^4(1+b^4+b^2)^2}{(1+b^2)^2} \end{pmatrix}\begin{pmatrix} -2 & -2b\frac{1+b^2}{1+b^2+b^4} & 0\cdots 0\\ \frac{1}{\sigma_0^2}\frac{-2b^3}{(1+b^2)(1+b^2+b^4)} & \frac{1}{\sigma_0^2}\frac{2b^2}{1+b^2+b^4} & 0\cdots 0 \end{pmatrix}\\
&= \begin{pmatrix} \frac{1+b^2+b^4}{(1+b^2)^2} & \frac{b}{1+b^2} & 0\cdots 0\\ \frac{-2b^3}{(1+b^2)^2}\sigma_\epsilon^2 & \frac{2b^2}{1+b^2}\sigma_\epsilon^2 & 0\cdots 0 \end{pmatrix}.
\end{aligned}
$$

It follows that

$$
\begin{aligned}
\boldsymbol{C}_m\boldsymbol{\Sigma}_{\boldsymbol{\theta}\boldsymbol{\Upsilon}} &= \begin{pmatrix} -1 & 0\\ \frac{-b(1+b^2)}{1+b^2+b^4} & 0\\ 0 & 0\\ \vdots & \vdots\\ 0 & 0 \end{pmatrix}\begin{pmatrix} \frac{1+b^2+b^4}{(1+b^2)^2} & \frac{b}{1+b^2} & 0\cdots 0\\ \frac{-2b^3}{(1+b^2)^2}\sigma_\epsilon^2 & \frac{2b^2}{1+b^2}\sigma_\epsilon^2 & 0\cdots 0 \end{pmatrix}\\
&= \begin{pmatrix} -\frac{1+b^2+b^4}{(1+b^2)^2} & \frac{-b}{1+b^2} & 0\cdots 0\\ \frac{-b}{1+b^2} & \frac{-b^2}{1+b^2+b^4} & 0\cdots 0\\ 0 & 0 & 0\cdots 0\\ \vdots & & \ddots\\ 0 & \cdots & 0 \end{pmatrix}.
\end{aligned}
$$

Moreover

$$\boldsymbol{C}_m \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{C}_m' = \frac{1+b^2+b^4}{(1+b^2)^2} \begin{pmatrix} 1 & -c_2^* & 0 \cdots 0 \\ -c_2^* & (c_2^*)^2 & 0 \cdots 0 \\ 0 & 0 & 0 \cdots 0 \\ \vdots & & \ddots \\ 0 & \cdots & 0 \end{pmatrix} = \begin{pmatrix} \frac{1+b^2+b^4}{(1+b^2)^2} & \frac{b}{1+b^2} & 0 \cdots 0 \\ \frac{b}{1+b^2} & \frac{b^2}{1+b^2+b^4} & 0 \cdots 0 \\ 0 & 0 & 0 \cdots 0 \\ \vdots & & \ddots \\ 0 & \cdots & 0 \end{pmatrix}.$$

Thus, we can obtain the expression of $\boldsymbol{I}_m + \boldsymbol{C}_m \boldsymbol{\Sigma}_{\boldsymbol{\theta}\boldsymbol{\Upsilon}} + \boldsymbol{\Sigma}_{\boldsymbol{\theta}\boldsymbol{\Upsilon}}' \boldsymbol{C}_m' + \boldsymbol{C}_m \boldsymbol{\Sigma}_{\boldsymbol{\theta}} \boldsymbol{C}_m'$. We have

$$E(\boldsymbol{\Upsilon}_t \boldsymbol{\Upsilon}_t') = \begin{pmatrix} 1 + \frac{2b^6}{(1+b^2)^2(1+b^2+b^4)^2} & \frac{b^3(1-b^2+b^4)}{(1+b^2)(1+b^2+b^4)^2} & 0 \cdots 0 \\ \frac{b^3(1-b^2+b^4)}{(1+b^2)(1+b^2+b^4)^2} & 1 + \frac{3b^4}{(1+b^2+b^4)^2} & 0 \cdots 0 \\ 0 & 0 & 1 \cdots 0 \\ \vdots & \vdots & \ddots \\ 0 & \cdots & 0 \cdots 1 \end{pmatrix},$$

so that we also have the expression of $\boldsymbol{\Sigma}_{\boldsymbol{\rho}_m^*}$. We thus have

$$\boldsymbol{\Sigma}_{\boldsymbol{\rho}_m^*}^{-1} = \begin{pmatrix} c(1+b^2)^2(1+b^2+5b^4+b^6+b^8) & cb(1+b^2+4b^4+b^6+b^8)(1+b^2) & 0 \cdots 0 \\ cb(1+b^2+4b^4+b^6+b^8)(1+b^2) & cb^2(1+2b^2+5b^4+2b^6+b^8) & 0 \cdots 0 \\ 0 & 0 & 1 \cdots 0 \\ \vdots & \vdots & \ddots \\ 0 & \cdots & 0 \cdots 1 \end{pmatrix}$$

where $c = \frac{(1+b^2+b^4)^2}{b^4(1+3b^2+8b^4+11b^6+8b^8+3b^{10}+b^{12})}$. The Bahadur slope of the Li test follows.

References

1. Akashi, F., Taniguchi, M., Monti, A.C. and Amano, T. (2021). *Diagnostic Methods in Time Series.* Springer, Singapore.
2. Amendola, A. and Francq, C. (2009). Concepts of and tools for nonlinear time series modelling. *Handbook of Computational Econometrics*, Edts: D. Belsley and E. Kontoghiorghes, Wiley.
3. Andrews, D.W.K. (1987). Asymptotic results for generalized Wald tests. *Econometric Theory* 3, 348–358.
4. Bahadur, R.R. (1960). Stochastic comparison of tests. *The Annals of Mathematical Statistics* **31** 276–295.
5. Billingsley, P. (1961). The Lindeberg–Levy theorem for martingales. *Proceedings of the American Mathematical Society* **12** 788–792.
6. Billingsley, P. (1995). *Probability and Measure.* John Wiley & Sons, New York.
7. Bollerslev, T. (2009). Glossary to ARCH (GARCH). *In: Bollerslev, T., Russell, J. R., Watson, M. (Eds.), Volatility and Time Series Econometrics: Essays in Honour of Robert F. Engle*. Oxford University Press, Oxford.

8. Boubacar Maïnassara, Y. and Saussereau, B. (2018). Diagnostic checking in multivariate ARMA models with dependent errors using normalized residual autocorrelations. *Journal of the American Statistical Association* **113** 1813–1827.
9. Box, G.E.P. and Pierce, D.A. (1970). Distribution of residual autocorrelations in autoregressive-integreted moving average time series models. *Journal of the American Statistical Association* **65** 1509–1526.
10. Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods*. Springer-Verlag, New-York.
11. Chitturi, R. V. (1974). Distribution of residual autocorrelations in multiple autoregressive schemes. *Journal of the American Statistical Association* **65** 1509–1526.
12. Duchesne, P. and Francq, C. (2008). On diagnostic checking time series models with portmanteau test statistics based on generalized inverses and 2-inverses. *COMPSTAT 2008, Proceedings in Computational Statistics*, 143–154.
13. Duchesne, P. and Lafaye de Micheaux, P. (2010). Computing the distribution of quadratic forms: Further comparisons between the Liu-Tang-Zhang approximation and exact methods. *Computational Statistics and Data Analysis* **54** 858–862.
14. Fisher, T.J. and Gallagher, C.M. (2012). New weighted portmanteau statistics for time series goodness of fit testing. *Journal of the American Statistical Association* **107** 777–787.
15. Francq, C., Roy, R. and Zakoïan, J- M. (2005). Diagnostic checking in ARMA models with uncorrelated errors. *Journal of the American Statistical Association* **100** 532–544.
16. Francq, C. and Zakoïan, J- M. (1998). Estimating linear representations of nonlinear processes. *Journal of Statistical Planning and Inference*. **68** 145–165.
17. Francq, C. and Zakoïan, J.- M. (2004). Maximum likelihood estimation of pure GARCH and ARMA-GARCH processes. *Bernoulli* **10** 605–637.
18. Harville, D.A. (1997). *Matrix Algebra From a Statistician's Perspective.* Springer-Verlag, New York.
19. Hosking, J.R.M. (1980). The multivariate portmanteau statistic. *Journal of the American Statistical Association* **75** 343–386.
20. Imhof, J.P. (1961). Computing the distribution of quadratic forms in normal variables. *Biometrika* **48** 419–426.
21. Katayama, N. (2008). An improvement of the portmanteau statistic. *Journal of Time Series Analysis* **29** 359–370.
22. Lafaye de Micheaux, P. (2017). Package CompQuadForm. *CRAN Repository*.
23. Li, W.K. (1992). On the asymptotic standard errors of residual autocorrelations in nonlinear time series modelling. *Biometrika* **79** 435-437.
24. Li, W.K. (2004). *Diagnostic Checks in Time Series*. Chapman & Hall/CRC, New York.
25. Li, W.K. and Mak, T.K. (1994). On the squared residual autocorrelations in non-linear time series with conditional heteroskedasticity. *Journal of Time Series Analysis* **15** 627–636.
26. Ling, S. and Li, W.K. (1997). On fractionally integrated autoregressive moving-average time series models with conditional heteroscedasticity. *Journal of the American Statistical Association* **92** 1184–1194.
27. Ljung, G.M. and Box, G.E.P. (1978). On measure of lack of fit in time series models. *Biometrika* **65** 297–303.
28. Lütkepohl, H. (1993). *Introduction to Multiple Time Series Analysis.* Springer Verlag, Berlin.
29. Mcleod, A.I. (1978). On the distribution of residual autocorrelations in Box–Jenkins method. *Journal of the Royal Statistical Society: Series B* **40** 296-302.
30. McLeod, A.I. and Li, W.K. (1983). Diagnostic checking ARMA time series models using squared-residual autocorrelations. *Journal of Time Series Analysis* 4, 269–273.
31. Pérez, A. and Ruiz, E. (2003). Properties of the sample autocorrelations of nonlinear transformations in long-memory stochastic volatility models. *Journal of Financial Econometrics* **1** 420–444.
32. Pötscher, B.M. and Prucha, I.R. (1997). *Dynamic Nonlinear Econometric Models*. Springer, Berlin.

33. Romano, J.L. and Thombs, L.A. (1996). Inference for autocorrelations under weak assumptions. *Journal of the American Statistical Association* **91** 590–600.
34. Shao, X. (2011). Testing for white noise under unknown dependence and its applications to diagnostic checking for time series models. *Econometric Theory* **27** 312–343.
35. Taniguchi, M. and Kakizawa, Y. (2000). *Asymptotic Theory of Statistical Inference for Time Series*. Springer, New York.
36. Tsay, R.S. (2010). *Analysis of Financial Time Series*, 3rd Edition, John Wiley & Sons.
37. van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press.
38. Wald, A. (1949). Note on the consistency of the maximum likelihood estimate. *Annals of Mathematical Statistics* **20** 595–601.
39. Wang, X. and Sun, Y. (2020). A simple asymptotically F-distributed portmanteau test for diagnostic checking of time series models with uncorrelated innovations. *Journal of Business & Economic Statistics* **40** 1–17.
40. Zhu, K. (2016). Bootstrapping the portmanteau tests in weak autoregressive moving average models. *Journal of the Royal Statistical Society: Series B* **78** 463–485.
41. Zhu, K. and Li, W.K. (2015). A bootstrapped spectral test for adequacy in weak ARMA models. *Journal of Econometrics* **187** 113–130.
42. Zolotarev, V. M. (1961). Concerning a certain probability problem. *Theory of Probability and Its Applications* **6** 201–204.

Chapter 6
Parameter Estimation of Standard AR(1) and MA(1) Models Driven by a Non-I.I.D. Noise

Violetta Dalla, Liudas Giraitis, and Murad S. Taqqu

Abstract The use of a non-i.i.d. noise in parametric modeling of stationary time series can lead to unexpected distortions of the standard errors and confidence intervals in parameter estimation. We consider AR(1) and MA(1) models and motivate the need for correction of standard errors when these are generated by a non-i.i.d. noise. The impact of the noise on the standard errors and confidence intervals is illustrated with Monte Carlo simulations using various types of noise.

6.1 Introduction

Many applications in statistics and econometrics involve the use of parametric stationary time series models. The well-known parametric ARMA models popularized by [2] and long-memory ARFIMA models introduced by [6, 7] can be represented as a linear process

$$X_t = \sum_{j=0}^{\infty} a_{\theta,j}\eta_{t-j}, \quad t = ..., -1, 0, 1, ..., \tag{6.1}$$

where $\{\eta_j\}$ is an underlying uncorrelated noise with $E[\eta_j] = 0$, $E[\eta_j^2] = \sigma_\eta^2$, and weights $a_{\theta,j}$. These weights are parameterized by a parameter θ and satisfy $\sum_{j=0}^{\infty} a_{\theta,j}^2 < \infty$, with $a_{\theta,0} = 1$.

V. Dalla
National and Kapodistrian University of Athens, Sofokleous 1, Athens 10559, Greece
e-mail: vidalla@econ.uoa.gr

L. Giraitis (✉)
Queen Mary University of London, Mile End Road, London E1 4NS, UK
e-mail: l.giraitis@qmul.ac.uk

M. S. Taqqu
Boston University, 111 Cummington Mall, Boston, MA 02215, USA
e-mail: murad@bu.edu

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_6

Our contribution is built on the work by [5], who analyzed the impact of a non-i.i.d. noise on the asymptotic properties of the Whittle estimator of the parameter θ for time series as in (6.1). The paper [5] extended the literature on the Whittle estimation by deriving analytical expressions for the standard errors when using the normal approximation for the Whittle estimator of the parameter θ, for a class of time series which includes stationary ARMA models. They showed that the distortions of the standard errors, caused by the presence of a non-i.i.d. noise $\{\eta_j\}$, can be large. They showed further that the standard errors of the estimators are related to the dependence structure of an underlying uncorrelated noise $\{\eta_j\}$ in a complex manner. The paper [5] provided the estimators of the standard errors for dynamic parameters of AR(1) and MA(1) models with zero mean, $E[X_t] = 0$. The estimation of the standard errors for other ARMA models remains an open problem.

In this paper, we focus solely on the estimation of AR(1) and MA(1) models and extend the work [5] in a number of directions. Firstly, we derive asymptotic normality and analytical expressions of the standard errors for the estimators of the parameters for the case of a non-zero mean. Secondly, we provide an extensive simulation study on coverage intervals for the parameters based on the estimated standard errors. This study shows that the empirical confidence intervals adjusted for a non-i.i.d. noise produce coverage probabilities that are close to the nominal. Simulations confirm that significant coverage distortions may arise when the standard confidence intervals, valid for an i.i.d. noise, are employed and the noise is not i.i.d. We show finally that the estimation of the mean of X_t is not affected by a potential latent dependence in the noise and it does not require adjustment.

Throughout this paper, we denote by $\to_p$ and $\to_D$ the convergence in probability and distribution, respectively, while C denotes generic constants.

6.2 Main Results

It has long been known that the Whittle estimation method is a convenient alternative to the maximum-likelihood method for estimating stationary parametric time series models with a linear representation as in (6.1), see [3, 8, 10] and [4, Chap. 8]. While the existing literature on the Whittle estimation enables us to understand the impact of dependence on the estimation, in order to obtain the asymptotic distribution of the Whittle estimator, one typically assumes that the noise $\{\eta_j\}$ is i.i.d.

Recently, [5] presented the corresponding asymptotic results on the Whittle estimation for linear time series with a non-i.i.d. noise. They imposed strong assumptions on the weights $a_{\theta,j}$, which are satisfied by stationary ARMA models. The paper [5, Theorem 2.2] also showed that in the presence of a non-i.i.d. noise, the asymptotic variance of the Whittle estimator has an infeasible structure. This variance might be substantially different from the variance derived for the case of i.i.d. noise, which subsequently complicates the estimation of the standard errors and building confidence intervals for parameters. In addition, [5] examined estimation of parameters of stationary AR(1) and MA(1) models with zero mean, and showed how to estimate the standard errors from the data.

Our primary interest is to provide asymptotic theory and feasible confidence intervals for parameters of the AR(1) and MA(1) models with non-zero mean, and to verify the accuracy of the theoretical asymptotic results using Monte Carlo simulations. We focus on the models:

$$\text{AR(1) model}: \quad X_t = \alpha_0 + \phi_0 X_{t-1} + \eta_t, \quad |\phi_0| < 1, \tag{6.2}$$

$$\text{MA(1) model}: \quad X_t = \alpha_0 + \eta_t - \theta_0 \eta_{t-1}, \quad |\theta_0| < 1, \tag{6.3}$$

where $\{\eta_t\}$ is a stationary uncorrelated noise with $E[\eta_t] = 0$ and $E[\eta_t^2] = \sigma_\eta^2$. We make the following assumptions:

– A1: $\{\eta_j\}$ is a stationary ergodic martingale difference sequence with respect to some natural filtration $\mathcal{F}_j$, namely $E[\eta_j|\mathcal{F}_{j-1}] = 0$,

– A2: $E[\eta_j^4] < \infty$.

For example, $\mathcal{F}_j$ could denote the σ-field generated by $(\eta_j, \eta_{j-1}, ...)$.

We estimate the unknown mean EX_t by the sample mean $\bar{X} = n^{-1} \sum_{j=1}^n X_j$. If $EX_j \neq 0$, then the Whittle estimator based on the demeaned data $X_j - \bar{X}$ has the same asymptotic properties as that based on the centered data $X_j - EX_j$, see [4, Chap. 8].

Estimation of AR(1) model. The Whittle estimator of the parameter ϕ_0 is the sample autocorrelation at the lag 1, see [5]:

$$\widehat{\phi} = \frac{\sum_{t=2}^n (X_t - \bar{X})(X_{t-1} - \bar{X})}{\sum_{t=2}^n (X_{t-1} - \bar{X})^2}.$$

We derive an estimator for the intercept α_0 by combining (6.2) and $\widehat{\phi}$, as follows:

$$\widehat{\alpha} = (n-1)^{-1} \sum_{t=2}^n (X_t - \widehat{\phi} X_{t-1}).$$

Recall that $E[X_j] = \alpha_0(1-\phi_0)^{-1}$ and $\text{var}(X_j) = \sigma_\eta^2(1-\phi_0^2)^{-1}$. We estimate the mean EX_j by the sample mean $\bar{X}$.

Theorem 6.1 *The estimators $\widehat{\phi}$, $\widehat{\alpha}$ of the parameters ϕ_0, α_0 in the AR(1) model (6.2) have the following properties:*

$$n^{1/2}(\widehat{\phi} - \phi_0) \to_D \mathcal{N}(0, v_\phi^2), \tag{6.4}$$

$$v_\phi^2 := \frac{E[\eta_1^2 (X_0 - EX_0)^2]}{\text{var}^2(X_1)} = (1-\phi_0^2) + \frac{\text{cov}(\eta_1^2,\ (X_0 - EX_0)^2)}{\text{var}^2(X_1)},$$

$$n^{1/2}(\widehat{\alpha} - \alpha_0) \to_D \mathcal{N}(0, v_\alpha^2), \tag{6.5}$$

$$v_\alpha^2 := E\big[\eta_1^2\big(\frac{E[X_1](X_0 - EX_0)}{\text{var}(X_1)} + 1\big)^2\big].$$

Moreover,

$$n^{1/2}(\bar{X} - EX_1) \to_D \mathcal{N}(0, s_x^2), \tag{6.6}$$

where

$$s_x^2 := \sum_{k=-\infty}^{\infty} \mathrm{cov}(X_k, X_0) = \frac{\sigma_\eta^2}{(1-\phi_0)^2}.$$

The impact of the noise on the asymptotic variances in (6.4) and (6.5) is complex. If $\{\eta_j\}$ is an i.i.d. noise, then in Theorem 6.1,

$$v_\phi^2 = 1 - \phi_0^2, \tag{6.7}$$
$$v_\alpha^2 = E[\eta_1^2]\Big(\frac{(E[X_1])^2}{\mathrm{var}(X_1)} + 1\Big) = \frac{\alpha_0^2(1+\phi_0)}{1-\phi_0} + \sigma_\eta^2.$$

The long-run variance s_x^2 in (6.6) depends on the variance σ_η^2 of the noise and is not affected by the hidden dependence in $\{\eta_j\}$.

The unknown variances v_ϕ^2 and v_α^2 can be consistently estimated by

$$\widehat{v}_\phi^2 = \frac{n^{-1}\sum_{t=2}^n \widehat{\eta}_t^2 (X_{t-1} - \bar{X})^2}{\widehat{\sigma}_x^4}, \quad \widehat{\sigma}_x^2 = n^{-1}\sum_{t=1}^n (X_t - \bar{X})^2, \tag{6.8}$$
$$\widehat{v}_\alpha^2 = (n-1)^{-1}\sum_{t=2}^n \widehat{\eta}_t^2 \Big(\frac{(X_{t-1} - \bar{X})\bar{X}}{\widehat{\sigma}_x^2} + 1\Big)^2,$$

where

$$\widehat{\eta}_t = X_t - \bar{X} - \widehat{\phi}(X_{t-1} - \bar{X}).$$

These estimators can be used to evaluate the standard errors $SE_\phi = v_\phi/\sqrt{n}$ and $SE_\alpha = v_\alpha/\sqrt{n}$ required to build the confidence intervals for the parameters ϕ and α.

Corollary 6.1 *Under the assumptions of Theorem 6.1, as $n \to \infty$,*

$$\widehat{v}_\phi^2 \to_p v_\phi^2, \qquad \widehat{v}_\alpha^2 \to_p v_\alpha^2. \tag{6.9}$$

In the proof of Theorem 6.1, we will use the following useful result.

Lemma 6.1 ([1, Theorem 2.1]) *Suppose that $\{Y_t\}$ is a linear process*

$$Y_t = \sum_{j=0}^{\infty} b_j \xi_{t-j}, \tag{6.10}$$

where $\{\xi_j\}$ *is a stationary ergodic martingale difference noise with* $E[\xi_j] = 0$, $E[\xi_j^2] < \infty$ *and the* b_j*'s are non-random weights,* $\sum_{j=0}^{\infty} b_j^2 < \infty$. *Then*

$$v_n^2 = \mathrm{var}\Big(\sum_{t=1}^{n} Y_t\Big) \to \infty, \quad n \to \infty, \tag{6.11}$$

implies

$$v_n^{-1} \sum_{t=1}^{n} Y_t \to_D \mathcal{N}(0, 1). \tag{6.12}$$

Proof of Theorem 6.1 We start with the proof of (6.4). The AR(1) model (6.2) can be written as a linear process

$$X_t - EX_t = \sum_{j=0}^{\infty} \phi_0^j \eta_{t-j}, \tag{6.13}$$

where

$$n^{-1}\mathrm{var}\Big(\sum_{t=1}^{n}(X_t - EX_t)\Big) \to s_x^2 = \sum_{k=-\infty}^{\infty} \mathrm{cov}(X_k, X_0) = \frac{\sigma_\eta^2}{(1-\phi_0)^2}. \tag{6.14}$$

It is well known that centering the data X_t by the sample mean $\bar{X}$ does not change the asymptotic properties of the Whittle estimators of the parameters of the AR(1) model, see [4, Chap. 8]. Therefore, the claim (6.4) follows from [5, Corollary 5.1].

Next, we prove (6.5). Since $X_t - \widehat{\phi} X_{t-1} = \alpha_0 + (\phi_0 - \widehat{\phi}) X_{t-1} + \eta_t$, then by definition of $\widehat{\alpha}$,

$$\begin{aligned}\widehat{\alpha} - \alpha_0 &= (n-1)^{-1} \sum_{t=2}^{n} \{(\phi_0 - \widehat{\phi}) X_{t-1} + \eta_t\} \\ &= (\phi_0 - \widehat{\phi}) E[X_1] + (n-1)^{-1} \sum_{t=2}^{n} \eta_t + o_p(n^{-1/2}),\end{aligned} \tag{6.15}$$

noting that $\phi_0 - \widehat{\phi} = O_p(n^{-1/2})$ by (6.4) and

$$(n-1)^{-1} \sum_{t=2}^{n} X_{t-1} = E[X_1] + o_p(1)$$

by (6.14). We will show below that

$$\widehat{\phi} - \phi_0 = (n-1)^{-1} \sum_{t=2}^{n} \eta_t \frac{X_{t-1} - EX_{t-1}}{\mathrm{var}(X_1)} + o_p(1). \tag{6.16}$$

Set $\gamma = E[X_1]/\mathrm{var}(X_1)$. Together with (6.15), this implies

$$\widehat{\alpha} - \alpha_0 = (n-1)^{-1} \sum_{t=2}^{n} z_t + o_p(1), \tag{6.17}$$

where

$$z_t := \eta_t\{\gamma(X_{t-1} - EX_{t-1}) + 1\}.$$

The sequence $\{z_t\}$ is a stationary martingale difference sequence. Since $\{X_t\}$ is a causal linear process (6.13), by [9, Theorem 3.5.8], $\{z_t\}$ is an ergodic sequence. As seen below, $E[X_t^4] < \infty$. Therefore, $E[z_t^2] < \infty$, and

$$v_n^2 = E(\sum_{t=2}^{n} z_t)^2 = \sum_{t=2}^{n} E[z_t^2] = (n-1)E[z_1^2] \to \infty.$$

Hence, by Lemma 6.1,

$$(nE[z_1^2])^{-1/2} \sum_{t=2}^{n} z_t \to_D \mathcal{N}(0, 1),$$

where

$$E[z_1^2] = E\Big[\eta_1^2\Big(\frac{E[X_1](X_0 - EX_0)}{\mathrm{var}(X_1)} + 1\Big)^2\Big],$$

which together with (6.17) proves (6.5).

Proof of (6.16). Recall $\mu = E[X_t] = \alpha_0(1-\phi_0)^{-1}$. Then we can write

$$\begin{aligned} X_t - \bar{X} &= \alpha_0 + \phi_0 X_{t-1} + \eta_t - \bar{X} \\ &= \alpha_0 + \phi_0(X_{t-1} - \bar{X}) + \eta_t + (\bar{X} - EX_t)(\phi_0 - 1) + E[X_t](\phi_0 - 1) \\ &= \phi_0(X_{t-1} - \bar{X}) + \eta_t + a_n, \end{aligned}$$

where $a_n = (\bar{X} - EX_t)(\phi_0 - 1)$. This together with the definition of $\widehat{\phi}$ gives

$$\begin{aligned} \widehat{\phi} - \phi_0 &= \frac{\sum_{t=2}^{n}(X_t - \bar{X})(X_{t-1} - \bar{X})}{\sum_{t=2}^{n}(X_{t-1} - \bar{X})^2} - \phi_0 \\ &= \frac{\sum_{t=2}^{n} \eta_t(X_{t-1} - \bar{X})}{\sum_{t=2}^{n}(X_{t-1} - \bar{X})^2} + \frac{a_n \sum_{t=2}^{n}(X_{t-1} - \bar{X})}{\sum_{t=2}^{n}(X_{t-1} - \bar{X})^2}. \end{aligned}$$

We will show that

$$\sum_{t=2}^{n} \eta_t (X_{t-1} - \bar{X}) = \sum_{t=2}^{n} \eta_t (X_{t-1} - \mu) + o_p(n), \tag{6.18}$$

$$a_n \sum_{t=2}^{n} (X_{t-1} - \bar{X}) = o_p(n), \tag{6.19}$$

$$(n-1)^{-1} \sum_{t=2}^{n} (X_{t-1} - \bar{X})^2 \to_p \operatorname{var}(X_1), \tag{6.20}$$

which implies (6.16).

To verify (6.18), it suffices to note that

$$(\mu - \bar{X})(\sum_{t=2}^{n} \eta_t) = O_p(n^{-1/2}) O_p(n^{1/2}) = O_p(1)$$

by (6.6), and noting that

$$E[(\sum_{t=2}^{n} \eta_t)^2] = (n-1)\sigma_\eta^2$$

implies $\sum_{t=2}^{n} \eta_t = O_p(n^{1/2})$. Also, (6.19) holds because (6.6) implies $a_n = O_p(n^{-1/2})$ and

$$\begin{aligned} \sum_{t=2}^{n} (X_{t-1} - \bar{X}) &= \sum_{t=2}^{n+1} (X_{t-1} - \bar{X}) - (X_n - \bar{X}) \\ &= -(X_n - \bar{X}) = O_p(1). \end{aligned}$$

Finally, to prove (6.20), it suffices to notice that

$$n^{-1} \sum_{t=1}^{n} X_t^k \to_p E[X_1^k], \quad \text{for } k = 1, 2, 4, \tag{6.21}$$

which follows from [9, Theorem 3.5.7] using the following facts:

– $\{X_t\}$ has the MA representation (6.13),
– $\{X_t^k\}$ is a stationary ergodic sequence, see [9, Theorem 3.5.8],
– $E[X_t^4] < \infty$ by [1, Lemma 3.1]

using (6.13) and $E[\eta_j^4] < \infty$.

Further, convergence (6.6) follows using the equality

$$n^{1/2}(\bar{X} - EX_1) = n^{-1/2} \sum_{t=1}^{n} (X_t - EX_t),$$

Equation (6.13) and Lemma 6.1. This completes the proof of Theorem 6.1. □

Proof of Corollary 6.1 The same argument as in the proof of (6.21) implies that sequences $\{\eta_t^2 X_{t-1}^k\}$, $k = 1, 2$ are stationary and ergodic, and

$$E[\eta_t^2 X_{t-1}^k] \leq (E[\eta_t^4] E[X_{t-1}^{2k}])^{1/2} < \infty.$$

By [9, Theorem 3.5.7], we have

$$n^{-1} \sum_{t=1}^{n} \eta_t^2 X_{t-1}^k \to_p E[\eta_1^2 X_0^k], \quad \text{for } k = 1, 2. \tag{6.22}$$

Using (6.8), (6.21), (6.22) and the property $\widehat{\phi} - \phi_0 = O_p(n^{-1/2})$, we obtain (6.9). □

Estimation of MA(1) model. The model (6.3) has the spectral density

$$f(u) = \sigma_\eta^2 k_{\theta_0}(u),$$
$$k_\theta(u) = 1 - 2\theta \cos(u) + \theta^2.$$

The Whittle estimator of parameter θ_0 in the MA(1) model is obtained by minimizing the Whittle objective function, see [5]:

$$\widehat{\theta} = \text{argmin}_{\theta \in [-a,\, a]} Q_n(\theta), \tag{6.23}$$

where

$$Q_n(\theta) = \int_{-\pi}^{\pi} \frac{I_n(u)}{k_\theta(u)} \mathrm{d}u,$$
$$I_n(u) = (2\pi n)^{-1} \Big| \sum_{t=1}^{n} (X_t - \bar{X}) e^{\mathrm{i}tu} \Big|^2, \quad |u| \leq \pi,$$

and $[-a,\, a] \subset (-1, 1)$. Recall that $\mu = E[X_j] = \alpha_0$ and $\text{var}(X_j) = \sigma_\eta^2(1 + \theta_0^2)$. We estimate the mean $E[X_t] = \alpha_0$ by the sample mean $\bar{X}$. Denote

$$Z_t = \sum_{s=0}^{\infty} \theta_0^s \eta_{t-s}, \quad t = ..., -1, 0, 1, ...$$

Theorem 6.2 *The estimators* $\widehat{\theta}$, $\bar{X}$ *of the parameters* θ_0, α_0 *in the MA(*1*) model (6.3) have the following properties:*

$$n^{1/2}(\widehat{\theta} - \theta_0) \to_D \mathcal{N}(0, v_\theta^2), \tag{6.24}$$

$$v_\theta^2 := \frac{E[\eta_1^2 Z_0^2]}{\mathrm{var}^2(Z_0)} = (1 - \theta_0^2) + \frac{\mathrm{cov}(\eta_1^2,\ Z_0^2)}{\mathrm{var}^2(Z_0)},$$

$$n^{1/2}(\bar{X} - EX_1) \to_D \mathcal{N}(0, s_x^2), \tag{6.25}$$

$$s_x^2 = \sum_{k=-\infty}^{\infty} \mathrm{cov}(X_k, X_0) = \sigma_\eta^2 (1 - \theta_0)^2.$$

The impact of the noise $\{\eta_j\}$ on v_θ^2 in (6.24) is complex, while for an i.i.d. noise, $v_\theta^2 = 1 - \theta_0^2$. The long-run variance s_x^2 in (6.25) depends on the variance σ_η^2 of the noise and is not affected by its latent dependence structure.

The paper [5] showed that the unknown variance v_θ^2 can be estimated consistently. They assumed $EX_t = \alpha_0 = 0$. Notice that

$$\eta_t = (1 - \theta_0 L)^{-1}(X_t - EX_t) = \sum_{s=0}^{\infty} \theta_0^s (X_{t-s} - EX_{t-s}),$$

$$Z_t = (1 - \theta_0 L)^{-1}\eta_t = \sum_{s=0}^{\infty} (s+1)\theta_0^s (X_{t-s} - EX_{t-s}).$$

The following estimator of v_θ^2 extends [5] to the case $EX_j \neq 0$:

$$\widehat{v}_\theta^2 = \frac{n^{-1} \sum_{j=2}^{n} \widehat{Z}_{j-1}^2 \widehat{\eta}_j^2}{\left(n^{-1} \sum_{j=1}^{n} \widehat{Z}_j^2\right)^2}, \tag{6.26}$$

where

$$\widehat{Z}_t = \sum_{s=0}^{t-1} (s+1)\widehat{\theta}^s (X_{t-s} - \bar{X}), \quad \widehat{\eta}_t = \sum_{s=0}^{t-1} \widehat{\theta}^s (X_{t-s} - \bar{X}).$$

This estimator allows us to evaluate the standard error $SE_\theta = v_\theta/\sqrt{n}$ and hence to build confidence intervals for the parameter θ_0.

Corollary 6.2 *Under the assumptions of Theorem 6.2, as $n \to \infty$,*

$$\widehat{v}_\theta^2 \to_p v_\theta^2. \tag{6.27}$$

***Proof of Theorem* 6.2** The MA(1) model (6.3) satisfies

$$X_t - EX_t = \eta_t - \theta_0 \eta_{t-1}, \tag{6.28}$$

$$n^{-1}\mathrm{var}\Big(\sum_{t=1}^{n} (X_t - EX_t)\Big) \to s_x^2 = \sum_{k=-\infty}^{\infty} \mathrm{cov}(X_k, X_0) = \sigma_\eta^2 (1 - \theta_0)^2.$$

Hence, similarly to the AR(1) model, centering X_j by the sample mean $\bar{X}$ does not affect the asymptotic properties of the Whittle estimator of the parameter θ_0 of the MA(1) model, see [4, Chap. 8]. Therefore, the claim (6.24) follows from [5, Corollary 5.2].

The proof of (6.25) is the same as that of (6.6) in Theorem 6.1. This completes the proof of Theorem 6.2. □

Proof of Corollary 6.2 By assumption, the noise $\{\eta_j\}$ is a stationary ergodic process with $E[\eta_j^4] < \infty$. Therefore, by [9, Theorem 3.5.7 and 3.5.8], $\{X_t\}$ is a stationary ergodic process with $E[X_t^4] < \infty$. These theorems also imply that the processes

$$Z_t = \sum_{s=0}^{\infty}(s+1)\theta_0^s(X_{t-s} - \mu), \quad \eta_t = \sum_{s=0}^{\infty}\theta_0^s(X_{t-s} - \mu),$$

and $\{Z_{t-1}^2\eta_t^2\}$, $\{Z_t^2\}$ are stationary and ergodic. Since $|\theta_0| < 1$, then

$$E[Z_t^4] \le \Big(\sum_{s=0}^{\infty}(s+1)|\theta_0|^s\Big)^4 E[X_1^4] < \infty, \tag{6.29}$$

$$E[\eta_t^4] \le \Big(\sum_{s=0}^{\infty}|\theta_0|^s\Big)^4 E[X_1^4] < \infty, \quad E[Z_{t-1}^2\eta_t^2] < \infty.$$

Then, by [9, Theorem 3.5.7 and 3.5.8] again,

$$\frac{n^{-1}\sum_{t=2}^{n} Z_{t-1}^2\eta_t^2}{\big(n^{-1}\sum_{t=1}^{n} Z_t^2\big)^2} \to_p \frac{E[\eta_1^2 Z_0^2]}{(E[Z_0^2])^2} = v_\eta^2. \tag{6.30}$$

It suffices to show that

$$n^{-1}\sum_{t=2}^{n}(\widehat{Z}_{t-1}^2\widehat{\eta}_t^2 - Z_{t-1}^2\eta_t^2) = o_p(1), \tag{6.31}$$

$$n^{-1}\sum_{t=1}^{n}(\widehat{Z}_t^2 - Z_t^2) = o_p(1),$$

which together with (6.30) proves (6.27).

In view of the definition of $\widehat{Z}_t$ and Z_t, we can bound

$$|\widehat{Z}_t - Z_t| \le r_{t,1} + r_{t,2}, \quad r_{t,1} = \sum_{s=0}^{t-1}(s+1)|\widehat{\theta}^s - \theta_0^s|\,|X_{t-s} - \mu|,$$

$$r_{t,2} = \sum_{s=t}^{\infty}(s+1)|\theta_0^s|\,|X_{t-s} - \mu|.$$

Observe that $|\theta_0| < \delta < 1$ for some δ. By (6.24), $\widehat{\theta} - \theta_0 = O_p(n^{-1/2})$. Therefore, $P(|\widehat{\theta}| \geq \delta) = o(1)$. Hence, without loss of generality, we can assume that $|\widehat{\theta}| \leq \delta$. First we bound $r_{t,1}$. By the mean value theorem,

$$|\widehat{\theta}^s - \theta_0^s| \leq |\widehat{\theta} - \theta_0| s(|\widehat{\theta}|^{s-1} + |\theta_0|^{s-1}) \leq |\widehat{\theta} - \theta_0| 2s\delta^{s-1}, \ \ s \geq 1.$$

Therefore,

$$r_{t,1} \leq |\widehat{\theta} - \theta_0| h_{t,1}, \quad h_{t,1} = \sum_{s=1}^{\infty} (s+1) 2s\delta^{s-1} |X_{t-s} - \mu|.$$

On the other hand, $r_{t,2} \leq \delta^{t/2} h_{t,2}$ and

$$h_{t,2} = \sum_{s=0}^{\infty} (s+1)\delta^{s/2} |X_{t-s} - \mu|.$$

So,

$$|\widehat{Z}_t - Z_t| \leq |\widehat{\theta} - \theta_0| h_{t,1} + \delta^{t/2} h_{t,2}, \quad |\widehat{Z}_t| \leq h_{t,2},$$
$$|\widehat{\eta}_t - \eta_t| \leq |\widehat{\theta} - \theta_0| h_{t,1} + \delta^{t/2} h_{t,2}, \quad |\widehat{\eta}_t| \leq h_{t,2},$$

where stationary processes $\{h_{t,1}\}, \{h_{t,2}\}$ satisfy $E[h_{t,1}^4] < \infty$, $E[h_{t,2}^4] < \infty$. Together with (6.29), these facts imply (6.31). This completes the proof of corollary. □

6.3 Monte Carlo Experiments

In this section, we conduct simulations to verify the theoretical properties of the estimators of the AR(1) and MA(1) models, derived in [5] and in this paper. We use 95% coverage intervals for the parameters based on the estimated standard errors, to examine the potential impact of a non-i.i.d. noise on the coverage probabilities in finite samples, and to show that the use of the standard errors corresponding to i.i.d. noise may lead to significant coverage distortions.

We set the nominal coverage probability to 95% and conduct 1000 replications. We consider sample sizes $n = 300, 500$ and generate data using the following three uncorrelated stationary noises:

1. $\eta_t = \varepsilon_t \sim$ i.i.d. $\mathcal{N}(0, 1)$ – i.i.d. noise,
2. $\eta_t = \varepsilon_t \varepsilon_{t-1}$ – martingale difference noise,
3. $\eta_t = \sigma_t \varepsilon_t$, $\sigma_t^2 = 0.1 + 0.2\eta_{t-1}^2 + 0.7\sigma_{t-1}^2$ – GARCH(1, 1) noise.

They have properties $\mathrm{var}(\eta_t) = 1$ and $E[\eta_t^4] < \infty$. Only the first noise is i.i.d.

Simulation results for AR(1) model. We generate an array of samples $X_1, ..., X_n$ following the AR(1) model

$$X_t = \alpha_0 + \phi_0 X_{t-1} + \eta_t, \tag{6.32}$$

with parameters $\alpha_0 = 1$ and $\phi_0 = 0, 0.5, 0.8, -0.5, -0.8$ and $\{\eta_t\}$ as above. Define the t-statistics,

$$t_\phi = \frac{\widehat{\phi} - \phi_0}{\widehat{SE}_\phi}, \quad t_\alpha = \frac{\widehat{\alpha} - \alpha_0}{\widehat{SE}_\alpha},$$

computed using the standard errors $\widehat{SE}_\phi = \widehat{v}_\phi/\sqrt{n}$ and $\widehat{SE}_\alpha = \widehat{v}_\alpha/\sqrt{n}$ which permit for a non-i.i.d. noise. By Theorem 6.1 and Corollary 6.1, for the chosen set of the parameter values and all three noises, these statistics have property:

$$t_\phi \to_D \mathcal{N}(0, 1), \quad t_\alpha \to_D \mathcal{N}(0, 1).$$

Define also the t-statistics

$$t_{0,\phi} = \frac{\widehat{\phi} - \phi_0}{\widehat{SE}_{0,\phi}}, \quad t_{0,\alpha} = \frac{\widehat{\alpha} - \alpha_0}{\widehat{SE}_{0,\alpha}}.$$

Here, the standard errors, $\widehat{SE}_{0,\phi} = \widehat{v}_{0,\phi}/\sqrt{n}$ and $\widehat{SE}_{0,\alpha} = \widehat{v}_{0,\alpha}/\sqrt{n}$, are computed under the assumption that the noise is i.i.d., using (6.7),

$$\widehat{v}^2_{0,\phi} = 1 - \widehat{\phi}^2, \quad \widehat{v}^2_{0,\alpha} = \widehat{\alpha}^2 \frac{1 + \widehat{\phi}}{1 - \widehat{\phi}} + \widehat{\sigma}^2_\eta, \quad \widehat{\sigma}^2_\eta = (n-1)^{-1} \sum_{t=2}^{n} \widehat{\eta}^2_t.$$

These statistics have the standard normal limiting distribution when the noise is i.i.d.

In addition, we conduct the simulations to evaluate the standard deviations $\mathrm{sd}(\nu_\phi)$ and $\mathrm{sd}(\nu_\alpha)$ of the statistics

$$\nu_\phi = \sqrt{n}\big(\widehat{\phi} - \phi_0\big), \quad \nu_\alpha = \sqrt{n}\big(\widehat{\alpha} - \alpha_0\big).$$

We use three types of noise η_t all with variance $\sigma^2_\eta = 1$. The difference in values of sd for i.i.d. and non-i.i.d. noises will reveal the impact of the latent dependence in the noise on the standard deviation of the estimators of the parameters, shown in Theorem 6.1.

We recall from Theorem 6.1 that the asymptotic distribution of the sample mean $\bar{X}$ of $\mu = EX_t$ does not depend on the latent dependence structure of the noise. To confirm this theoretical finding in simulations, we define the t-statistic

$$t_\mu = \frac{\sqrt{n}\big(\bar{X} - \mu\big)}{\widehat{SE}_\mu}, \quad \widehat{SE}_\mu = \frac{\widehat{\sigma}_\eta}{1 - \widehat{\phi}}.$$

By (6.6), $t_\mu \to_D \mathcal{N}(0, 1)$ irrespective of whether or not the noise $\{\eta_j\}$ is i.i.d.

In Tables 6.1 and 6.2, we examine the accuracy of the theoretical results on the estimation of the parameters ϕ, α in the AR(1) model in finite samples. We report the standard deviations $\mathrm{sd}(\nu_\phi)$ and $\mathrm{sd}(\nu_\alpha)$ and the empirical coverage probabilities for 95% confidence intervals. Table 6.1 presents the coverage probabilities $\mathrm{CP}(t_\phi)$ and $\mathrm{CP}(t_{0,\phi})$ for confidence intervals for ϕ_0, constructed using the approximations $t_\phi \sim \mathcal{N}(0, 1)$ and $t_{0,\phi} \sim \mathcal{N}(0, 1)$. Table 6.2 reports the empirical coverage probabilities $\mathrm{CP}(t_\alpha)$ and $\mathrm{CP}(t_{0,\alpha})$ for confidence intervals for α_0, built using the approximations $t_\alpha \sim \mathcal{N}(0, 1)$ and $t_{0,\alpha} \sim \mathcal{N}(0, 1)$.

The Monte Carlo experiments confirm the theoretical findings of [5] on properties of the estimator $\widehat{\phi}$ presented in Sect. 6.2. In Table 6.1, for fixed value of ϕ_0, the standard deviation $\mathrm{sd}(\nu_\phi)$ varies depending on whether or not the noise $\{\eta_t\}$ is i.i.d. It shows that $\mathrm{sd}(\nu_\phi)$ takes different values for i.i.d. and non-i.i.d. noises. The empirical coverage probability $\mathrm{CP}(t_\phi)$ for 95% confidence interval for ϕ_0 that takes into account potential latent dependence in the noise, is close to 94% for sample size $n = 300$ and tends to reach the 95% level for $n = 500$. However, the coverage for the confidence intervals, built upon the assumption that the noise is i.i.d., is close to the nominal only for i.i.d. noise, and drops to 80%–90% when the noise is non i.i.d. The problem of the coverage distortion persists even when the sample size increases from 300 to 500. The coverage probability $\mathrm{CP}(t_{0,\phi})$ of the 95% confidence interval for ϕ_0 built upon the assumption that the noise is i.i.d. is close to $\mathrm{CP}(t_\phi)$ only for i.i.d. noise. The coverage may drop to 80%–90% when the noise is not i.i.d.

Table 6.2 exposes similar patterns of the empirical coverage for 95% confidence intervals for the AR(1) parameter α_0. The coverage $\mathrm{CP}(t_\alpha)$ for intervals that account for non-i.i.d. property of the noise is close to the nominal 95% for $n = 500$ and is in the range of 93%–95% for $n = 300$. If the noise is not i.i.d., the coverage distortion for confidence intervals for α_0 built upon the assumption that the noise is i.i.d. appears to be slightly milder than that for the dynamic parameter ϕ_0.

In Table 6.3, the empirical coverage probabilities for 95% confidence interval for the mean $\mu = EX_t$, built using the normal approximation $t_\mu \sim \mathcal{N}(0, 1)$ are close to the nominal 95% coverage irrespective of whether or not the noise is i.i.d. Given that, the standard errors and the long-run variance s_x^2 depend only on the variance of the noise which is set to 1, this result on coverage is expected.

Simulation results for MA(1) model. We perform a similar simulation on the coverage precision of the confidence intervals for the parameters in the MA(1) model:

$$X_t = \alpha_0 + \eta_t - \theta_0 \eta_{t-1}. \tag{6.33}$$

We generate 1000 samples $X_1, \ldots, X_n$ following the MA(1) model with the parameters $\alpha_0 = 1$ and $\theta_0 = 0, -0.5, -0.8, 0.5, 0.8$, and we consider three types of uncorrelated noise $\{\eta_t\}$ as above.

Similarly to the AR(1) model, we define the t-statistics,

Table 6.1 Estimation of the parameter ϕ_0 in the AR(1) model (6.32). Standard deviation sd(ν_ϕ), empirical coverage probabilities CP(t_ϕ) and CP($t_{0,\phi}$) (in %) of 95% confidence intervals

$\eta_t = \varepsilon_t$						
	$n = 300$			$n = 500$		
ϕ_0	sd(ν_ϕ)	CP(t_ϕ)	CP($t_{0,\phi}$)	sd(ν_ϕ)	CP(t_ϕ)	CP($t_{0,\phi}$)
0	0.97	95.5	96.2	0.97	95.2	95.7
0.5	0.86	94.4	94.9	0.85	95.1	95.3
0.8	0.63	93.7	94.0	0.61	95.0	95.2
-0.5	0.85	94.9	95.3	0.84	95.9	96.0
-0.8	0.60	94.2	94.5	0.59	95.5	95.9
$\eta_t = \varepsilon_t \varepsilon_{t-1}$						
	$n = 300$			$n = 500$		
ϕ_0	sd(ν_ϕ)	CP(t_ϕ)	CP($t_{0,\phi}$)	sd(ν_ϕ)	CP(t_ϕ)	CP($t_{0,\phi}$)
0	1.68	93.9	74.4	1.66	95.0	75.0
0.5	1.33	93.6	80.1	1.33	94.6	78.9
0.8	0.79	94.2	87.2	0.80	95.1	87.1
-0.5	1.35	93.4	79.3	1.33	94.9	79.4
-0.8	0.81	94.8	87.0	0.80	94.9	86.7
$\eta_t = \sigma_t \varepsilon_t$ GARCH						
	$n = 300$			$n = 500$		
ϕ_0	sd(ν_ϕ)	CP(t_ϕ)	CP($t_{0,\phi}$)	sd(ν_ϕ)	CP(t_ϕ)	CP($t_{0,\phi}$)
0	1.27	93.9	87.5	1.31	94.9	86.5
0.5	1.11	93.8	88.1	1.14	94.5	86.9
0.8	0.77	94.1	88.1	0.77	95.6	88.8
-0.5	1.10	93.8	87.9	1.10	94.8	87.2
-0.8	0.74	94.5	89.9	0.73	95.2	89.6

$$t_\theta = \frac{\widehat{\theta} - \theta_0}{\widehat{SE}_\theta}, \quad t_{0,\theta} = \frac{\widehat{\theta} - \theta_0}{\widehat{SE}_{0,\theta}}$$

computed using the standard errors $\widehat{SE}_\theta = \widehat{v}_\theta/\sqrt{n}$ and $\widehat{SE}_{0,\theta} = \widehat{v}_{0,\theta}/\sqrt{n}$ with $\widehat{v}_\theta$ as in (6.26) and $\widehat{v}^2_{0,\theta} = 1 - \widehat{\theta}^2$. The standard error $\widehat{SE}_\theta$ permits the presence of a non-i.i.d. noise in the model (6.33). By Theorem 6.2 and Corollary 6.2, the convergence $t_\theta \to_D \mathcal{N}(0, 1)$ holds for all values of θ_0 and all three noises $\{\eta_t\}$. The standard error $\widehat{SE}_{0,\theta}$ yields the convergence $t_{0,\theta} \to_D \mathcal{N}(0, 1)$ when the noise $\{\eta_t\}$ is i.i.d.

Table 6.2 Estimation of the parameter α_0 in the AR(1) model (6.32). Standard deviation sd(ν_α), empirical coverage probabilities CP(t_α) and CP($t_{0,\alpha}$) (in %) of 95% confidence intervals

$\eta_t = \varepsilon_t$						
	$n = 300$			$n = 500$		
ϕ_0	sd(ν_α)	CP(t_α)	CP($t_{0,\alpha}$)	sd(ν_α)	CP(t_α)	CP($t_{0,\alpha}$)
0	1.44	94.8	95.1	1.40	95.2	95.1
0.5	2.05	93.8	94.0	1.98	94.9	95.3
0.8	3.34	93.6	94.4	3.21	95.0	95.3
–0.5	1.18	93.9	94.1	1.14	94.9	94.5
–0.8	1.06	95.4	95.9	1.04	95.0	94.9
$\eta_t = \varepsilon_t \varepsilon_{t-1}$						
	$n = 300$			$n = 500$		
ϕ_0	sd(ν_α)	CP(t_α)	CP($t_{0,\alpha}$)	sd(ν_α)	CP(t_α)	CP($t_{0,\alpha}$)
0	1.94	94.3	84.3	1.93	95.3	84.7
0.5	2.83	94.0	82.7	2.83	94.7	83.9
0.8	4.08	93.8	87.8	4.08	94.8	86.8
–0.5	1.32	93.9	90.7	1.31	94.8	92.4
–0.8	1.06	95.4	95.4	1.07	95.8	95.2
$\eta_t = \sigma_t \varepsilon_t$ GARCH						
	$n = 300$			$n = 500$		
ϕ_0	sd(ν_α)	CP(t_α)	CP($t_{0,\alpha}$)	sd(ν_α)	CP(t_α)	CP($t_{0,\alpha}$)
0	1.68	93.2	89.8	1.68	94.8	89.7
0.5	2.51	93.1	87.4	2.53	94.4	87.1
0.8	4.04	93.8	88.1	4.02	94.9	89.3
–0.5	1.27	93.6	91.9	1.26	94.4	92.4
–0.8	1.10	94.5	94.3	1.08	94.8	93.8

Table 6.3 Estimation of the mean $\mu = EX_t$ in the AR(1) model (6.32). Empirical coverage probabilities CP(t_μ) (in %) of 95% confidence intervals

	$\eta_t = \varepsilon_t$		$\eta_t = \varepsilon_t \varepsilon_{t-1}$		$\eta_t = \sigma_t \varepsilon_t$ GARCH	
ϕ_0	$n = 300$	500	300	500	300	500
0	94.7	94.8	96.6	95.9	94.0	94.7
0.5	94.2	94.7	95.7	95.7	93.8	94.7
0.8	93.5	94.1	94.5	95.2	93.1	94.9
–0.5	95.1	94.9	96.6	95.7	94.4	94.7
–0.8	95.2	94.6	95.9	95.6	94.6	95.0

We also evaluate the standard deviation $\mathrm{sd}(\nu_\phi)$ of the statistic $\nu_\theta = \sqrt{n}\big(\widehat{\theta} - \theta_0\big)$ to confirm the theoretical finding that the dependence in the noise $\{\eta_t\}$ has impact on the variance of the estimator $\widehat{\theta}$, established in Theorem 6.2.

Recall from Theorem 6.2 that the asymptotic variance s_x^2 of the sample mean $\bar{X}$ of $\mu = EX_t$ depends only on the variance σ_η^2 of the noise, and the t-statistic

$$t_\mu = \frac{\sqrt{n}\big(\bar{X} - \mu\big)}{\widehat{SE}_\mu}, \quad \widehat{SE}_\mu = \widehat{\sigma}_\eta(1 - \widehat{\theta})$$

has property $t_\mu \to_D \mathcal{N}(0, 1)$. We use it to build confidence intervals for the mean EX_t and to show that they are not affected by the latent dependence in the noise.

Table 6.4 reveals similar patterns of the empirical coverage for the MA(1) parameter θ_0 as for the AR(1) parameter ϕ_0 above. The empirical coverage $\mathrm{CP}(t_\theta)$ of the confidence intervals built to accommodate the potential dependence in the noise $\{\eta_t\}$ is close to the nominal 95%, and the coverage improves when the sample size increases. Some coverage distortions are observed for the parameter $\theta_0 = 0.8$ disappear when data is centered by the true mean (simulations not presented here).

As seen from $\mathrm{CP}(t_{0,\theta})$, the coverage distortions for the confidence intervals based on the assumption that the noise is i.i.d. range from mild to severe. Simulations confirm that such intervals do not produce adequate coverage $\mathrm{CP}(t_{0,\theta})$, unless the noise is i.i.d. Table 6.4 also shows that the standard deviation $\mathrm{sd}(\nu_\theta)$ varies depending on whether the noise is i.i.d. or not, which affirms the theoretical findings of Theorem 6.2. Table 6.5 confirms that the confidence intervals for the mean $E[X_t]$ are not affected by the latent dependence on the noise.

In summary, the simulation experiments clearly demonstrate some standardization problems in the estimation of the AR(1) and MA(1) models and overall in the Whittle estimation of linear processes generated by a noise which is not i.i.d. Parameter estimation and building of confidence intervals need corrected standardization. For the AR(1) and MA(1) models, the corrected standard errors and asymptotic variances can be estimated. Simulations show that the empirical coverage for confidence intervals with the estimated standard errors is close to the nominal which confirms the theoretical findings of [5]. The problem of the estimation of the standard errors and asymptotic variances of the Whittle estimators for the AR(p), $p \geq 2$ and other ARMA models remains open.

Acknowledgements Murad Taqqu research was supported in part by the SIMONS grant 569118.

Table 6.4 Estimation of the parameter θ_0 in the MA(1) model (6.33). Standard deviation sd(ν_θ), empirical coverage probabilities CP(t_θ) and CP($t_{0,\theta}$) (in %) of 95% confidence intervals

$\eta_t = \varepsilon_t$						
	$n = 300$			$n = 500$		
θ_0	sd(ν_θ)	CP(t_θ)	CP($t_{0,\theta}$)	sd(ν_θ)	CP(t_θ)	CP($t_{0,\theta}$)
0	1.01	94.6	94.6	0.99	94.9	95.0
–0.5	0.89	93.8	94.5	0.85	95.6	95.3
–0.8	0.63	94.8	94.7	0.60	95.4	95.3
0.5	0.88	92.4	92.5	0.87	94.1	94.0
0.8	0.67	91.6	91.2	0.67	92.2	92.6
$\eta_t = \varepsilon_t \varepsilon_{t-1}$						
	$n = 300$			$n = 500$		
θ_0	sd(ν_θ)	CP(t_θ)	CP($t_{0,\theta}$)	sd(ν_θ)	CP(t_θ)	CP($t_{0,\theta}$)
0	1.70	91.5	73.5	1.67	93.9	74.5
–0.5	1.40	93.0	77.2	1.37	93.9	77.6
–0.8	0.93	90.8	82.9	0.88	93.3	83.2
0.5	1.36	92.3	79.5	1.35	93.4	79.8
0.8	0.91	89.8	81.8	0.85	91.9	83.8
$\eta_t = \sigma_t \varepsilon_t$ GARCH						
	$n = 300$			$n = 500$		
θ_0	sd(ν_θ)	CP(t_θ)	CP($t_{0,\theta}$)	sd(ν_θ)	CP(t_θ)	CP($t_{0,\theta}$)
0	1.33	92.0	86.7	1.33	93.8	86.3
–0.5	1.21	92.3	87.2	1.11	94.0	87.0
–0.8	0.76	94.5	88.6	0.75	95.1	89.3
0.5	1.13	93.0	85.3	1.15	93.6	85.1
0.8	0.81	90.1	85.4	0.81	93.2	86.9

Table 6.5 Estimation of the mean $\mu = EX_t$ in the MA(1) model (6.33). Empirical coverage probabilities CP(t_μ) (in %) of 95% confidence intervals

	$\eta_t = \varepsilon_t$		$\eta_t = \varepsilon_t \varepsilon_{t-1}$		$\eta_t = \sigma_t \varepsilon_t$ GARCH	
θ_0	$n = 300$	500	300	500	300	500
0	94.5	94.8	96.4	95.6	94.1	94.6
–0.5	95.1	94.8	96.5	95.4	94.0	94.5
–0.8	95.3	94.9	96.1	95.6	94.2	94.5
0.5	94.1	94.6	95.5	95.8	93.3	94.1
0.8	92.6	93.4	93.7	95.2	92.2	93.8

References

1. Abadir, K. M., Distaso, W., Giraitis, L. and Koul, H. L. (2014). Asymptotic normality for weighted sums of linear processes. *Econometric Theory* **30** 252–284.
2. Box, G. E. P. and Jenkins, G. M. (1970). *T*ime Series Analysis: Forecasting and Control. Holden-Day, San Francisco.
3. Fox, R. and Taqqu, M. S. (1986). Large-sample properties of parameter estimates for strongly dependent stationary Gaussian time series. *Annals of Statistics* **14** 517–532.
4. Giraitis, L., Koul, H. L. and Surgailis, D. (2012). *L*arge Sample Inference for Long Memory Processes. Imperial College Press, London.
5. Giraitis, L., Taniguchi, M. and Taqqu, M. S. (2018). Estimation pitfalls when the noise is not i.i.d. *Japanese Journal of Statistics and Data Science* **1** 59–80.
6. Granger, C. W. J. and Joyeux, R. (1980). An introduction to long-memory time series models and fractional differencing. *Journal of Time Series Analysis* **1** 15–29.
7. Hosking, J. R. M. (1981). Fractional differencing. *Biometrika* **68** 165–176.
8. Hosoya, Y. (1997). A limit theory for long-range dependence and statistical inference on related models. *Annals of Statistics* **25** 105–137.
9. Stout, W. F. (1974). *A*lmost Sure Convergence. Academic Press, New York.
10. Whittle, P. (1953). Estimation and information in stationary time series. *Arkiv főr Matematik* **2** 423–434.

Chapter 7
Tests for a Structural Break for Nonnegative Integer-Valued Time Series

Yuichi Goto

Abstract We investigate tests for a structural break for nonnegative integer-valued time series. This topic has been intensively studied in recent years. We deal with the model whose conditional expectation is endowed with dependence structures. Unknown parameters of the model are estimated by an M-estimator. Then, we study three types of test statistics: the Wald type, score type, and residual type. First, we show the asymptotic null distributions of these three test statistics, which enable us to construct asymptotically size α tests. Next, we show the consistency of the tests, that is, the power of the tests converges to one as sample size increases. Finally, numerical study illustrates the finite-sample performance of the tests.

7.1 Introduction

Nonnegative integer-valued time series have attracted great attention and appeared in a variety of fields, for example, the number of corporate defaults [1], transactions of stocks [14], earthquakes [34], and the infected people [13, 31–33].

Sufficient conditions for stationarity, ergodicity, and β-mixing for nonlinear Poisson integer-valued generalized autoregressive conditional heteroskedastic models with orders 1 and 1 (INGARCH(1,1)) were given by [26]. The existence of stationary and ergodic solutions of general nonlinear Poisson AR models with any finite moment under a contractive condition was clarified by [12]. Nonlinear Poisson INGARCH(p,q) models with exogenous variables were dealt by [1]. A one-parameter exponential family for nonnegative integer-valued time series, which includes the Poisson, negative binomial (NB), and Bernoulli distributions was introduced by [8] together with fundamental properties of the models.

Tests for structural breaks have been studied from the 1950s. A pioneer study of the detection of structural breaks was given by [30]. A cumulative sum (CUSUM) test for time series models was proposed by [21]. For nonnegative integer-valued time series, the residual-based CUSUM test using a conditional least square esti-

---

Y. Goto (✉)
Kyushu University, 744 Motooka, Nishi-ku, Fukuoka 819-0395, Japan
e-mail: yuichi.goto@math.kyushu-u.ac.jp

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_7

mator for non-linear Poisson AR(1) models was proposed by [15]. The Wald type and residual type of the tests using a conditional maximum likelihood estimator were constructed for nonlinear Poisson INGARCH(1,1) models [19]. Change tests for nonlinear zero-inflated generalized Poisson INGARCH(1,1) models and linear bivariate Poisson INGARCH(1,1) models were studied by [22, 24], respectively. Several tests including the Wald type, score type, and residual type of CUSUM tests were compared by [23]. A change test for general nonlinear Poisson AR models was developed by [11]. The paper also investigated asymptotics under the null and alternative. A change detection problem for one-parameter exponential INGARCH(1,1) models was studied by [9]. The paper also constructed a consistent test. A test for a structural break using probability generating functions for the first-order integer-valued autoregressive (INAR(1)) models was advocated by [17]. The estimator of a change point for an one-parameter exponential family as well as its asymptotic behavior was investigated by [7].

However, in practice, it is unrealistic to assume the knowledge of the underlying conditional distribution. To the best of my knowledge, a change detection problem without making assumptions on the conditional distributions for nonnegative integer-valued time series has not yet been investigated except for [10]. Few studies have examined the consistency of tests. The exceptions were [10, 11]. Moreover, although several change test procedures for INGARCH(1,1) models and other simple models were well studied, those procedures for higher order models have not yet been discussed except for [11].

On the other hand, the assumptions on conditional distributions were relaxed, and Poisson quasi-maximum likelihood estimators (QMLE) were proposed for general nonlinear AR models in [2]. They showed strong consistency and asymptotic normality (CAN) of Poisson QMLE. In addition, negative binomial QMLEs and exponential QMLEs were proposed and CAN was shown in [3, 4]. Sufficient conditions for count time series to have the properties of strictly stationarity, ergodicity, and β-mixing were elucidated by [3]. The essential condition is called *a stochastic-equal-mean order property* many models satisfy.

In this paper, we tackle a change detection problem using the M-estimator proposed by [16] for general models. We emphasize our model includes the nonlinear INGARCH(p,q) models and need not specify a conditional distribution. We investigate the Wald type, score type, and residual type of CUSUM tests and derive asymptotic null distributions. As a result, we obtain asymptotically size α tests. Moreover, we establish the consistency of these tests, which is our main result.

The rest of the paper is organized as follows. In Sect. 7.2, we introduce nonnegative integer-valued time series with a conditional expectation endowed with dependence structures. Next, we review the M-estimator, which includes the Poisson QMLE as a special case, and make some assumptions for CAN of the M-estimator. In Sect. 7.3, we formulate a test for a structural break and define three test statistics; the Wald type, score type, and residual type. We derive asymptotic distributions of these statistics under the null hypothesis. These results enable us to construct asymptotically size α tests. Next, we show the consistency of the tests. We illustrate the finite-sample performance in Sect. 7.4 and reveal that the classical Wald type test is sensitive to the misspecification of conditional distributions and has size distortion because of the instability of the estimator for small samples. In contrast, the score type and residual type tests provide good empirical size. Our simulation study suggests the residual type test is superior to the score type test in terms of the power. All proofs of theorems are provided in Sect. 7.5.

In what follows, the following notations will be used. The symbol $\top$ is the transpose of vector or matrix. For $\boldsymbol{x} = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$, $\|\boldsymbol{x}\| := \sum_{i=1}^{d} |x_i|$ and $\|\boldsymbol{x}\|_2 := \left(\sum_{i=1}^{d} x_i^2\right)^{1/2}$, and for a matrix $\boldsymbol{A} = (a_{ij})_{i,j=1,\ldots,d}$, $\|\boldsymbol{A}\| := \sum_{i,j=1}^{d} |a_{ij}|$. For a sequence of random variables $\{X_n\}$ and a random variable X, let $X_n \to X$ in probability, $X_n \to X$ a.s., and $X_n \Rightarrow X$ denote X_n converges in probability to X, X_n converges to X a.s., and X_n converges in distribution to X, respectively.

7.2 Settings

In this section, we introduce general models which encompass popular models such as INGARCH models and INAR models for count data and review the M-estimator introduced by [16].

Let $\{Z_t\}$ be a nonnegative integer-valued time series whose conditional expectation is given by, for any $t \in \mathbb{Z}$,

$$\mathrm{E}\left(Z_t|\mathcal{F}_{t-1}\right) := \lambda(Z_{t-1}, Z_{t-2}, \ldots, ; \boldsymbol{\theta}_0),$$

where $\mathcal{F}_{t-1}$ is the σ-field generated by $\{Z_s, s \le t-1\}$, $\boldsymbol{\theta}_0$ is an unknown parameter in a parameter space $\boldsymbol{\Theta} \subset \mathbb{R}^d$, and λ is a known measurable function on $[0, \infty)^\infty \times \boldsymbol{\Theta}$ to (δ, ∞) for some $\delta > 0$.

For the sake of notational simplicity, we define, for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $t \in \mathbb{N} \cup \{0\}$,

$$\lambda_t(\boldsymbol{\theta}) := \lambda(Z_{t-1}, Z_{t-2}, \ldots, ; \boldsymbol{\theta}) \text{ and } \tilde{\lambda}_t(\boldsymbol{\theta}) := \lambda(Z_{t-1}, Z_{t-2}, \ldots, Z_1, \mathbf{x}_0; \boldsymbol{\theta}),$$

where $\mathbf{x}_0 \in [0, \infty)^\infty$ is an initial value. The function $\tilde{\lambda}_t$, which can be calculated from observations, is a proxy for λ_t, which includes unobservable values. Since we use specific models like the linear INGARCH(p,q) model as λ in practice, $\mathbf{x}_0 \in [0, \infty)^\infty$ reduces to a finite dimensional vector. Some examples of an initial value $\mathbf{x}_0$ were given by [2, 11]. Note that the impact of a choice $\mathbf{x}_0 \in [0, \infty)^\infty$ is asymptotically negligible as $n \to \infty$ by the assumptions we impose afterward. Instead of making assumptions on conditional distributions, we assume stationarity and ergodicity of $\{Z_t\}$.

Assumption 7.1 $\{Z_t\}$ is ergodic and strictly stationary.

Remark 7.1 This assumption is satisfied by a broad class of the time series of counts. Sufficient conditions of Assumption 1 for the non-linear INGARCH(p,q) models were given by [3], including the exponential family and the zero-inflated distributions.

The strictly stationarity and ergodicity of multivariate generalized INAR models were shown by [20].

The M-estimator for nonnegative integer-valued time series, proposed by [16], is defined as follows:

$$\hat{\boldsymbol{\theta}}_n := \arg\max_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \tilde{L}_n(\boldsymbol{\theta}), \quad \tilde{L}_n(\boldsymbol{\theta}) := \frac{1}{n}\sum_{t=1}^{n} \ell(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta})),$$

where $\ell(\cdot,\cdot)$ is a measurable function. The M-estimator reduces to Poisson QMLE or conditional least square estimator if we choose $\ell(Z_t, \lambda_t(\boldsymbol{\theta}))$ as $-\lambda_t(\boldsymbol{\theta}) + Z_t \log \lambda_t(\boldsymbol{\theta})$ or $-(Z_t - \lambda_t(\boldsymbol{\theta}))^2$, respectively. Let $\ell'(\cdot,\cdot)$ and $\ell''(\cdot,\cdot)$ denote the first and second derivatives of $\ell(\cdot,\cdot)$ with respect to the second component, respectively.

We make the following assumptions to ensure the CAN of the M-estimator.

Assumption 7.2 **(B1)** The functions $\ell(\cdot,\cdot)$ with respect to its second component and $\lambda_t(\boldsymbol{\theta})$ are continuous a.s.

(B2) The quantity $\left|\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta})) - \sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \ell(Z_t, \lambda_t(\boldsymbol{\theta}))\right|$ converges to 0 a.s., as $t \to \infty$.

(B3) There uniquely exists $\boldsymbol{\theta}_0 \in \mathring{\boldsymbol{\Theta}}$ such that $\mathrm{E}\ell(Z_t, \lambda_t(\boldsymbol{\theta})) < \mathrm{E}\ell(Z_t, \lambda_t(\boldsymbol{\theta}_0)) < \infty$ for any $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, where $\mathring{\boldsymbol{\Theta}}$ denotes the interior of $\boldsymbol{\Theta}$.

(B4) The parameter space $\boldsymbol{\Theta}$ is compact.

(C5) The functions $\ell(\cdot,\cdot)$ with respect to its second component and $\lambda_t(\boldsymbol{\theta})$ are twice continuously differentiable.

(C6) For some $\delta > 0$,

$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\| \partial\lambda_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta} \left(\ell'(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta})) - \ell'(Z_t, \lambda_t(\boldsymbol{\theta}))\right)\right\|,$$
$$\sup_{\boldsymbol{\theta}\in\boldsymbol{\Theta}} \left\| \ell'(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta})) \left(\partial\tilde{\lambda}_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta} - \partial\lambda_t(\boldsymbol{\theta})/\partial\boldsymbol{\theta}\right)\right\|$$

are of order $O(t^{-1/2-\delta})$ a.s., as $n \to \infty$.

(C7) There exist a neighborhood $\mathcal{V}(\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}_0$ and some $\delta > 0$ such that

$$\sum_{i,j=1}^{d} \mathrm{E}\left(\left|\partial\ell(Z_t, \lambda_t(\boldsymbol{\theta}_0))/\partial\theta_i \partial\ell(Z_t, \lambda_t(\boldsymbol{\theta}_0))/\partial\theta_j\right|^{1+\delta}\right) < \infty,$$
$$\sum_{i,j=1}^{d} \mathrm{E}\left(\sup_{\boldsymbol{\theta}\in\mathcal{V}(\boldsymbol{\theta}_0)} \left|\partial^2\ell(Z_t, \lambda_t(\boldsymbol{\theta}))/(\partial\theta_i\partial\theta_j)\right|\right) < \infty.$$

(C8) The sequence $\{\ell'(Z_t, \lambda_t(\boldsymbol{\theta}_0))\}_{t\in\mathbb{Z}}$ is martingale difference relative to $\{\mathcal{F}_t\}_{t\in\mathbb{Z}}$.

(C9) The quantity $\mathrm{E}\left(\ell''(Z_t, \lambda_t(\boldsymbol{\theta}_0))|\mathcal{F}_{t-1}\right)$ is not equal to zero a.s. for any $t \in \mathbb{Z}$, and if $\mathbf{s}^\top \partial\lambda_t(\boldsymbol{\theta}_0)/\partial\boldsymbol{\theta} = 0$, then $\mathbf{s} = \mathbf{0}$.

Remark 7.2 The asymptotic normality of $\hat{\boldsymbol{\theta}}_n$ is ensured by (C8), and (C9) guarantees J is a positive definite matrix.

Lemma 7.1 *Under Assumptions 7.1 and 7.2 (B1)–(B4),*

$$\hat{\boldsymbol{\theta}}_n \to \boldsymbol{\theta}_0 \quad a.s., \text{ as } n \to \infty.$$

Lemma 7.2 *Under Assumptions 7.1 and 7.2,*

$$\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\right) \Rightarrow N(0, J^{-1} I J^{-1}) \quad as\ n \to \infty,$$

where

$$\begin{aligned} I &:= \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\theta}} \ell(Z_t, \lambda_t(\boldsymbol{\theta}_0)) \frac{\partial}{\partial \boldsymbol{\theta}^\top} \ell(Z_t, \lambda_t(\boldsymbol{\theta}_0))\right) \\ &= \mathrm{E}\left(\left(\ell'(Z_t, \lambda_t(\boldsymbol{\theta}_0))\right)^2 \frac{\partial}{\partial \boldsymbol{\theta}} \lambda_t(\boldsymbol{\theta}_0) \frac{\partial}{\partial \boldsymbol{\theta}^\top} \lambda_t(\boldsymbol{\theta}_0)\right), \\ J &:= -\mathrm{E}\left(\frac{\partial^2}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \ell(Z_t, \lambda_t(\boldsymbol{\theta}_0))\right) \\ &= -\mathrm{E}\left(\ell''(Z_t, \lambda_t(\boldsymbol{\theta}_0)) \frac{\partial}{\partial \boldsymbol{\theta}} \lambda_t(\boldsymbol{\theta}_0) \frac{\partial}{\partial \boldsymbol{\theta}^\top} \lambda_t(\boldsymbol{\theta}_0)\right). \end{aligned}$$

7.3 Detection of a Structural Break

In this section, we consider tests for structural breaks, which is the main contribution of this paper. The null hypothesis H_0 is that there is no change in the true parameter, and the alternative H_1 is that the true parameter changes after an unknown point $\lfloor n\tau \rfloor$, where $\tau \in (0, 1)$. This can be formulated as follows: The null hypothesis and the alternative hypothesis are defined, for some $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$ and $\boldsymbol{\theta}_1 (\neq \boldsymbol{\theta}_0) \in \boldsymbol{\Theta}$, as

$$H_0: \ \mathrm{E}\left(Z_t | \mathcal{F}_{t-1}\right) := \lambda(Z_{t-1}, Z_{t-2}, \ldots; \boldsymbol{\theta}_0) \quad \text{for any } t \in \{1, \ldots, n\}$$

and

$$\begin{aligned} H_1: \mathrm{E}\left(Z_t | \mathcal{F}_{t-1}\right) &:= \lambda(Z_{t-1}, Z_{t-2}, \ldots; \boldsymbol{\theta}_0) \quad \text{for any } t \in \{1, \ldots, \lfloor n\tau \rfloor\}, \\ \mathrm{E}\left(Z'_t | \mathcal{F}'_{t-1}\right) &:= \lambda(Z'_{t-1}, Z'_{t-2}, \ldots; \boldsymbol{\theta}_1) \quad \text{for any } t \in \{\lfloor n\tau \rfloor + 1, \ldots, n\}, \end{aligned}$$

where $\mathcal{F}_{t-1}$ and $\mathcal{F}'_{t-1}$ are the σ-field generated by $\{Z_s, s \leq t-1\}$ and $\{Z'_s, s \leq t-1\}$, respectively. Let v_n be a nonnegative integer-valued sequence such that $v_n \to \infty$ and $v_n/n \to 0$ as $n \to \infty$.

To investigate tests for structural break, we introduce the following three test statistics. The Wald type test was investigated by [19, 22, 24]:

$$T_{n,\text{Wald}} := \max_{v_n \le k \le n} \frac{k^2}{n} (\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_n)^\top \hat{J}_n \hat{I}_n^{-1} \hat{J}_n (\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_n),$$

where

$$\hat{I}_n := \frac{1}{n} \sum_{t=1}^{n} \tilde{I}_t(\hat{\boldsymbol{\theta}}_n), \quad \hat{J}_n := \frac{1}{n} \sum_{t=1}^{n} \tilde{J}_t(\hat{\boldsymbol{\theta}}_n),$$

$$\tilde{I}_t(\boldsymbol{\theta}) := \frac{\partial \ell(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}} \frac{\partial \ell(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta}^\top}, \text{ and } \tilde{J}_t(\boldsymbol{\theta}) := \frac{\partial^2 \ell(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta}))}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top}.$$

The score-type test was studied by [5, 23, 28]:

$$T_{n,\text{score}} := \max_{v_n \le k \le n} \frac{1}{n} \left(\sum_{t=1}^{k} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(Z_t, \tilde{\lambda}_t(\hat{\boldsymbol{\theta}}_n)) \right)^\top \hat{I}_n^{-1} \left(\sum_{t=1}^{k} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(Z_t, \tilde{\lambda}_t(\hat{\boldsymbol{\theta}}_n)) \right).$$

The residual-type test was discussed by [15, 19]:

$$T_{n,\text{res}} := \max_{v_n \le k \le n} \frac{1}{\sqrt{\frac{1}{n} \sum_{t=1}^{n} \tilde{\epsilon}_t^2(\hat{\boldsymbol{\theta}}_n)}} \frac{1}{\sqrt{n}} \left| \sum_{t=1}^{k} \tilde{\epsilon}_t(\hat{\boldsymbol{\theta}}_n) - \frac{k}{n} \sum_{t=1}^{n} \tilde{\epsilon}_t(\hat{\boldsymbol{\theta}}_n) \right|,$$

where $\tilde{\epsilon}_t(\hat{\boldsymbol{\theta}}_n) := Z_t - \tilde{\lambda}_t(\hat{\boldsymbol{\theta}}_n)$.

The following assumptions are required to establish theoretical results.

Assumption 7.3 **(D10)** The following conditions hold true:

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\| \left(\ell'(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta})) \right)^2 \left(\partial \tilde{\lambda}_t(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \tilde{\lambda}_t(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^\top - \partial \lambda_t(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \lambda_t(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^\top \right) \right\|,$$

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\| \partial \lambda_t(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \lambda_t(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^\top \left(\left(\ell'(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta})) \right)^2 - \left(\ell'(Z_t, \lambda_t(\boldsymbol{\theta})) \right)^2 \right) \right\|,$$

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\| \ell''(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta})) \left(\partial \tilde{\lambda}_t(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \tilde{\lambda}_t(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^\top - \partial \lambda_t(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \lambda_t(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^\top \right) \right\|,$$

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\| \left(\ell''(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta})) - \ell''(Z_t, \lambda_t(\boldsymbol{\theta})) \right) \partial \lambda_t(\boldsymbol{\theta}) / \partial \boldsymbol{\theta} \partial \lambda_t(\boldsymbol{\theta}) / \partial \boldsymbol{\theta}^\top \right\|,$$

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\| \ell'(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta})) \left(\partial^2 \tilde{\lambda}_t(\boldsymbol{\theta}) / (\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top) - \partial \lambda_t(\boldsymbol{\theta}) / (\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top) \right) \right\|,$$

and

$$\sup_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} \left\| \left(\ell'(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta})) - \ell'(Z_t, \lambda_t(\boldsymbol{\theta})) \right) \partial \lambda_t(\boldsymbol{\theta}) / (\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top) \right\|$$

are of order $O(t^{-1/2-\delta})$ a.s., as $n \to \infty$, and there exist a neighborhood $\mathcal{V}(\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}_0$ and some $\delta > 0$ such that

$$\sum_{i,j=1}^{d} \mathrm{E} \sup_{\boldsymbol{\theta}\in\mathcal{V}(\boldsymbol{\theta}_0)} \left|\partial\ell(Z_t,\lambda_t(\boldsymbol{\theta}))/\partial\theta_i \partial\ell(Z_t,\lambda_t(\boldsymbol{\theta}))/\partial\theta_j\right| < \infty.$$

(D11) There exist a neighborhood $\mathcal{V}(\boldsymbol{\theta}_0)$ of $\boldsymbol{\theta}_0$ and some $\delta > 0$ such that

$$\sum_{i,j=1}^{d} \mathrm{E}\left(\sup_{\boldsymbol{\theta}\in\mathcal{V}(\boldsymbol{\theta}_0)} |\partial\lambda_t(\boldsymbol{\theta})/\partial\theta_i|^2 \right) < \infty, \quad \mathrm{E}\,|Z_t - \lambda_t(\boldsymbol{\theta}_0)|^2 < \infty,$$

and, as $n \to \infty$,

$$\sup_{\boldsymbol{\theta}\in\Theta} \left|\tilde{\lambda}_t(\boldsymbol{\theta}) - \lambda_t(\boldsymbol{\theta})\right| = O(t^{-1/2-\delta}) \quad \text{a.s.}$$

(D12) The matrix I is positive definite.

Remark 7.3 In conjunction with Assumptions 7.1 and 7.2, (D10) is sufficient to ensure that $\hat{I}_n$ and $\hat{J}_n$ converge to I and J a.s., as $n \to \infty$, respectively. Also, (D11) guarantees a convergence of $T_{n,\mathrm{res}}$ under H_0.

We are ready to state the asymptotic null distributions.

Theorem 7.1 *Suppose Assumptions 7.1–7.3. Then, under the null hypothesis H_0, we have, as $n \to \infty$,*

$$T_{n,\mathrm{Wald}},\ T_{n,\mathrm{score}} \Rightarrow \sup_{0\le s\le 1} \|B_d^\circ(s)\|_2^2, \quad T_{n,\mathrm{res}} \Rightarrow \sup_{0\le s\le 1} |B_1^\circ(s)|^2,$$

where $B_d^\circ(s)$ is a d-dimensional standard Brownian bridge.

From Theorem 7.1, we can construct the asymptotically distribution-free tests, which reject the null hypothesis H_0 when $T_{n,\mathrm{Wald}} > C$, $T_{n,\mathrm{score}} > C$, and $T_{n,\mathrm{res}} > C'$, respectively, where C and C' are critical values satisfying, for significance level α, $\mathrm{P}\left(\sup_{0\le s\le 1} \|B_d^\circ(s)\|_2^2 > C\right) = \alpha$ and $\mathrm{P}\left(\sup_{0\le s\le 1} |B_1^\circ(s)|_2^2 > C'\right) = \alpha$. Then, these tests have asymptotic size α, respectively.

Next, the power of the tests under the alternative H_1 is scrutinized. For notational simplicity, let $\lambda_t'(\boldsymbol{\theta}){:=}\lambda(Z_{t-1}', Z_{t-2}', \ldots; \boldsymbol{\theta})$. From the compactness of the parameter space (B4), the extreme value theorem in calculus ensures the existence of one or more maximizers of $\tau\mathrm{E}\ell(Z_t, \lambda_t(\boldsymbol{\theta})) + (1-\tau)\mathrm{E}\ell(Z_t', \lambda_t'(\boldsymbol{\theta}))$.

The following lemma can be shown along the lines with [16, Theorem 1].

Lemma 7.3 *Suppose that $\tau\mathrm{E}\ell(Z_t, \lambda_t(\boldsymbol{\theta})) + (1-\tau)\mathrm{E}\ell(Z_t', \lambda_t'(\boldsymbol{\theta}))$ has a unique maximum at $\bar{\boldsymbol{\theta}}$. Under Assumptions 7.1 and 7.2 (B1)–(B4) and those corresponding assumptions for Z_t' and λ_t',*

$$\hat{\boldsymbol{\theta}}_n \to \bar{\boldsymbol{\theta}} \quad a.s.$$

Remark 7.4 The uniqueness of maximizers for $\tau \mathrm{E}\ell(Z_t, \lambda_t(\boldsymbol{\theta})) + (1-\tau)\mathrm{E}\,\ell(Z_t', \lambda_t'(\boldsymbol{\theta}))$ can be relaxed to a finite number of minimizers. In that case, we can show that $\hat{\boldsymbol{\theta}}_n$ belongs to the set of maximizers a.s., as $n \to \infty$.

To prove the consistency of the tests, we define the following quantities:

$$J^* := \tau J(\bar{\boldsymbol{\theta}}) + (1-\tau)J'(\bar{\boldsymbol{\theta}}), \quad I^* := \tau I(\bar{\boldsymbol{\theta}}) + (1-\tau)I'(\bar{\boldsymbol{\theta}}),$$

where

$$I(\boldsymbol{\theta}) := \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\theta}}\ell(Z_t,\lambda_t(\boldsymbol{\theta}))\frac{\partial}{\partial \boldsymbol{\theta}^\top}\ell(Z_t,\lambda_t(\boldsymbol{\theta}))\right), \quad J(\boldsymbol{\theta}) := -\mathrm{E}\left(\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}\ell(Z_t,\lambda_t(\boldsymbol{\theta}))\right),$$

$$I'(\boldsymbol{\theta}) := \mathrm{E}\left(\frac{\partial}{\partial \boldsymbol{\theta}}\ell(Z_t',\lambda_t'(\boldsymbol{\theta}))\frac{\partial}{\partial \boldsymbol{\theta}^\top}\ell(Z_t',\lambda_t'(\boldsymbol{\theta}))\right),$$
$$J'(\boldsymbol{\theta}) := -\mathrm{E}\left(\frac{\partial^2}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top}\ell(Z_t',\lambda_t'(\boldsymbol{\theta}))\right).$$

To state the consistency of the tests, we make the following assumptions.

Assumption 7.4 **(E13)** The matrices I^* and J^* are non-singular.

(E14) Appropriate moment conditions such that $\hat{I}_n$ and $\hat{J}_n$ converge to I^* and J^* a.s., respectively.

(E15) The following conditions hold true: $\boldsymbol{\theta}_0 \neq \bar{\boldsymbol{\theta}}$, $\mathrm{E}\left(\partial \ell_t(Z_t, \lambda_t(\bar{\boldsymbol{\theta}}))/\partial \boldsymbol{\theta}\right) \neq \mathbf{0}$, and $\mathrm{E}\epsilon_t(\bar{\boldsymbol{\theta}}) \neq \mathrm{E}\epsilon_t'(\bar{\boldsymbol{\theta}})$, where $\epsilon_t(\boldsymbol{\theta}) := Z_t - \lambda_t(\boldsymbol{\theta})$ and $\epsilon_t'(\boldsymbol{\theta}) := Z_t' - \lambda_t'(\boldsymbol{\theta})$.

Now, we are ready to state the consistency of the tests.

Theorem 7.2 *Suppose the assumptions of Lemma 7.3 and Assumption 7.4. Then, the Wald-type, score-type, and residual type tests are consistent, that is, under the alternative hypothesis H_1, the powers of the tests converge to one as $n \to \infty$.*

Remark 7.5 Existing papers did not investigate the consistency of tests for a structural break for nonnegative integer-valued time series except for [11] for Poisson INGARCH models and [10] for general INGARCH models. They considered the following test statistic:

$$T_{n,\mathrm{DK}} := \max_{v_n \le k \le n - v_n} \frac{1}{q^2\left(\frac{k}{n}\right)} \frac{k^2(n-k)^2}{n^3} (\hat{\boldsymbol{\theta}}_{1:k} - \hat{\boldsymbol{\theta}}_{k+1:n})^\top \hat{J}_{\mathrm{DK}} \hat{I}_{\mathrm{DK}}^{-1} \hat{J}_{\mathrm{DK}} (\hat{\boldsymbol{\theta}}_{1:k} - \hat{\boldsymbol{\theta}}_{k+1:n}),$$

where

$$\hat{I}_{\rm DK} := \frac{1}{2}\left(\sum_{t=1}^{u_n} \tilde{I}_t(\hat{\boldsymbol\theta}_{1:u_n})/u_n + \sum_{t=u_n+1}^{n} \tilde{I}_t(\hat{\boldsymbol\theta}_{u_n+1:n})/(n-u_n)\right),$$

$$\hat{J}_{\rm DK} := \frac{1}{2}\left(\sum_{t=1}^{u_n} \tilde{J}_t(\hat{\boldsymbol\theta}_{1:u_n})/u_n + \sum_{t=u_n+1}^{n} \tilde{J}_t(\hat{\boldsymbol\theta}_{u_n+1:n})/n-u_n\right),$$

u_n and v_n are nonnegative integer-valued sequences such that $u_n, v_n \to \infty$ and $u_n/n, v_n/n \to 0$ as $n \to \infty$, $q(\cdot)$ is an appropriate weight function, and $\hat{\boldsymbol\theta}_{a:b}$ is the Poisson QMLE based on $\{Z_a, \ldots, Z_b\}$.

Remark 7.6 The score type test statistic under general conditions was discussed by [25]. The residual type test statistic for ARMA-GARCH models was considered by [29].

7.4 Numerical Study

In this section, we investigate the finite-sample performance of the proposed tests. We use the linear INGARCH(1,1) model $\lambda_t = \omega + \alpha Z_{t-1} + \beta\lambda_{t-1}$, i.e., Poisson distribution and negative binomial distribution with the parameter r, denoted by Pois(λ_t) and NB(4, $r/(r+\lambda_t)$), respectively.

The critical values are calculated by generating 10000 realizations of the standard Brownian bridge by the R function *BBridge* in the R package *sde* [18] and these are 1.336995 and 3.011263 for $\max_{k=1,\ldots,10000} |B_1^\circ(k/10000)|$ and $\max_{k=1,\ldots,10000} \|B_3^\circ(k/10000)\|_2^2$ at a significance level of 0.05, respectively. Set $v_n = \lfloor(\log n)^2\rfloor$. As a competitor, we use the following test statistic:

$$T_{n,\rm KL} := \max_{v_n\le k\le n} \frac{k^2}{n}(\hat{\boldsymbol\theta}_k - \hat{\boldsymbol\theta}_n)^\top \hat{I}_n(\hat{\boldsymbol\theta}_k - \hat{\boldsymbol\theta}_n),$$

proposed by [19] for the situation where the conditional distribution of a process is the Poisson.

The simulation procedure is as follows: First, we investigate the empirical sizes of the tests. Generate n ($n = 300, 600, 900$) samples from the Poisson or negative binomial INGARCH(1,1) with $r = 4$ and $(\omega, \alpha, \beta) = (1, 0.3, 0.2)$, and estimate parameters by the Poisson QMLE. Then, we calculate the test statistics and replicate these steps 200 times to get rejection probabilities. Second, we simulate the cases that true parameters change from $(\omega, \alpha, \beta) = (1, 0.3, 0.2)$ to $(1, 0.3, 0.4)$ and $(1.5, 0.3, 0.2)$ at $\lfloor n/2\rfloor + 1$, respectively, to study the empirical powers of the tests. The rest of the procedure is the same as the null case.

The results are summarized in Tables 7.1 and 7.2. The existing test statistic $T_{n,\rm KL}$ is sensitive to the misspecification for the non-Poisson case. Our test based on $T_{n,\rm Wald}$ works better than the test based on $T_{n,\rm KL}$ since we do not assume underlying conditional distribution of the process. However, as [23] pointed out, the tests based on

Table 7.1 Empirical size at a nominal size $\alpha = 0.05$

$\lambda_t = 1 + 0.3Z_{t-1} + 0.2\lambda_{t-1}$ for all n					
Conditional distribution	n	$T_{n,\mathrm{KL}}$	$T_{n,\mathrm{Wald}}$	$T_{n,\mathrm{score}}$	$T_{n,\mathrm{res}}$
Pois(λ_t)	300	0.175	0.155	0.035	0.015
	600	0.170	0.180	0.025	0.025
	900	0.130	0.135	0.055	0.080
NB($4, 4/(4 + \lambda_t)$)	300	0.585	0.180	0.040	0.030
	600	0.590	0.145	0.025	0.005
	900	0.575	0.115	0.045	0.045

Table 7.2 The empirical power at the nominal size $\alpha = 0.05$

$\lambda_t = 1 + 0.3Z_{t-1} + 0.2\lambda_{t-1}$ for the first half of n					
$\lambda_t = 1 + 0.3Z_{t-1} + 0.4\lambda_{t-1}$ for the latter half of n					
Conditional distribution	n	$T_{n,\mathrm{KL}}$	$T_{n,\mathrm{Wald}}$	$T_{n,\mathrm{score}}$	$T_{n,\mathrm{res}}$
Pois(λ_t)	300	0.960	0.950	0.435	0.755
	600	1.000	1.000	0.970	1.000
	900	1.000	1.000	1.000	1.000
NB($4, 4/(4 + \lambda_t)$)	300	0.985	0.850	0.315	0.445
	600	1.000	0.990	0.840	0.945
	900	1.000	1.000	0.990	1.000
$\lambda_t = 1 + 0.3Z_{t-1} + 0.2\lambda_{t-1}$ for the first half of n					
$\lambda_t = 1.5 + 0.3Z_{t-1} + 0.2\lambda_{t-1}$ for the latter half of n					
Conditional distribution	n	$T_{n,\mathrm{KL}}$	$T_{n,\mathrm{Wald}}$	$T_{n,\mathrm{score}}$	$T_{n,\mathrm{res}}$
Pois(λ_t)	300	0.780	0.775	0.510	0.670
	600	0.985	0.980	0.950	0.990
	900	1.000	1.000	1.000	1.000
NB($4, 4/(4 + \lambda_t)$)	300	0.950	0.735	0.375	0.595
	600	1.000	0.865	0.865	0.905
	900	1.000	0.975	0.975	0.990

$T_{n,\mathrm{KL}}$ and $T_{n,\mathrm{Wald}}$ show severe size distortions in Table 7.1. This can be explained from the instability of the Poisson QMLE and the fact that the Wald type test statistic is calculated using the Poisson QMLE based on small samples (see Fig. 7.1). In contrast, the tests based on $T_{n,\mathrm{score}}$ and $T_{n,\mathrm{res}}$ provide good size. Regarding the empirical power, Table 7.2 shows the test based on $T_{n,\mathrm{res}}$ has better power than the test based on $T_{n,\mathrm{score}}$.

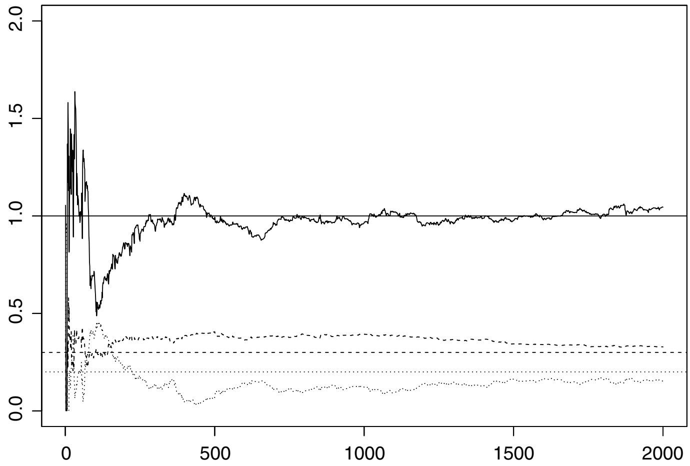


Fig. 7.1 Estimated values of Poisson QMLE based on $\{Z_1, \ldots, Z_k\}$ for $\lambda_t = 1 + 0.3Z_{t-1} + 0.2\lambda_{t-1}$: X-axis represents the number of observations (k). Y-axis represents estimated values $\hat{\boldsymbol{\theta}}_k := (\hat{\theta}_{k,1}, \hat{\theta}_{k,2}, \hat{\theta}_{k,3})$ for (0.1,0.3,0.2). The solid line, the dashed line, and the dotted line correspond to $\hat{\theta}_{k,1}$, $\hat{\theta}_{k,2}$, and $\hat{\theta}_{k,3}$, respectively

7.5 Proof

This section provides proofs of Theorems 7.1 and 7.2. An essential tool to prove Theorem 7.1 is the multi-dimensional martingale FCLT. The one-dimensional martingale FCLT on the Skorokhod space was shown by [6, Theorem 18.2]. For any d-dimensional bounded functions $\boldsymbol{x}(s) := (x_1(s), \ldots, x_d(s))$ and $\boldsymbol{y}(s) := (y_1(s), \ldots, y_d(s))$, we define the uniform metric $d(\cdot, \cdot)$ as

$$d(\boldsymbol{x}, \boldsymbol{y}) := \sup_{\substack{s \in [0,1] \\ i \in \{1,\ldots,d\}}} |x_i(s) - y_i(s)|.$$

The multi-dimensional martingale FCLT on the space of real-valued bounded functions $\ell^\infty([0, 1] \times \{1, \ldots, d\})$ with the uniform metric $d(\cdot, \cdot)$ was given by [27]. The following lemma is due to [27, Theorem 7.2.3].

Lemma 7.4 *Let* $\{\mathbf{m}_t \in \mathbb{R}^d : t \in \mathbb{Z}\}$ *be a martingale difference sequence, i.e.* $\mathrm{E}(\mathbf{m}_t | \mathcal{M}_{t-1}) = \mathbf{0}$, *where* $\mathcal{M}_{t-1}$ *is the* σ*-field generated by* $\{\mathbf{m}_s, s \le t-1\}$ *and* $\mathbf{M}_n(s) := \sum_{j=1}^{\lfloor ns \rfloor} \mathbf{m}_j / \sqrt{n}$. *If, for any* $\epsilon > 0$,

$$\frac{1}{n} \sum_{t=1}^{n} \mathrm{E}\left(\|\mathbf{m}_t\|_2^2 \mathbb{I}_{\{\|\mathbf{m}_t\|_2 > \sqrt{n}\epsilon\}} \mid \mathcal{M}_{t-1} \right) \to 0 \quad \textit{in probability, as } n \to \infty,$$

and

$$\frac{1}{n}\sum_{t=1}^{n} \mathrm{E}\left(\mathbf{m}_t\mathbf{m}_t^\top \mid \mathcal{M}_{t-1}\right) \to \mathbf{\Sigma} \quad \textit{in probability, as } n \to \infty,$$

where $\mathbf{\Sigma}$ *is a positive definite matrix, then,*

$$\mathbf{\Sigma}^{-1/2}\mathbf{M}_n(s) \Rightarrow \mathbf{B}_d(s) \quad \textit{in } \ell^\infty([0, 1] \times \{1, \ldots, d\}) \textit{ as } n \to \infty,$$

where B_d *is a* d*-dimensional standard Brownian motion.*

The multi-dimensional martingale FCLT holds for any ergodic, stationary, and martingale difference process $\{\boldsymbol{\xi}_t \in \mathbb{R}^d : t \in \mathbb{Z}\}$ with a positive definite covariance matrix and $\mathrm{E}\|\boldsymbol{\xi}_1\|_2^{2+\delta} < \infty$ for some $\delta > 0$. Actually, the conditions of Lemma 7.4 can be easily checked: for the σ-field $\mathcal{F}_{t-1}$ generated by $\{\xi_s, s \leq t-1\}$,

$$\frac{1}{n}\sum_{t=1}^{n} \mathrm{E}\left(\|\xi_t\|_2^2 \mathbb{I}_{\{\|\xi_t\|_2 > \sqrt{n}\epsilon\}} \mid \mathcal{F}_{t-1}\right) \leq \frac{1}{\epsilon^2 n^{1+\delta}} \sum_{t=1}^{n} \mathrm{E}\left(\|\xi_t\|_2^{2+\delta} \mid \mathcal{F}_{t-1}\right),$$

which, by the ergodic theorem, converges to 0. The second condition follows from the ergodic theorem. Therefore, we have

$$\left(\mathrm{E}\boldsymbol{\xi}_1\boldsymbol{\xi}_1^\top\right)^{-1/2} \frac{1}{\sqrt{n}} \sum_{j=1}^{\lfloor ns \rfloor} \boldsymbol{\xi}_j \Rightarrow \mathbf{B}_d(s) \quad \text{in } \ell^\infty([0, 1] \times \{1, \ldots, d\}), \text{ as } n \to \infty.$$

Hence,

$$\left(\mathrm{E}\boldsymbol{\xi}_1\boldsymbol{\xi}_1^\top\right)^{-1/2} \left(\frac{1}{\sqrt{n}} \sum_{j=1}^{\lfloor ns \rfloor} \boldsymbol{\xi}_j - \frac{s}{\sqrt{n}} \sum_{j=1}^{n} \boldsymbol{\xi}_j\right) \Rightarrow \mathbf{B}_d(s) - s\mathbf{B}_d(1)$$

in $\ell^\infty([0, 1] \times \{1, \ldots, d\})$, as $n \to \infty$.

7.5.1 Proof of Theorem 7.1

The Wald Type Test Statistic

First, from the definition of the M-estimator and Taylor's expansion, it follows that

$$\begin{aligned}\mathbf{0} &= \sqrt{n}\frac{\partial}{\partial \boldsymbol{\theta}} \tilde{L}_n(\hat{\boldsymbol{\theta}}_n) \\ &= \sqrt{n}\frac{\partial}{\partial \boldsymbol{\theta}} \tilde{L}_n(\boldsymbol{\theta}_0) + \frac{\partial}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \tilde{L}_n({\boldsymbol{\theta}_n}^*)\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0),\end{aligned}$$

where $\boldsymbol{\theta}_0 \lesseqgtr {\boldsymbol{\theta}_n}^* \lesseqgtr \hat{\boldsymbol{\theta}}_n$. Then, we can write the above equation as

$$J\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = \sqrt{n}\frac{\partial}{\partial \boldsymbol{\theta}}\tilde{L}_n(\boldsymbol{\theta}_0) + \Delta_n,$$

where

$$\Delta_n := \begin{cases} -\left(J + \frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\tilde{L}_n(\boldsymbol{\theta}_n{}^*)\right)\left(\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\tilde{L}_n(\boldsymbol{\theta}_n{}^*)\right)^{-1}\sqrt{n}\frac{\partial}{\partial\boldsymbol{\theta}}\tilde{L}_n(\boldsymbol{\theta}_0) & \text{if } \left(\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\tilde{L}_n(\boldsymbol{\theta}_n{}^*)\right)^{-1} \text{ exists,} \\ \left(J + \frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\tilde{L}_n(\boldsymbol{\theta}_n{}^*)\right)\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) & \text{otherwise.} \end{cases}$$

Hence, we obtain the following equation:

$$J\frac{\lfloor ns\rfloor}{\sqrt{n}}(\hat{\boldsymbol{\theta}}_{\lfloor ns\rfloor} - \hat{\boldsymbol{\theta}}_n) = \frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor ns\rfloor}\frac{\partial}{\partial\boldsymbol{\theta}}\tilde{\ell}(Z_t,\tilde{\lambda}_t(\boldsymbol{\theta}_0)) - \frac{\lfloor ns\rfloor}{n}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\tilde{\ell}(Z_t,\tilde{\lambda}_t(\boldsymbol{\theta}_0)) + \sqrt{\frac{\lfloor ns\rfloor}{n}}\Delta_{\lfloor n\tau\rfloor} - \frac{\lfloor ns\rfloor}{n}\Delta_n. \tag{7.1}$$

Second, from (C6),

$$\sup_{0\le s\le 1}\sup_{\boldsymbol{\theta}\in\Theta}\left\|\frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor ns\rfloor}\frac{\partial}{\partial\boldsymbol{\theta}}\ell(Z_t,\tilde{\lambda}_t(\boldsymbol{\theta})) - \frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor ns\rfloor}\frac{\partial}{\partial\boldsymbol{\theta}}\ell(Z_t,\lambda_t(\boldsymbol{\theta}))\right\|$$
$$\le \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\sup_{\boldsymbol{\theta}\in\Theta}\left\|\frac{\partial}{\partial\boldsymbol{\theta}}\ell(Z_t,\tilde{\lambda}_t(\boldsymbol{\theta})) - \frac{\partial}{\partial\boldsymbol{\theta}}\ell(Z_t,\lambda_t(\boldsymbol{\theta}))\right\| \to 0 \quad \text{a.s., as } n\to\infty. \tag{7.2}$$

Third, we shall show that

$$\max_{1\le k\le n}\sqrt{\frac{k}{n}}\,\|\Delta_k\| = o_p(1) \quad \text{as } n\to\infty. \tag{7.3}$$

We follow the proof of [19, Lemma 9]. By Assumption (C7), it can be shown that

$$-\frac{1}{n}\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\tilde{L}_n(\boldsymbol{\theta}_n{}^*) \to J \quad \text{a.s., as } n\to\infty$$

in the same way as [2, Theorem 2.2]. We can apply Egorov's theorem, that is, for any $\epsilon > 0$, there exists some Borel set $A \in \mathcal{F}$ such that $\mathrm{P}(A) < \epsilon$ and

$$-\frac{1}{n}\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\tilde{L}_n(\boldsymbol{\theta}_n{}^*) \to J \quad \text{uniformly on } \Omega\backslash A, \quad \text{as } n\to\infty.$$

There exists N_1 such that, for any $n \ge N_1$,

$$\left|\det(J) - \det\left(-\frac{1}{n}\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\tilde{L}_n(\boldsymbol{\theta}_n{}^*)\right)\right| < \frac{1}{2}|\det(J)| \quad \text{on } \Omega\backslash A,$$

and then

$$\left|\det\left(-\frac{1}{n}\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\tilde{L}_n(\boldsymbol{\theta}_n{}^*)\right)\right| > \frac{1}{2}|\det(J)| \quad \text{on } \Omega\backslash A.$$

Therefore, $\left(-\partial^2/(\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top)\tilde{L}_n(\boldsymbol{\theta}_n{}^*)/n\right)^{-1}$ exists on $\Omega\backslash A$ for any $n \geq N_1$. For any invertible matrix B_n and B such that $B_n \to B$ as $n \to \infty$, $\|B_n^{-1} - B^{-1}\| = \|B_n^{-1}(B_n - B)B^{-1}\| \leq \|B_n^{-1}\|\|(B_n - B)\|\|B^{-1}\| \to 0$. Thus, there exists $N_2 \in \mathbb{N}$ such that, for any $n \geq N_2 \geq N_1$,

$$\left\|\left(-\frac{1}{n}\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\tilde{L}_n(\boldsymbol{\theta}_n{}^*)\right)^{-1}\right\| < \frac{3}{2}\left\|J^{-1}\right\| \quad \text{on } \Omega\backslash A.$$

For any $\eta > 0$,

$$\begin{aligned}
&\mathrm{P}\left(\max_{1\leq k\leq n}\sqrt{\frac{k}{n}}\,\|\Delta_k\| > \eta\right)\\
\leq &\mathrm{P}\left(\max_{1\leq k\leq N_2}\sqrt{\frac{k}{n}}\,\|\Delta_k\| > \eta,\ \Omega\backslash A\right) + \mathrm{P}\left(\max_{N_2+1\leq k\leq n}\sqrt{\frac{k}{n}}\,\|\Delta_k\| > \eta,\ \Omega\backslash A\right) + \mathrm{P}(A).
\end{aligned} \tag{7.4}$$

From the definition of the tightness in $\mathbb{R}$, for large n,

$$\mathrm{P}\left(\max_{1\leq k\leq N_2}\sqrt{k}\,\|\Delta_k\| > n\eta,\ \Omega\backslash A\right) < \epsilon.$$

Therefore, the first term of (7.4) converges to 0. The second term of (7.4) is asymptotically negligible along the line of [19, Lemma 9], which concludes (7.2).

Fourth, the multi-dimensional martingale FCLT (Lemma 7.4) yields

$$\begin{aligned}
&\hat{I}_n^{-1/2}\frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor ns\rfloor}\frac{\partial}{\partial\boldsymbol{\theta}}\ell(Z_t,\lambda_t(\boldsymbol{\theta})) - \hat{I}_n^{-1/2}\frac{\lfloor ns\rfloor}{n}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\ell(Z_t,\lambda_t(\boldsymbol{\theta}))\\
&= I^{-1/2}\frac{1}{\sqrt{n}}\sum_{t=1}^{\lfloor ns\rfloor}\frac{\partial}{\partial\boldsymbol{\theta}}\ell(Z_t,\lambda_t(\boldsymbol{\theta})) - I^{-1/2}\frac{\lfloor ns\rfloor}{n}\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\ell(Z_t,\lambda_t(\boldsymbol{\theta})) + o_p(1)\\
&\Rightarrow \mathbf{B}_d(s) - s\mathbf{B}_d(1) \quad \text{in } \ell^\infty([0,1]\times\{1,\ldots,d\}), \text{ as } n\to\infty.
\end{aligned} \tag{7.5}$$

Hence, from (7.1)–(7.3), (7.5), and the continuous mapping theorem, we obtain, for $\iota_n := \inf\{s \in [0,1] : \nu_n \leq \lfloor ns\rfloor\}$,

$$
\begin{aligned}
&T_{n,\mathrm{Wald}} \\
&= \max_{\nu_n \le k \le n} \left\| \frac{k}{\sqrt{n}} \hat{I}_n^{-1/2} \hat{J}_n (\hat{\boldsymbol{\theta}}_k - \hat{\boldsymbol{\theta}}_n) \right\|_2^2 \\
&= \sup_{\iota_n \le s \le 1} \left\| I^{-1/2} \hat{J}_n J^{-1} \left(\frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor ns \rfloor} \frac{\partial}{\partial \boldsymbol{\theta}} \tilde{\ell}(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta})) - \frac{\lfloor ns \rfloor}{n} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta})) \right) \right\|_2^2 \\
&+ o_p(1) \\
&= \sup_{\iota_n \le s \le 1} \left\| I^{-1/2} \left(\frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor ns \rfloor} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta})) - \frac{\lfloor ns \rfloor}{n} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(Z_t, \lambda_t(\boldsymbol{\theta})) \right) \right\|_2^2 \\
&+ o_p(1) \\
\Rightarrow &\sup_{0 \le s \le 1} \| B_d^{\circ}(s) \|_2^2,
\end{aligned}
$$

as $n \to \infty$.

The Score-Type Test Statistic
First, we show that

$$
\begin{aligned}
\max_{1 \le k \le n} \frac{1}{\sqrt{n}} \Bigg\| &\sum_{t=1}^{k} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(Z_t, \tilde{\lambda}_t(\hat{\boldsymbol{\theta}}_n)) \\
&- \left(\sum_{t=1}^{k} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta}_0)) - \frac{k}{n} \sum_{t=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta}_0)) \right) \Bigg\| = o_p(1) \qquad (7.6)
\end{aligned}
$$

as $n \to \infty$.

By Taylor's expansion, there exist $\boldsymbol{\theta}_n{}^*$ and $\boldsymbol{\theta}_n{}^{**}$ such that $\boldsymbol{\theta}_0 \lesseqgtr \boldsymbol{\theta}_n{}^* \lesseqgtr \hat{\boldsymbol{\theta}}_n$, $\boldsymbol{\theta}_0 \lesseqgtr \boldsymbol{\theta}_n{}^{**} \lesseqgtr \hat{\boldsymbol{\theta}}_n$,

$$
\sum_{t=1}^{k} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(Z_t, \tilde{\lambda}_t(\hat{\boldsymbol{\theta}}_n)) - \sum_{t=1}^{k} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta}_0)) = \sum_{t=1}^{k} \frac{\partial}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \ell(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta}_n^{**})) \left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \right),
$$

and

$$
\sum_{t=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(Z_t, \tilde{\lambda}_t(\hat{\boldsymbol{\theta}}_n)) - \sum_{t=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta}_0)) = \sum_{t=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \ell(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta}_n^{*})) \left(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \right).
$$

Then, we have

$$
\begin{aligned}
&\max_{1\le k\le n}\frac{1}{\sqrt{n}}\Bigg\|\sum_{t=1}^{k}\frac{\partial}{\partial\boldsymbol{\theta}}\ell(Z_t,\tilde{\lambda}_t(\hat{\boldsymbol{\theta}}_n))\\
&\quad-\left(\sum_{t=1}^{k}\frac{\partial}{\partial\boldsymbol{\theta}}\ell(Z_t,\tilde{\lambda}_t(\boldsymbol{\theta}_0))-\frac{k}{n}\sum_{t=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\ell(Z_t,\tilde{\lambda}_t(\boldsymbol{\theta}_0))\right)\Bigg\|\\
&=\max_{1\le k\le n}\frac{1}{n}\Bigg\|\sum_{t=1}^{k}\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\top}}\ell(Z_t,\tilde{\lambda}_t(\boldsymbol{\theta}_n^{**}))\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)\\
&\quad-\frac{k}{n}\sum_{t=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\top}}\ell(Z_t,\tilde{\lambda}_t(\boldsymbol{\theta}_n^{*}))\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)\Bigg\|\\
&\le\left\|\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)\right\|\max_{1\le k\le n}\frac{k}{n}\left\|\frac{1}{k}\sum_{t=1}^{k}\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\top}}\ell(Z_t,\tilde{\lambda}_t(\boldsymbol{\theta}_n^{**}))+J\right\|\\
&\quad+\left\|\frac{1}{n}\sum_{t=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\top}}\ell(Z_t,\tilde{\lambda}_t(\boldsymbol{\theta}_n^{*}))+J\right\|\left\|\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)\right\|.
\end{aligned}
$$

Since $\left\|\sqrt{n}\left(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right)\right\|=O_p(1)$ and

$$
\left\|\frac{1}{n}\sum_{t=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\top}}\tilde{\ell}_t({\boldsymbol{\theta}_n}^{**})+J\right\|\to 0\quad\text{a.s., as } n\to\infty, \tag{7.7}
$$

the second term is asymptotically negligible. From (7.7), for any $\epsilon>0$, there exists $N_4>0$ such that, for $k\ge N_4$,

$$
\mathrm{P}\left(\left\|\frac{1}{k}\sum_{t=1}^{k}\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\top}}\ell(Z_t,\tilde{\lambda}_t(\boldsymbol{\theta}_n^{**}))+J\right\|>\epsilon\right)=0.
$$

Hence,

$$
\begin{aligned}
&\max_{1\le k\le n}\frac{k}{n}\left\|\frac{1}{k}\sum_{t=1}^{k}\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\top}}\ell(Z_t,\tilde{\lambda}_t(\boldsymbol{\theta}_n^{**}))+J\right\|\\
&=\max_{1\le k\le N_4}\frac{k}{n}\left\|\frac{1}{k}\sum_{t=1}^{k}\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\top}}\ell(Z_t,\tilde{\lambda}_t(\boldsymbol{\theta}_n^{**}))+J\right\|\\
&\quad+\max_{N_4\le k\le n}\frac{k}{n}\left\|\frac{1}{k}\sum_{t=1}^{k}\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\top}}\ell(Z_t,\tilde{\lambda}_t(\boldsymbol{\theta}_n^{**}))+J\right\|\\
&\le\frac{N_4}{n}\sum_{t=1}^{N_4}\left\|\frac{\partial}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^{\top}}\ell(Z_t,\tilde{\lambda}_t(\boldsymbol{\theta}_n^{**}))\right\|+\frac{N_4}{n}\left\|J\right\|
\end{aligned}
$$

$$+ \max_{N_4 \le k \le n} \left\| \frac{1}{k} \sum_{t=1}^{k} \frac{\partial}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \ell(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta}_n^{**})) + J \right\|,$$

which tends to 0 in probability since, for any ϵ', there exists M such that

$$\mathrm{P}\left(\sum_{t=1}^{N_4} \left\| \frac{\partial}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} \ell(Z_t, \tilde{\lambda}_t(\boldsymbol{\theta}_n^{**})) \right\| > M \right) \le \epsilon'.$$

Consequently, (7.6) is established.

Second, we know that

$$\begin{aligned} &\Bigg| \sup_{0 \le s \le 1} \left\| \hat{I}_n^{-1/2} \frac{1}{\sqrt{n}} \left(\sum_{t=1}^{\lfloor ns \rfloor} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(Z_t, \tilde{\lambda}_t(\hat{\boldsymbol{\theta}}_n)) \right) \right\| \\ &\quad - \sup_{0 \le s \le 1} \left\| \hat{I}_n^{-1/2} \frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor ns \rfloor} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(Z_t, \lambda_t(\hat{\boldsymbol{\theta}}_n)) \right\| \Bigg| \\ &\le \left\| \hat{I}_n^{-1/2} \right\| \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left\| \left(\frac{\partial}{\partial \boldsymbol{\theta}} \ell(Z_t, \tilde{\lambda}_t(\hat{\boldsymbol{\theta}}_n)) \right) - \left(\frac{\partial}{\partial \boldsymbol{\theta}} \ell(Z_t, \lambda_t(\hat{\boldsymbol{\theta}}_n)) \right) \right\| \to 0 \quad \text{a.s., as } n \to \infty. \end{aligned} \tag{7.8}$$

Third, by employing the multi-dimensional martingale FCLT, we can see that

$$\begin{aligned} &\hat{I}_n^{-1/2} \frac{1}{\sqrt{n}} \sum_{t=1}^{\lfloor ns \rfloor} \frac{\partial}{\partial \boldsymbol{\theta}} \ell_t(\boldsymbol{\theta}_0) - \hat{I}_n^{-1/2} \frac{\lfloor ns \rfloor}{n} \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \ell_t(\boldsymbol{\theta}_0) \\ \Rightarrow &\mathbf{B}_d(s) - s\mathbf{B}_d(1) \quad \text{in } \ell^\infty([0, 1] \times \{1, \ldots, d\}), \text{ as } n \to \infty. \end{aligned} \tag{7.9}$$

Hence, (7.6), (7.8), (7.9), the ergodic theorem, and the continuous mapping theorem yield, for $\iota_n := \inf\{s \in [0, 1] : v_n \le \lfloor ns \rfloor\}$,

$$\sup_{\iota_n \le s \le 1} \left\| \hat{I}_n^{-1/2} \frac{1}{\sqrt{n}} \left(\sum_{t=1}^{\lfloor ns \rfloor} \frac{\partial}{\partial \boldsymbol{\theta}} \ell(Z_t, \tilde{\lambda}_t(\hat{\boldsymbol{\theta}}_n)) \right) \right\|_2^2 \Rightarrow \sup_{0 \le s \le 1} \left\| \mathbf{B}_d^\circ(s) \right\|_2^2,$$

as $n \to \infty$.

The residual type test statistic

First, we shall show that, for $\epsilon_t := Z_t - \lambda(\boldsymbol{\theta}_0)$,

$$\max_{1 \le k \le n} \frac{1}{\sqrt{n}} \left| \sum_{t=1}^{k} (\tilde{\epsilon}_t(\hat{\boldsymbol{\theta}}_n) - \epsilon_t) - \frac{k}{n} \sum_{t=1}^{n} (\tilde{\epsilon}_t(\hat{\boldsymbol{\theta}}_n) - \epsilon_t) \right| = o_p(1) \quad \text{as } n \to \infty. \tag{7.10}$$

By Taylor's expansion, we have, for $a_t := \sup_{\boldsymbol{\theta}\in\Theta} \left|\lambda_t(\boldsymbol{\theta}) - \tilde{\lambda}_t(\boldsymbol{\theta})\right|$,

$$\begin{aligned}
&\max_{1\le k\le n} \frac{1}{\sqrt{n}} \left| \sum_{t=1}^{k} (\tilde{\epsilon}_t(\hat{\boldsymbol{\theta}}_n) - \epsilon_t) - \frac{k}{n} \sum_{t=1}^{n} (\tilde{\epsilon}_t(\hat{\boldsymbol{\theta}}_n) - \epsilon_t) \right| \\
\le& \frac{2}{\sqrt{n}} \sum_{t=1}^{n} a_t + \max_{1\le k\le n} \frac{1}{\sqrt{n}} \left| \sum_{t=1}^{k} (\lambda_t(\hat{\boldsymbol{\theta}}_n) - \lambda_t(\boldsymbol{\theta}_0)) - \frac{k}{n} \sum_{t=1}^{n} (\lambda_t(\hat{\boldsymbol{\theta}}_n) - \lambda_t(\boldsymbol{\theta}_0)) \right| \\
\le& \frac{2}{\sqrt{n}} \sum_{t=1}^{n} a_t + \max_{1\le k\le n} \frac{1}{\sqrt{n}} \left| \sum_{t=1}^{k} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \frac{\partial}{\partial \boldsymbol{\theta}} \lambda_t(\boldsymbol{\theta}_0) - \frac{k}{n} \sum_{t=1}^{n} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \frac{\partial}{\partial \boldsymbol{\theta}} \lambda_t(\boldsymbol{\theta}_0) \right| \\
&+ \max_{1\le k\le n} \frac{1}{\sqrt{n}} \left| \sum_{t=1}^{k} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \left(\frac{\partial}{\partial \boldsymbol{\theta}} \lambda_t(\boldsymbol{\theta}_n{}^*) - \frac{\partial}{\partial \boldsymbol{\theta}} \lambda_t(\boldsymbol{\theta}_0) \right) \right. \\
&\qquad \left. - \frac{k}{n} \sum_{t=1}^{n} (\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0)^\top \left(\frac{\partial}{\partial \boldsymbol{\theta}} \lambda_t(\boldsymbol{\theta}_n{}^*) - \frac{\partial}{\partial \boldsymbol{\theta}} \lambda_t(\boldsymbol{\theta}_0) \right) \right| \\
\le& \frac{2}{\sqrt{n}} \sum_{t=1}^{n} a_t + \sqrt{n} \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \max_{1\le k\le n} \frac{k}{n} \left\| \frac{1}{k} \sum_{t=1}^{k} \frac{\partial}{\partial \boldsymbol{\theta}} \lambda_t(\boldsymbol{\theta}_0) - \frac{1}{n} \sum_{t=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \lambda_t(\boldsymbol{\theta}_0) \right\| \\
&+ \sqrt{n} \|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| \frac{2}{n} \sum_{t=1}^{n} \left\| \frac{\partial}{\partial \boldsymbol{\theta}} \lambda_t(\boldsymbol{\theta}_n{}^*) - \frac{\partial}{\partial \boldsymbol{\theta}} \lambda_t(\boldsymbol{\theta}_0) \right\|,
\end{aligned}$$

where $\boldsymbol{\theta}_0 \leqslant \boldsymbol{\theta}_n{}^* \leqslant \hat{\boldsymbol{\theta}}_n$. Under the assumption (D11), we can show that

$$\frac{2}{n} \sum_{t=1}^{n} \left\| \frac{\partial}{\partial \boldsymbol{\theta}} \lambda_t(\boldsymbol{\theta}_n{}^*) - \frac{\partial}{\partial \boldsymbol{\theta}} \lambda_t(\boldsymbol{\theta}_0) \right\| \to 0 \quad \text{a.s., as } n \to \infty$$

along the line of [2, Theorem 2.2] and, by the ergodic theorem, it is easy to see that

$$\max_{1\le k\le n} \frac{k}{n} \left\| \frac{1}{k} \sum_{t=1}^{k} \frac{\partial}{\partial \boldsymbol{\theta}} \lambda_t(\boldsymbol{\theta}_0) - \frac{1}{n} \sum_{t=1}^{n} \frac{\partial}{\partial \boldsymbol{\theta}} \lambda_t(\boldsymbol{\theta}_0) \right\| \to 0 \quad \text{a.s., as } n \to \infty.$$

Using $\sqrt{n}\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| = O_p(1)$, the proof of (7.10) is completed.

Second, we prove that

$$\frac{1}{n} \sum_{t=1}^{n} \tilde{\epsilon}_t^2(\hat{\boldsymbol{\theta}}_n) \to \mathrm{E}\epsilon_t^2 \quad \text{in probability as } n \to \infty. \tag{7.11}$$

By Taylor's expansion, we obtain

$$
\begin{aligned}
&\left|\frac{1}{n}\sum_{t=1}^{n}(\hat{\epsilon}_t^2(\hat{\boldsymbol{\theta}}_n)-\epsilon_t^2)\right| \\
&\quad=\left|\frac{1}{n}\sum_{t=1}^{n}(\hat{\epsilon}_t(\hat{\boldsymbol{\theta}}_n)-\epsilon_t)(\hat{\epsilon}_t(\hat{\boldsymbol{\theta}}_n)+\epsilon_t)\right| \\
&\leq\frac{1}{n}\sum_{t=1}^{n}\left(a_t+\left|(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)^\top\frac{\partial}{\partial\boldsymbol{\theta}}\lambda_t(\boldsymbol{\theta}_n^*)\right|\right) \\
&\quad\times\left(2|Z_t-\lambda_t(\boldsymbol{\theta}_0)|+a_t+\left|(\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0)^\top\frac{\partial}{\partial\boldsymbol{\theta}}\lambda_t(\boldsymbol{\theta}_n^*)\right|\right) \\
&\leq\frac{1}{n}\sum_{t=1}^{n}\left(a_t^2+2a_t|Z_t-\lambda_t(\boldsymbol{\theta}_0)|+2a_t\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|\left\|\frac{\partial}{\partial\boldsymbol{\theta}}\lambda_t(\boldsymbol{\theta}_n^*)\right\|\right. \\
&\quad\left.+2|Z_t-\lambda_t(\boldsymbol{\theta}_0)|\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|\left\|\frac{\partial}{\partial\boldsymbol{\theta}}\lambda_t(\boldsymbol{\theta}_n^*)\right\|+\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|^2\left\|\frac{\partial}{\partial\boldsymbol{\theta}}\lambda_t(\boldsymbol{\theta}_n^*)\right\|^2\right) \\
&\leq\frac{1}{n}\sum_{t=1}^{n}a_t^2+2\sqrt{\frac{1}{n}\sum_{t=1}^{n}a_t^2}\sqrt{\frac{1}{n}\sum_{t=1}^{n}|Z_t-\lambda_t(\boldsymbol{\theta}_0)|^2} \\
&\quad+2\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|\sqrt{\frac{1}{n}\sum_{t=1}^{n}a_t^2}\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left\|\frac{\partial}{\partial\boldsymbol{\theta}}\lambda_t(\boldsymbol{\theta}_n^*)\right\|^2} \\
&\quad+2\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|\sqrt{\frac{1}{n}\sum_{t=1}^{n}|Z_t-\lambda_t(\boldsymbol{\theta}_0)|^2}\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left\|\frac{\partial}{\partial\boldsymbol{\theta}}\lambda_t(\boldsymbol{\theta}_n^*)\right\|^2} \\
&\quad+\left\|\hat{\boldsymbol{\theta}}_n-\boldsymbol{\theta}_0\right\|^2\frac{1}{n}\sum_{t=1}^{n}\left\|\frac{\partial}{\partial\boldsymbol{\theta}}\lambda_t(\boldsymbol{\theta}_n^*)\right\|^2,
\end{aligned}
$$

where $\boldsymbol{\theta}_0 \leqslant \boldsymbol{\theta}_n{}^* \leqslant \hat{\boldsymbol{\theta}}_n$. By the ergodic theorem, we observe

$$\frac{1}{n}\sum_{t=1}^{n}|Z_t-\lambda_t(\boldsymbol{\theta}_0)|^2 \to \mathrm{E}\epsilon_t^2 \quad \text{a.s., as } n\to\infty$$

and, by Assumption (D10), we can show

$$\frac{1}{n}\sum_{t=1}^{n}\left\|\frac{\partial}{\partial\boldsymbol{\theta}}\lambda_t(\boldsymbol{\theta}_n^*)\right\|^2 \to \mathrm{E}\left\|\frac{\partial}{\partial\boldsymbol{\theta}}\lambda_t(\boldsymbol{\theta}_0)\right\|^2 \quad \text{a.s., as } n\to\infty,$$

which shows (7.11).

Since ϵ_t is strictly stationary, ergodic, and martingale difference, we can apply the multi-dimensional martingale FCLT. From (7.10) and (7.11), we have, for $\iota_n := \inf\{s \in [0,1] : \nu_n \leq \lfloor ns \rfloor\}$,

$$
\begin{aligned}
&\max_{1\le k\le n} \frac{1}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\hat{\epsilon}_t^2(\hat{\boldsymbol{\theta}}_n)}} \frac{1}{\sqrt{n}} \left| \sum_{t=1}^{k} \tilde{\epsilon}_t - \frac{k}{n}\sum_{t=1}^{n}\tilde{\epsilon}_t \right| \\
&= \sup_{\iota_n \le s\le 1} \frac{1}{\sqrt{\mathrm{E}\epsilon_t^2}} \frac{1}{\sqrt{n}} \left| \sum_{t=1}^{\lfloor ns\rfloor} \epsilon_t - \frac{\lfloor ns\rfloor}{n}\sum_{t=1}^{n}\epsilon_t \right| + o_p(1) \\
&\Rightarrow \sup_{0\le s\le 1} |B_1^\circ(s)|,
\end{aligned}
$$

as $n \to \infty$. □

7.5.2 *Proof of Theorem 7.2*

For the test based on $T_{n,\mathrm{Wald}}$, it is easy to see that

$$
\begin{aligned}
P\left(\frac{1}{n}T_{n,\mathrm{Wald}} > \frac{C}{n}\right) &\ge P\left(\frac{\lfloor n\tau\rfloor^2}{n^2}(\hat{\boldsymbol{\theta}}_{\lfloor n\tau\rfloor} - \hat{\boldsymbol{\theta}}_n)^\top \hat{J}\hat{I}^{-1}\hat{J}(\hat{\boldsymbol{\theta}}_{\lfloor n\tau\rfloor} - \hat{\boldsymbol{\theta}}_n) > \frac{C}{n}\right) \\
&= P\left(\tau^2(\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}})^\top J^*(I^*)^{-1}J^*(\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}}) > 0\right) + o(1) \\
&\ge P\left(\tau^2\Lambda_{\min}\left(J^*(I^*)^{-1}J^*\right)\|\boldsymbol{\theta}_0 - \bar{\boldsymbol{\theta}}\|_2^2 > 0\right) + o(1), \quad (7.12)
\end{aligned}
$$

where $\Lambda_{\min}(\mathrm{A})$ denotes a minimum eigenvalue of a matrix A. The assumptions that J^* and I^* are positive definite matrices imply $J^*(I^*)^{-1}J^*$ is a positive definite matrix. Thus, $\Lambda_{\min}(J^*(I^*)^{-1}J^*) > 0$, which concludes that (7.12) converges to 1 as $n \to \infty$.

For the test based on $T_{n,\mathrm{score}}$, we observe that

$$
\begin{aligned}
&P\left(\frac{1}{n}T_{n,\mathrm{score}} > \frac{C}{n}\right) \\
&\ge P\left(\left(\frac{1}{n}\sum_{t=1}^{\lfloor n\tau\rfloor}\frac{\partial}{\partial\boldsymbol{\theta}}\tilde{\ell}_t(Z_t,\lambda_t(\hat{\boldsymbol{\theta}}_n))\right)^\top \hat{I}^{-1}\left(\frac{1}{n}\sum_{t=1}^{\lfloor n\tau\rfloor}\frac{\partial}{\partial\boldsymbol{\theta}}\tilde{\ell}_t(Z_t,\lambda_t(\hat{\boldsymbol{\theta}}_n))\right) > \frac{C}{n}\right) \\
&= P\left(\tau^2\mathrm{E}\left(\frac{\partial}{\partial\boldsymbol{\theta}^\top}\ell_t(Z_t,\lambda_t(\bar{\boldsymbol{\theta}}))\right)(I^*)^{-1}\mathrm{E}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\ell_t(Z_t,\lambda_t(\bar{\boldsymbol{\theta}}))\right) > 0\right) + o(1) \\
&\ge P\left(\tau^2\Lambda_{\min}\left((I^*)^{-1}\right)\left\|\mathrm{E}\left(\frac{\partial}{\partial\boldsymbol{\theta}}\ell_t(Z_t,\lambda_t(\bar{\boldsymbol{\theta}}))\right)\right\|_2^2 > 0\right) + o(1),
\end{aligned}
$$

which converges to 1 as $n \to \infty$, since I^* is a positive definite matrix.

For the test based on $T_{n,\mathrm{res}}$, with $\epsilon_t(\boldsymbol{\theta}) := Z_t - \lambda_t(\boldsymbol{\theta})$ and $\epsilon_t'(\boldsymbol{\theta}) := Z_t' - \lambda_t'(\boldsymbol{\theta})$, we have

$$
\begin{aligned}
&P\left(\frac{1}{\sqrt{n}}T_{n,\mathrm{res}} > \frac{C}{\sqrt{n}}\right)\\
\geq &P\left(\frac{1}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\tilde{\epsilon}_t^2(\hat{\boldsymbol{\theta}}_n)}}\frac{1}{n}\left|\sum_{t=1}^{\lfloor n\tau\rfloor}\tilde{\epsilon}_t(\hat{\boldsymbol{\theta}}_n) - \frac{\lfloor n\tau\rfloor}{n}\sum_{t=1}^{n}\tilde{\epsilon}_t(\hat{\boldsymbol{\theta}}_n)\right| > \frac{C'}{\sqrt{n}}\right)\\
= &P\left(\frac{\tau(1-\tau)}{\sqrt{\tau\mathrm{E}\epsilon_t^2(\bar{\boldsymbol{\theta}}) + (1-\tau)\mathrm{E}\epsilon_t'^2(\bar{\boldsymbol{\theta}})}}\left|\mathrm{E}\epsilon_t(\bar{\boldsymbol{\theta}}) - \mathrm{E}\epsilon_t'(\bar{\boldsymbol{\theta}})\right| > 0\right) + o(1),
\end{aligned}
$$

which converges to 1 as $n \to \infty$. □

Acknowledgements The author would like to express my deepest gratitude to Professor Masanobu Taniguchi for all your support. Your kind and warm guidance always encouraged the author very much. The author is so proud of being your last disciple. The author is grateful to the editors and referee for their instructive comments. The author would also like to thank Doctor Yan Liu and Doctor Akitoshi Kimura for their encouragements and comments. This research was supported by Grant-in-Aid for JSPS Research Fellow Grant Number JP201920060 and Grant-in-Aid for Research Activity Start-up JP21K20338.

References

1. Agosto, A., Cavaliere, G., Kristensen, D. and Rahbek, A. (2016). Modeling corporate defaults: Poisson autoregressions with exogenous covariates (PARX). *J Empirical Finance* **38** 640–663.
2. Ahmad, A. and Francq, C. (2016). Poisson QMLE of count time series models. *J Time Series Anal* **37** 291–314.
3. Aknouche, A. and Francq, C. (2020). Count and duration time series with equal conditional stochastic and mean orders. *Econom Theory* **37** 248–280.
4. Aknouche, A., Bendjeddou, S. and Touche, N. (2018). Negative binomial quasi-likelihood inference for general integer-valued time series models. *J Time Series Anal* **39** 192–211.
5. Berkes, I., Horváth, L. and Kokoszka, P. (2004). Testing for parameter constancy in GARCH(p, q) models. *Statist Prob Lett* **70** 263–273.
6. Billinsley, P. (1999). *Convergence of Probability Measures*. 2nd edn, New York: Wiley.
7. Cui, Y., Wu, R. and Zheng, Q. (2020). Estimation of change-point for a class of count time series models. *Scand J Statist* **48** 1277–1313.
8. Davis, R. A. and Liu, H. (2016). Theory and inference for a class of nonlinear models with application to time series of counts. *Stat Sin* **26** 1673–1707.
9. Diop, M. L. and Kengne, W. (2017). Testing parameter change in general integer-valued time series. *J Time Series Anal* **38** 880–894.
10. Diop, M. L. and Kengne, W. (2022). Poisson QMLE for change-point detection in general integer-valued time series models. *Metrika* **85** 373–403.
11. Doukhan, P. and Kengne, W. (2015). Inference and testing for structural change in general Poisson autoregressive models. *Electron J Statist* **9** 1267–1314.
12. Doukhan, P., Fokianos, K. and Tjstheim, D. (2013). Correction to "On weak dependence conditions for Poisson autoregressions" [Statist. Probab. Lett. 82 (2012) 942–948]. *Statist Probab Lett* **83** 1926–1927.

13. Ferland, R., Latour, A. and Oraichi, D. (2006). Integer-valued GARCH process. *J Time Series Anal* **27** 923–942.
14. Fokianos, K., Rahbek, A. and Tjstheim, D. (2009). Poisson autoregression. *J Amer Statist Assoc* **104** 1430–1439.
15. Franke, J., Kirch, C. and Kamgaing, J. T. (2012). Changepoints in times series of counts. *J Time Series Anal* **33** 757–770.
16. Goto, Y. and Fujimori, K. (2023). Test for conditional variance of integer-valued time series. *to appear in Stat Sin* **33**.
17. Hudecová, Š., Hušková, M. and Meintanis, S. G. (2017). Tests for structural changes in time series of counts. *Scand J Statist* **44** 843–865.
18. Iacus, S. M. (2016). sde: Simulation and inference for stochastic differential equations. URL https://CRAN.R-project.org/package=sde, R package version 2.0.15
19. Kang, J. and Lee, S. (2014). Parameter change test for Poisson autoregressive models. *Scand J Statist* **41** 1136–1152.
20. Latour, A. (1997). The multivariate GINAR(p) process. *Adv Appl Probab* **29** 228–248.
21. Lee, S., Ha, J., Na, O. and Na, S. (2003). The cusum test for parameter change in time series models. *Scand J Statist* **30** 781–796.
22. Lee, S., Lee, Y. and Chen, C. W. (2016). Parameter change test for zero-inflated generalized Poisson autoregressive models. *Statistics* **50** 540–557.
23. Lee, Y. and Lee, S. (2019). CUSUM test for general nonlinear integer-valued GARCH models: comparison study. *Ann Inst Statist Math* **71** 1033–1057.
24. Lee, Y., Lee, S. and Tjstheim, D. (2018). Asymptotic normality and parameter change test for bivariate Poisson INGARCH models. *Test* **27** 52–69.
25. Negri, I. and Nishiyama, Y. (2017). Z-process method for change point problems with applications to discretely observed diffusion processes. *Stat Methods Appt* **26** 231–250.
26. Neumann, M. H. (2011). Absolute regularity and ergodicity of Poisson count processes. *Bernoulli* **17** 1268–1284.
27. Nishiyama, Y. (2021). *Martingale Methods in Statistics*. Chapman and Hall/CRC.
28. Oh, H. and Lee, S. (2018). On score vector-and residual-based CUSUM tests in ARMA–GARCH models. *Stat Methods Appt* **27** 385–406.
29. Oh, H. and Lee, S. (2019). Modified residual CUSUM test for location-scale time series models with heteroscedasticity. *Ann Inst Statist Math* **71** 1059–1091.
30. Page, E. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika* **42** 523–527.
31. Parker, M. R. P., Li, Y., Elliott, L. T., Ma, J. and Cowen, L. L. E. (2021). Under-reporting of COVID-19 in the Northern Health Authority region of British Columbia. *Can J Stat* **49** 1018–1038.
32. Pedeli, X., Davison, A. C. and Fokianos, K. (2015). Likelihood estimation for the INAR(p) model by saddlepoint approximation. *J Amer Statist Assoc* **110** 1229–1238.
33. Schmidt, A. M. and Pereira, J. B. M. (2011). Modelling time series of counts in epidemiology. *Int Stat Rev* **79** 48–69.
34. Wang, C., Liu, H., Yao, J.-F., Davis, R. A. and Li, W. K. (2014). Self-excited threshold Poisson autoregression. *J Amer Statist Assoc* **109** 777–787.

Chapter 8
M-Estimation in GARCH Models in the Absence of Higher-Order Moments

Marc Hallin, Hang Liu, and Kanchan Mukherjee

Abstract We consider a class of M-estimators of the parameters of a GARCH(p, q) model. These estimators are asymptotically normal, depending on score functions, under milder moment assumptions than the usual quasi-maximum likelihood, which makes them more reliable in the presence of heavy tails. We also consider weighted bootstrap approximations of the distributions of these M-estimators and establish their validity. Through extensive simulations, we demonstrate the robustness of these M-estimators under heavy tails and conduct a comparative study of the performance (biases and mean squared errors) for various score functions and the accuracy (confidence interval coverage probabilities) of their bootstrap approximations. In addition to the GARCH(1,1) model, our simulations also involve higher-order models such as GARCH(2,1) and GARCH(1,2) which so far have received relatively little attention in the literature. We also consider the case of order-misspecified models. Finally, we analyze two real financial time series datasets by fitting GARCH(1,1) or GARCH(2,1) models with our M-estimators.

M. Hallin (✉)
ECARES and Département de Mathématique, Université Libre de Bruxelles CP 114/4, Avenue F.D. Roosevelt 50, 1050 Bruxelles, Belgium
e-mail: mhallin@ulb.ac.be

H. Liu
International Institute of Finance, School of Management, University of Science and Technology of China, Hefei, Anhui 230026, China
e-mail: hliu01@ustc.edu.cn

K. Mukherjee
Department of Mathematics and Statistics, Lancaster University, LA1 4YF Lancaster, United Kingdom
e-mail: k.mukherjee@lancaster.ac.uk

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_8

8.1 Introduction

Generalized AutoRegressive Conditional Heteroscedastic (GARCH) models have been used extensively to analyze the volatility or the instantaneous variability in financial time series. This is a domain in which Prof. Masanobu Taniguchi and his coauthors published impactful papers [6, 10] and two influential monographs [11, 12].

A stochastic process $\mathcal{X} := \{X_t; t \in \mathbb{Z}\}$ is said to follow a GARCH(p, q) model if

$$X_t = \sigma_t \epsilon_t, \quad t \in \mathbb{Z}, \tag{8.1}$$

where $\{\epsilon_t; t \in \mathbb{Z}\}$ is a sequence of (unobservable) i.i.d. errors with symmetric distribution around zero and $\{\sigma_t; t \in \mathbb{Z}\}$ is a solution of

$$\sigma_t = \left(\omega_0 + \sum_{i=1}^{p} \alpha_{0i} X_{t-i}^2 + \sum_{j=1}^{q} \beta_{0j} \sigma_{t-j}^2\right)^{1/2}, \quad t \in \mathbb{Z}, \tag{8.2}$$

for some $\omega_0 > 0$, $\alpha_{0i} > 0$, $i = 1, \ldots, p$, and $\beta_{0j} > 0$, $j = 1, \ldots, q$. Mukherjee [7] introduced a class of M-estimators for estimating the GARCH parameter

$$\boldsymbol{\theta}_0 := (\omega_0, \alpha_{01}, \ldots, \alpha_{0p}, \beta_{01}, \ldots, \beta_{0q})' \tag{8.3}$$

based on an observed finite realization $\{X_t; 1 \leq t \leq n\}$ of $\mathcal{X}$. Depending on the choice of a score function, these M-estimators are asymptotically normal under milder moment assumptions on the error distribution than the commonly used quasi-maximum likelihood estimator (QMLE). Mukherjee [8] further considered a class of weighted bootstrap methods to approximate the distributions of these estimators and established their asymptotic validity. In this paper, we discuss an iteratively re-weighted algorithm to compute these M-estimators and the corresponding bootstrap estimators with emphasis on the so-called Huber, μ-, and Cauchy estimators, which so far were not given much attention in the literature. This iteratively re-weighted algorithm turns out to be particularly useful in the computation of bootstrap replicates since it avoids re-evaluating some core quantities for each new bootstrap sample.

The class of M-estimators in Mukherjee [7] includes the (Gaussian) QMLE as a special case. The asymptotic normality of the QMLE and the asymptotic validity of bootstrapping it are well-known classical results which, however, require the existence of fourth-order moment of the error distribution. The same class also contains other less-known M-estimators, such as the μ-estimator and Cauchy estimator, which are asymptotically normal under milder moment assumptions and hence should be considered as attractive alternatives to the QMLE. One of the objectives of this paper is to study the performance of these estimators through simulations and use them for the empirical study of some interesting datasets.

In an earlier work, Muler and Yohai [9] analyzed the Electric Fuel Corporation (EFCX) time series by fitting a GARCH(1,1) model to the data. Using exploratory analysis, they detected the presence of outliers and considered the estimation of the GARCH parameters based on various robust methods. It turned out that the estimates based on different methods vary widely, which makes their study somewhat inconclusive as to which robust methods should be preferred in similar situations. In this paper, we show how M-estimators can be used for making such choices.

In a different direction, Francq and Zakoïan [4] stressed the importance of considering higher-order GARCH models such as the GARCH(2,1) in the context of financial data. Computational results and simulation studies for such models, however, are rather scarce in the literature. Our simulation study and empirical applications therefore include higher-order models such as GARCH(2,1) and GARCH(1,2).

The main contributions of the paper are as follows. We implement a very general algorithm for computing a variety of M-estimators and demonstrate their importance in the analysis of real data. We consider situations when the error distributions are possibly heavy-tailed or require a higher-order GARCH model fitting. We provide results and analysis of an extensive simulation study based on M-estimators which are asymptotically normal under weak moment assumptions on error distribution. Finally, we study the effectiveness of the bootstrap approximation of the distribution of M-estimators.

The rest of the paper is organized as follows. Sections 8.2 and 8.3 set the background. In particular, Sect. 8.2 considers the class of M-estimators and provides several examples. Section 8.3 contains the bootstrap formulation and its asymptotic validity. Section 8.4 discusses some of the computational aspects of M-estimators and their bootstrapped versions. Section 8.5 reports simulation results for various M-estimators. Section 8.6 compares the bootstrap approximations of M-estimators with the classical asymptotic normal approximation. Section 8.7 analyzes two real financial time series data.

8.2 M-Estimation of GARCH Parameters

8.2.1 A Class of M-Estimators

Throughout the paper, we write $\dot{g}$ for the derivative and $\dot{\mathbf{g}}$ for the gradient of a differentiable function g, $\text{sign}(x)$ for $I(x > 0) - I(x < 0)$, and $\log^+(x)$ for $I(x > 1)\log(x)$ when $x > 0$, where the symbol I stands for the indicator. Also, ϵ represents a generic random variable with the same distribution as the errors $\{\epsilon_t\}$ in (8.1).

Consider $H(x) := x\psi(x)$, $x \in \mathbb{R}$, where $\psi : \mathbb{R} \to \mathbb{R}$ is an odd and differentiable function at all but possibly a finite number of points; denote by $\mathcal{D} \subseteq \mathbb{R}$ the set of points where ψ is differentiable and by $\bar{\mathcal{D}}$ its complement. Since ψ is an odd function, H is an even function. Functions H of this type will be used as *score functions* in the M-estimation procedures described below. Examples are as follows.

Example 8.1

(i) QMLE score function: $\psi(x) = x$; $\bar{\mathcal{D}} = \phi$ (the empty set), $H(x) = x^2$.

(ii) Least absolute deviation (LAD) score function: $\psi(x) = \text{sign}\,(x)$; $\bar{\mathcal{D}} = \{0\}$, and $H(x) = |x|$.

(iii) Huber's k score function: $\psi(x) = xI(|x| \leq k) + k\,\text{sign}\,(x)I(|x| > k)$, where $k > 0$ is a known constant ($\bar{\mathcal{D}} = \{-k, k\}$), and

$$H(x) = x^2 I(|x| \leq k) + k|x|I(|x| > k).$$

(iv) Maximum likelihood score function: $\psi(x) = -\dot{f}(x)/f(x)$, where f is the actual density of ϵ, assumed to be known, and $H(x) = x\{-\dot{f}(x)/f(x)\}$.

(v) μ score function: $\psi(x) = \mu\,\text{sign}(x)/(1 + |x|)$, where $\mu > 1$ is a known constant ($\bar{\mathcal{D}} = \{0\}$), and $H(x) = \mu|x|/(1 + |x|)$ (a bounded score function).

(vi) Cauchy score function: $\psi(x) = 2x/(1 + x^2)$, $H(x) = 2x^2/(1 + x^2)$ (a bounded score function).

(vii) Exponential pseudo-maximum likelihood score function:

$$\psi(x) = \delta_1 |x|^{\delta_2 - 1}\text{sign}(x),$$

where $\delta_1 > 0$ and $1 < \delta_2 \leq 2$ are known constants; $\bar{\mathcal{D}} = \{0\}$, and $H(x) = \delta_1 |x|^{\delta_2}$.

Assume that for some $\kappa_1 \geq 2$ and $\kappa_2 > 0$,

$$\text{E}[|\epsilon|^{\kappa_1}] < \infty \quad \text{and} \quad \lim_{t \to 0} \text{P}[\epsilon^2 < t]/t^{\kappa_2} = 0. \tag{8.4}$$

Then σ_t^2 (see (8.2)) admits the unique almost sure representation

$$\sigma_t^2 = c_0 + \sum_{j=1}^{\infty} c_j X_{t-j}^2, \quad t \in \mathbb{Z}, \tag{8.5}$$

where $\{c_j;\ j \geq 0\}$ are defined in Berkes et al. ([1]; (2.9)–(2.16)). Let $\boldsymbol{\Theta}$ be a compact subset of $(0, \infty)^{1+p} \times (0, 1)^q$ and denote by $\boldsymbol{\theta} = (\omega, \alpha_1, \ldots, \alpha_p, \beta_1, \ldots, \beta_q)'$ a typical element in $\boldsymbol{\Theta}$. Define the *variance function* $v_t : \boldsymbol{\Theta} \to \mathbb{R}^+$ by

$$v_t(\boldsymbol{\theta}) = c_0(\boldsymbol{\theta}) + \sum_{j=1}^{\infty} c_j(\boldsymbol{\theta}) X_{t-j}^2, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta},\ t \in \mathbb{Z}, \tag{8.6}$$

where the coefficients $\{c_j(\boldsymbol{\theta});\ j \geq 0\}$ are such that, for $\boldsymbol{\theta} = \boldsymbol{\theta}_0$,

$$c_j(\boldsymbol{\theta}_0) = c_j, \quad j \geq 0 \tag{8.7}$$

(Berkes et al. ([1]; (3.1))). Hence the variance function satisfies $v_t(\boldsymbol{\theta}_0) = \sigma_t^2$, $t \in \mathbb{Z}$ and (8.1) can be rewritten as

$$X_t = \{v_t(\boldsymbol{\theta}_0)\}^{1/2}\epsilon_t, \quad 1 \le t \le n. \tag{8.8}$$

Let H denote a score function. The M-estimators are defined as the solutions $\hat{\boldsymbol{\theta}}_n$ of $\widehat{\boldsymbol{M}}_{n,H}(\boldsymbol{\theta}) = \mathbf{0}$, where

$$\widehat{\boldsymbol{M}}_{n,H}(\boldsymbol{\theta}) := \sum_{t=1}^{n} \left\{1 - H\{X_t/\hat{v}_t^{1/2}(\boldsymbol{\theta})\}\right\}\{\dot{\hat{\boldsymbol{v}}}_t(\boldsymbol{\theta})/\hat{v}_t(\boldsymbol{\theta})\} \tag{8.9}$$

and

$$\hat{v}_t(\boldsymbol{\theta}) := c_0(\boldsymbol{\theta}) + I(2 \le t)\sum_{j=1}^{t-1} c_j(\boldsymbol{\theta})X_{t-j}^2, \quad \boldsymbol{\theta} \in \boldsymbol{\Theta}, \ 1 \le t \le n \tag{8.10}$$

is the observable approximation of the variance function $v_t(\boldsymbol{\theta})$ defined in (8.6).

The recursive nature of the coefficients $\{c_j(\boldsymbol{\theta})\}$ greatly simplifies the computation of M-estimators, as discussed in Sect. 8.4. For $p, q = 1$ or 2, these coefficients, for instance, satisfy the following recursions.

Example 8.2

(i) GARCH(1,1) model: with $\boldsymbol{\theta} = (\omega, \alpha, \beta)'$,

$$c_0(\omega, \alpha, \beta) = \omega/(1-\beta), \ \ c_j(\omega, \alpha, \beta) = \alpha\beta^{j-1}, \ \ j \ge 1.$$

(ii) GARCH(2,1) model: with $\boldsymbol{\theta} = (\omega, \alpha_1, \alpha_2, \beta)'$,

$$c_0(\boldsymbol{\theta}) = \omega/(1-\beta), \ \ c_1(\boldsymbol{\theta}) = \alpha_1, \ \ c_2(\boldsymbol{\theta}) = \alpha_2 + \beta c_1(\boldsymbol{\theta}) = \alpha_2 + \beta\alpha_1,$$

and

$$c_j(\boldsymbol{\theta}) = \beta c_{j-1}(\boldsymbol{\theta}), \quad j \ge 3.$$

(iii) GARCH(1, 2) model: with $\boldsymbol{\theta} = (\omega, \alpha, \beta_1, \beta_2)'$,

$$c_0(\boldsymbol{\theta}) = \omega/(1-\beta_1-\beta_2), \ \ c_1(\boldsymbol{\theta}) = \alpha, \ \ c_2(\boldsymbol{\theta}) = \beta_1 c_1(\boldsymbol{\theta}) = \beta_1\alpha,$$

and

$$c_j(\boldsymbol{\theta}) = \beta_1 c_{j-1}(\boldsymbol{\theta}) + \beta_2 c_{j-2}(\boldsymbol{\theta}), \quad j \ge 3.$$

(iv) GARCH(2,2) model: with $\boldsymbol{\theta} = (\omega, \alpha_1, \alpha_2, \beta_1, \beta_2)'$,

$$c_0(\boldsymbol{\theta}) = \omega/(1-\beta_1-\beta_2), \ \ c_1(\boldsymbol{\theta}) = \alpha_1, \ \ c_2(\boldsymbol{\theta}) = \alpha_2 + \beta_1\alpha_1,$$

and

$$c_j(\boldsymbol{\theta}) = \beta_1 c_{j-1}(\boldsymbol{\theta}) + \beta_2 c_{j-2}(\boldsymbol{\theta}), \quad j \geq 3.$$

8.2.2 Asymptotic Distribution of M-Estimators

The asymptotic distribution of M-estimators is derived under the following assumptions.

Assumption 8.1 (*Model assumptions*) The parameter space $\boldsymbol{\Theta}$ is compact and $\boldsymbol{\theta}_0$ defined in (8.3) belongs to its interior; (8.4), (8.6), and (8.8) hold; $\{X_t\}$ is stationary and ergodic.

Assumption 8.2 (*Score function assumptions*)
(8.2.1) Associated with the score function H, there exists a unique number $c_H > 0$ such that

$$\mathrm{E}[H(\epsilon/c_H^{1/2})] = 1, \;\; \mathrm{E}[H(\epsilon/c_H^{1/2})]^2 < \infty, \;\; \text{and} \;\; 0 < \mathrm{E}\{(\epsilon/c_H^{1/2})\dot{H}(\epsilon/c_H^{1/2})\} < \infty; \tag{8.11}$$

and the transformed parameter

$$\boldsymbol{\theta}_{0H} := (c_H\omega_0, c_H\alpha_{01}, \ldots, c_H\alpha_{0p}, \beta_{01}, \ldots, \beta_{0q})' \tag{8.12}$$

is in the interior of $\boldsymbol{\Theta}$.
(8.2.2) (Smoothness conditions)[1]

(i) There exists a function L satisfying

$$|H(es) - H(e)| \leq L(e)|s^2 - 1|, \; e \in \mathbb{R}^1, \; s > 0,$$

where $\mathrm{E}\log^+\{L(\epsilon/c_H^{1/2})\} < \infty$;
(ii) there exists a function Λ such that for $e \in \mathbb{R}^1, \; s > 0, es, e \in \mathcal{D}$,

$$|\dot{H}(es) - \dot{H}(e)| \leq \Lambda(e)|s - 1|,$$

where $\mathrm{E}\{|\epsilon/c_H^{1/2}|\Lambda(\epsilon/c_H^{1/2})\} < \infty$;
(iii) there exists a function Λ^* satisfying

$$|\Lambda(e + es) - \Lambda(e)| \leq \Lambda^*(e)s, \; e \in \mathbb{R}^1, \; s > 0,$$

where $\mathrm{E}\log^+\{\Lambda^*(\epsilon/c_H^{1/2})\} < \infty$.

Define the *score function factor*

[1] These conditions are trivially satisfied by all the examples of score functions H considered above.

Table 8.1 Values of c_H for various M-estimators (Huber, μ-, Cauchy) under normal, double exponential, logistic, $t(3)$, and $t(2.2)$ error distributions, where $t(d)$ is the t-distribution with d degrees of freedom

	Huber	μ-estimator	Cauchy
Normal	0.825	1.692	0.377
Double exponential	0.677	1.045	0.207
Logistic	0.781	1.487	0.316
$t(3)$	0.533	0.850	0.172
$t(2.2)$	0.204	0.274	0.053

$$\sigma^2(H) := 4\ \mathrm{Var}\{H(\epsilon/c_H^{1/2})\}/[\mathrm{E}\{(\epsilon/c_H^{1/2})\dot{H}(\epsilon/c_H^{1/2})\}]^2,$$

and the matrix

$$\boldsymbol{G} := \mathrm{E}\{\dot{\boldsymbol{v}}_1(\boldsymbol{\theta}_{0H})\dot{\boldsymbol{v}}_1'(\boldsymbol{\theta}_{0H})/v_1^2(\boldsymbol{\theta}_{0H})\},$$

the following result on the asymptotic distribution of the M-estimator $\hat{\boldsymbol{\theta}}_n$ defined in (8.9) and (8.10) holds (Mukherjee [7]).

Theorem 8.1 *Suppose that Assumptions 8.1 and 8.2 hold. Then, as $n \to \infty$,*

$$n^{1/2}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_{0H}) \to \mathcal{N}(\mathbf{0}, \sigma^2(H)\boldsymbol{G}^{-1}).$$

Remark 8.1 The coefficients c_H in Assumption (8.2.1) take the values $c_H = \mathrm{E}(\epsilon^2)$ for the QMLE and $c_H = (\mathrm{E}|\epsilon|)^2$ for the LAD. For the Huber, μ-, Cauchy, and other scores, c_H does not have a closed-form expression but the corresponding numerical values can be computed from (8.11) for various error distributions as follows. Fix a large positive integer I and generate $\{\epsilon_i; 1 \le i \le I\}$ from the error distribution. Then, using the bisection method on $c > 0$, solve the equation

$$(1/I)\sum_{i=1}^{I}\left\{H\left(\epsilon_i/c^{1/2}\right)\right\} - 1 = 0.$$

In Table 8.1, we provide c_H for some further error distributions and score functions such as Huber's k-score with $k = 1.5$ and the μ-estimator with $\mu = 3$, which are used in simulations and data analysis in subsequent sections.

8.3 Bootstrapping M-Estimators

Let $\{w_{nt}; 1 \le t \le n, n \ge 1\}$ be a triangular array of random variables such that

(i) for each $n \ge 1$, $\{w_{nt}; 1 \le t \le n\}$ are exchangeable and independent of $\{X_t; t \ge 1\}$ and $\{\epsilon_t; t \ge 1\}$, and

(ii) $w_{nt} \geq 0$ and $\mathrm{E}(w_{nt}) = 1$ for all $t \geq 1$.

Based on these weights w_{nt}, a bootstrap estimator $\hat{\boldsymbol{\theta}}_{*n}$ is defined as a solution of $\widehat{\boldsymbol{M}}^*_{n,H}(\boldsymbol{\theta}) = \mathbf{0}$, where

$$\widehat{\boldsymbol{M}}^*_{n,H}(\boldsymbol{\theta}) := \sum_{t=1}^{n} w_{nt} \left\{ 1 - H\{X_t/\hat{v}_t^{1/2}(\boldsymbol{\theta})\} \right\} \{\dot{\hat{\boldsymbol{v}}}_t(\boldsymbol{\theta})/\hat{v}_t(\boldsymbol{\theta})\}. \tag{8.13}$$

From various available choices of bootstrap weights, we consider, for the sake of comparison, the following three bootstrapping schemes.

Scheme M. The sequence $\{w_{n1}, \ldots, w_{nn}\}$ has a multinomial $(n, 1/n, \ldots, 1/n)$ distribution, which essentially yields the classical paired bootstrap.

Scheme E. The weights are of the form $w_{nt} = (nE_t)/\sum_{i=1}^{n} E_i$, where $\{E_t\}$ are i.i.d. exponential with mean 1.

Scheme U. The weights are of the form $w_{nt} = (nU_t)/\sum_{i=1}^{n} U_i$, where $\{U_t\}$ are i.i.d. uniform on (0.5, 1.5).

A large number of other bootstrap methods in the literature are special cases of the above formulation. Such general formulation of weighted bootstrap offers a unified way of studying several bootstrap schemes simultaneously. See, for instance, Chatterjee and Bose [3] for details in different contexts.

We assume that the weights satisfy the basic conditions

$$\mathrm{E}(w_{n1}) = 1, \;\; 0 < k_3 < \sigma_n^2 = o(n), \;\; \text{and } \mathrm{Corr}\,(w_{n1}, w_{n2}) = O(1/n), \tag{8.14}$$

where $\sigma_n^2 = \mathrm{Var}(w_{ni})$ and $k_3 > 0$ is a constant (Chatterjee and Bose ([3]; Conditions BW)). We also assume additional smoothness and moment conditions.

Assumption 8.3 $H(x)$ is twice differentiable at all but a finite number of points and for some $\delta > 2$, $\mathrm{E}[H(\epsilon/c_H^{1/2})]^{\delta} < \infty$.

Under Assumptions 8.1–8.3 and (8.14), the weighted bootstrap is asymptotically valid (Mukherjee [8]).

Theorem 8.2 *Suppose that Assumptions 8.1–8.3 and (8.14) hold. Then, for almost all data,*[2] *as $n \to \infty$,*

$$\sigma_n^{-1} n^{1/2}(\hat{\boldsymbol{\theta}}_{*n} - \hat{\boldsymbol{\theta}}_n) \to \mathcal{N}(\mathbf{0}, \sigma^2(H)\boldsymbol{G}^{-1}). \tag{8.15}$$

Since $0 < 1/\sigma_n < 1/\sqrt{k_3}$, the probability of convergence of the bootstrap estimator is the same as that of the original M-estimator, although the scaling σ_n^{-1} reflects the impact of the chosen weights.

The distributional result of (8.15) is useful for constructing confidence intervals for the GARCH parameters. Let B be the number of bootstrap replicates. Consider the true value γ_0 of a generic parameter (either ω_0, α_{0i}, or β_{0j}), and let $\hat{\gamma}_n$

[2] Namely, outside a set of probability zero in the bootstrap distribution induced by the data.

and $\hat{\gamma}_{*nb}$ denote its M-estimator and b-th bootstrap estimator ($1 \le b \le B$), respectively. Let γ_{0H} denote the corresponding transformed parameter (either $c_H\omega_0$, $c_H\alpha_{0i}$, or β_{0j}; see (8.12)); the value of this γ_{0H} is known in simulation experiments.

Using the approximation of $\sqrt{n}(\hat{\gamma}_n - \gamma_{0H})$ by $\sigma_n^{-1}n^{1/2}(\hat{\gamma}_{*n} - \hat{\gamma}_n)$, the bootstrap confidence interval (with level $1-\alpha$) for γ_{0H} is of the form

$$\Big[\hat{\gamma}_n - n^{-1/2}\{\sigma_n^{-1}n^{1/2}(\hat{\gamma}_{*n,\alpha/2} - \hat{\gamma}_n)\},\ \hat{\gamma}_n + n^{-1/2}\{\sigma_n^{-1}n^{1/2}(\hat{\gamma}_{*n,1-\alpha/2} - \hat{\gamma}_n)\}\Big], \tag{8.16}$$

where $\hat{\gamma}_{*n,\beta}$ is the β-th quantile of the numbers $\{\hat{\gamma}_{*nb}, 1 \le b \le B\}$. Consequently, the bootstrap coverage probability is evaluated by the proportion of intervals of the form (8.16) containing γ_{0H}.

The asymptotic normality result of Theorem 8.1 also yields a confidence interval

$$\Big[\hat{\gamma}_n - n^{-1/2}\hat{d}z_{1-\alpha/2},\ \hat{\gamma}_n + n^{-1/2}\hat{d}z_{1-\alpha/2}\Big] \tag{8.17}$$

(with level $1-\alpha$) for γ_{0H}, where $\hat{d}^2$ is the estimated variance of $\hat{\gamma}_n$ obtained as the appropriate diagonal entry of the estimator of $\sigma^2(H)\boldsymbol{G}^{-1}$ and $z_{1-\alpha/2}$ is the quantile of order $(1-\alpha/2)$ of the standard normal distribution. Call it the *normal confidence interval*.

In Sect. 8.6, we compare the accuracy of the bootstrap-based and normal confidence intervals (8.16) and (8.17).

8.4 Computational Issues

This section discusses, in detail, the implementation of an iteratively re-weighted algorithm for the computation of M-estimators proposed in Mukherjee [8]. In particular, we highlight the μ- and Cauchy estimators, since their asymptotic distributions are derived under mild moment assumptions. We also consider the bootstrap estimators based on the corresponding score functions.

8.4.1 Computation of the M-Estimators

For notational convenience, let $\alpha(c) := \mathrm{E}[H(c\epsilon)]$ for $c > 0$. Using a Taylor expansion of $\widehat{\boldsymbol{M}}_{n,H}$, we obtain the updating equation

$$\tilde{\boldsymbol{\theta}} = \check{\boldsymbol{\theta}} + \{\dot{\alpha}(1)/2\}^{-1}\Big[\sum_{t=1}^{n} \dot{\boldsymbol{v}}_t(\check{\boldsymbol{\theta}})\dot{\boldsymbol{v}}_t'(\check{\boldsymbol{\theta}})/\hat{v}_t^2(\check{\boldsymbol{\theta}})\Big]^{-1} \times \sum_{t=1}^{n} \Big\{H\{X_t/\hat{v}_t^{1/2}(\check{\boldsymbol{\theta}})\} - 1\Big\}\{\dot{\boldsymbol{v}}_t(\check{\boldsymbol{\theta}})/\hat{v}_t(\check{\boldsymbol{\theta}}\}, \quad (8.18)$$

where $\dot{\alpha}(1) = \mathrm{E}\{\epsilon \dot{H}(\epsilon)\}$ (this expectation exists under the smoothness conditions in Assumption 8.2) and $\tilde{\boldsymbol{\theta}}$ is the updated estimator of $\hat{\boldsymbol{\theta}}_n$ as a function of the current one $\check{\boldsymbol{\theta}}$, say. We now discuss two aspects regarding the implementation of the algorithm in (8.18). First, the initial value of $\check{\boldsymbol{\theta}}$ for the iteration, in principle, should be a $\sqrt{n}$-consistent estimator of $\boldsymbol{\theta}_{0H}$. However, we observe in our extensive simulation study that, irrespective of the choice of the QMLE, LAD, $\boldsymbol{\theta}_0$ or even values very different from $\boldsymbol{\theta}_{0H}$ as initial estimates, only few iterations are needed for the convergence to the same estimates. Second, we cannot, in general, estimate $\dot{\alpha}(1)$ from the data using the GARCH residuals $\{X_t/\hat{v}_t^{1/2}(\hat{\boldsymbol{\theta}}_n)\}$ as they are close to $\{\epsilon_t/c_H^{1/2}\}$, an unknown multiplicative factor of the errors. Therefore, we use ad-hoc techniques such as simulating $\{\tilde{\epsilon}_t;\, 1 \le t \le n\}$ from $\mathcal{N}(0, 1)$ or standardized double exponential distributions and then use $n^{-1}\sum_{t=1}^{n} \tilde{\epsilon}\dot{H}(\tilde{\epsilon})$ to carry out the iterations. Note that if the iterations in (8.18) converge, then $\tilde{\boldsymbol{\theta}} - \check{\boldsymbol{\theta}} \approx \mathbf{0}$, hence $\widehat{\boldsymbol{M}}_{n,H}(\check{\boldsymbol{\theta}}) \approx \mathbf{0}$, and $\tilde{\boldsymbol{\theta}}$ is the desired $\hat{\boldsymbol{\theta}}_n$. Based on our extensive simulation study and real data analysis, this algorithm appears to be robust enough to converge to the same value of $\hat{\boldsymbol{\theta}}_n$ irrespective of the evaluations of the unknown value of $\dot{\alpha}(1)$ used in the computation.

In the following examples, we discuss (8.18) when specialized to the M-estimators computed in this paper.

(a) **QMLE**. Let $H(x) = x^2$ and $\alpha(c) = c^2\mathrm{E}(\epsilon^2)$. Hence $\dot{\alpha}(1)/2 = \mathrm{E}(\epsilon^2)$ and (8.18) takes the form

$$\tilde{\boldsymbol{\theta}} = \check{\boldsymbol{\theta}} + \Big\{\mathrm{E}(\epsilon^2)\Big\}^{-1}\Big[\sum_{t=1}^{n} \Big\{\dot{\boldsymbol{v}}_t(\check{\boldsymbol{\theta}})\dot{\boldsymbol{v}}_t'(\check{\boldsymbol{\theta}})/\hat{v}_t^2(\check{\boldsymbol{\theta}})\Big\}\Big]^{-1} \times \sum_{t=1}^{n} \Big[\{X_t^2/\hat{v}_t(\check{\boldsymbol{\theta}})\} - 1\Big]\{\dot{\boldsymbol{v}}_t(\check{\boldsymbol{\theta}})/\hat{v}_t(\check{\boldsymbol{\theta}})\}.$$

With $W_t = 1/\hat{v}_t^2(\tilde{\boldsymbol{\theta}}_{(r)})$, $x_t = \dot{\boldsymbol{v}}_t(\tilde{\boldsymbol{\theta}}_{(r)})$, and $y_t = X_t^2 - \hat{v}_t(\tilde{\boldsymbol{\theta}}_{(r)})$, we compute (iteration $r + 1$)

$$\tilde{\boldsymbol{\theta}}_{(r+1)} = \tilde{\boldsymbol{\theta}}_{(r)} + \Big\{\mathrm{E}(\epsilon^2)\Big\}^{-1}\Big\{\sum_t W_t x_t x_t'\Big\}^{-1}\Big\{\sum_t W_t x_t y_t\Big\}.$$

Note that when $\mathrm{E}(\epsilon^2) = 1$, this coincides with the formula obtained through the BHHH algorithm proposed by Berndt et al. [2].

(b) **LAD.** Let $H(x) = |x|$ and $\alpha(c) = c\mathrm{E}|\epsilon|$. Hence $\dot{\alpha}(1) = \mathrm{E}|\epsilon|$ and (8.18) takes the form

$$\begin{aligned}
\tilde{\boldsymbol{\theta}} &= \check{\boldsymbol{\theta}} + \{2/\mathrm{E}|\epsilon|\}\Big[\sum_{t=1}^{n}\Big\{\dot{\hat{\boldsymbol{v}}}_t(\check{\boldsymbol{\theta}}))\dot{\hat{\boldsymbol{v}}}_t'(\check{\boldsymbol{\theta}}))/\hat{v}_t^2(\check{\boldsymbol{\theta}}))\Big\}\Big]^{-1} \\
&\qquad \times \sum_{t=1}^{n}\Big[|X_t|/\hat{v}_t^{1/2}(\check{\boldsymbol{\theta}})) - 1\Big]\{\dot{\hat{\boldsymbol{v}}}_t(\check{\boldsymbol{\theta}}))/\hat{v}_t(\check{\boldsymbol{\theta}}))\} \\
&= \check{\boldsymbol{\theta}} + \{2/\mathrm{E}|\epsilon|\}\Big[\sum_{t=1}^{n}\Big\{\dot{\hat{\boldsymbol{v}}}_t(\check{\boldsymbol{\theta}})\dot{\hat{\boldsymbol{v}}}_t'(\check{\boldsymbol{\theta}})/\hat{v}_t^2(\check{\boldsymbol{\theta}})\Big\}\Big]^{-1} \\
&\qquad \times \sum_{t=1}^{n}\Big\{|X_t| - \hat{v}_t^{1/2}(\check{\boldsymbol{\theta}})\Big\}\{\dot{\hat{\boldsymbol{v}}}_t(\check{\boldsymbol{\theta}})/\hat{v}_t^{3/2}(\check{\boldsymbol{\theta}})\} \\
&= \check{\boldsymbol{\theta}} + \{2/\mathrm{E}|\epsilon|\}\Big[\sum_{t=1}^{n}\Big\{\dot{\hat{\boldsymbol{v}}}_t(\check{\boldsymbol{\theta}})\dot{\hat{\boldsymbol{v}}}_t'(\check{\boldsymbol{\theta}})/\hat{v}_t^2(\check{\boldsymbol{\theta}})\Big\}\Big]^{-1} \\
&\qquad \times \sum_{t=1}^{n}\Big\{\hat{v}_t^{1/2}(\check{\boldsymbol{\theta}})(|X_t| - \hat{v}_t^{1/2}(\check{\boldsymbol{\theta}}))\Big\}\{\dot{\hat{\boldsymbol{v}}}_t(\check{\boldsymbol{\theta}})/\hat{v}_t^2(\check{\boldsymbol{\theta}})\}.
\end{aligned}$$

With $W_t = 1/\hat{v}_t^2(\tilde{\boldsymbol{\theta}}_{(r)})$, $x_t = \dot{\hat{\boldsymbol{v}}}_t(\tilde{\boldsymbol{\theta}}_{(r)})$, and $y_t = \hat{v}_t^{1/2}(\tilde{\boldsymbol{\theta}}_{(r)})(|X_t| - \hat{v}_t^{1/2}(\tilde{\boldsymbol{\theta}}_{(r)}))$, we compute (iteration $r+1$)

$$\tilde{\boldsymbol{\theta}}_{(r+1)} = \tilde{\boldsymbol{\theta}}_{(r)} + \{2/\mathrm{E}|\epsilon|\}\left\{\sum_t W_t x_t x_t'\right\}^{-1}\left\{\sum_t W_t x_t y_t\right\}.$$

(c) **Huber.** Let $H(x) = x^2 I(|x| \le k) + k|x|I(|x| > k)$ and

$$\alpha(c) = \mathrm{E}\left[(c\epsilon)^2 I(|c\epsilon| \le k) + k|c\epsilon|I(|c\epsilon| > k)\right].$$

Hence

$$\dot{\alpha}(1) = \mathrm{E}\left[2\epsilon^2 I(|\epsilon| \le k) + k|\epsilon|I(|\epsilon| > k)\right]$$

and (8.18) takes the form

$$\begin{aligned}
\tilde{\boldsymbol{\theta}} &= \check{\boldsymbol{\theta}} - \Big\{\dot{\alpha}(1)/2\Big\}^{-1}\Big[\sum_{t=1}^{n}\Big\{\frac{\dot{\hat{\boldsymbol{v}}}_t(\check{\boldsymbol{\theta}})\dot{\hat{\boldsymbol{v}}}_t'(\check{\boldsymbol{\theta}})}{\hat{v}_t^2(\check{\boldsymbol{\theta}})}\Big\}\Big]^{-1} \\
&\quad \times \sum_{t=1}^{n}\left[1 - \frac{X_t^2}{\hat{v}_t(\check{\boldsymbol{\theta}})} I\left(\frac{|X_t|}{\hat{v}_t^{1/2}(\check{\boldsymbol{\theta}})} \le k\right) - k\frac{|X_t|}{\hat{v}_t^{1/2}(\check{\boldsymbol{\theta}})} I\left(\frac{|X_t|}{\hat{v}_t^{1/2}(\check{\boldsymbol{\theta}})} > k\right)\right]\left\{\frac{\dot{\hat{\boldsymbol{v}}}_t(\check{\boldsymbol{\theta}})}{\hat{v}_t(\check{\boldsymbol{\theta}})}\right\}.
\end{aligned}$$

With $W_t = 1/\hat{v}_t^2(\tilde{\boldsymbol{\theta}}_{(r)})$, $x_t = \dot{\boldsymbol{v}}_t(\tilde{\boldsymbol{\theta}}_{(r)})$ and

$$y_t = X_t^2 I\left(|X_t|/\hat{v}_t^{1/2}(\tilde{\boldsymbol{\theta}}_{(r)}) \le k\right) + k|X_t|\hat{v}_t^{1/2}(\tilde{\boldsymbol{\theta}}_{(r)}) I\left(|X_t|/\hat{v}_t^{1/2}(\tilde{\boldsymbol{\theta}}_{(r)}) > k\right) - \hat{v}_t(\tilde{\boldsymbol{\theta}}_{(r)}),$$

we compute (iteration $r+1$)

$$\tilde{\boldsymbol{\theta}}_{(r+1)} = \tilde{\boldsymbol{\theta}}_{(r)} + \left\{\dot{\alpha}(1)/2\right\}^{-1} \left\{\sum_t W_t x_t x_t'\right\}^{-1} \left\{\sum_t W_t x_t y_t\right\}.$$

(d) **$\boldsymbol{\mu}$-estimator.** Let $H(x) = \mu|x|/(1+|x|)$ and $\alpha(c) = \mu - \mu\mathrm{E}\,[1/(1+|c\epsilon|)]$. Hence

$$\dot{\alpha}(1) = \mu\mathrm{E}\left[|\epsilon|/(1+|\epsilon|)^2\right]$$

and (8.18) takes the form

$$\tilde{\boldsymbol{\theta}} = \check{\boldsymbol{\theta}} + \left\{\frac{\mu}{2}\mathrm{E}\left[\frac{|\epsilon|}{(1+|\epsilon|)^2}\right]\right\}^{-1} \times \Big[\sum_{t=1}^n \Big\{\frac{\dot{\boldsymbol{v}}_t(\check{\boldsymbol{\theta}})\dot{\boldsymbol{v}}_t'(\check{\boldsymbol{\theta}})}{\hat{v}_t^2(\check{\boldsymbol{\theta}})}\Big\}\Big]^{-1} \sum_{t=1}^n \left[\frac{\mu|X_t|}{\hat{v}_t^{1/2}(\check{\boldsymbol{\theta}}) + |X_t|} - 1\right] \left\{\frac{\dot{\boldsymbol{v}}_t(\check{\boldsymbol{\theta}})}{\hat{v}_t(\check{\boldsymbol{\theta}})}\right\}.$$

With $W_t = 1/\hat{v}_t^2(\tilde{\boldsymbol{\theta}}_{(r)})$, $x_t = \dot{\boldsymbol{v}}_t(\tilde{\boldsymbol{\theta}}_{(r)})$, and $y_t = \dfrac{\mu|X_t|\hat{v}_t(\tilde{\boldsymbol{\theta}}_{(r)})}{\hat{v}_t^{1/2}(\tilde{\boldsymbol{\theta}}_{(r)}) + |X_t|} - \hat{v}_t(\tilde{\boldsymbol{\theta}}_{(r)})$, we compute (iteration $r+1$)

$$\tilde{\boldsymbol{\theta}}_{(r+1)} = \tilde{\boldsymbol{\theta}}_{(r)} + \left\{\frac{\mu}{2}\mathrm{E}\left[\frac{|\epsilon|}{(1+|\epsilon|)^2}\right]\right\}^{-1} \left\{\sum_t W_t x_t x_t'\right\}^{-1} \left\{\sum_t W_t x_t y_t\right\}.$$

(e) **Cauchy estimator.** Let $H(x) = 2x^2/(1+x^2)$ and $\alpha(c) = \mathrm{E}\left[2c^2\epsilon^2/(1+c^2\epsilon^2)\right]$. Hence

$$\dot{\alpha}(1) = \mathrm{E}\left[4\epsilon^2/(1+\epsilon^2)^2\right]$$

and

$$\tilde{\boldsymbol{\theta}} = \check{\boldsymbol{\theta}} - \left\{2\mathrm{E}\left[\frac{\epsilon^2}{(1+\epsilon^2)^2}\right]\right\}^{-1} \Big[\sum_{t=1}^n \Big\{\frac{\dot{\boldsymbol{v}}_t(\check{\boldsymbol{\theta}})\dot{\boldsymbol{v}}_t'(\check{\boldsymbol{\theta}})}{\hat{v}_t^2(\check{\boldsymbol{\theta}})}\Big\}\Big]^{-1} \sum_{t=1}^n \left[1 - \frac{2X_t^2}{\hat{v}_t(\check{\boldsymbol{\theta}}) + X_t^2}\right] \left\{\frac{\dot{\boldsymbol{v}}_t(\check{\boldsymbol{\theta}})}{\hat{v}_t(\check{\boldsymbol{\theta}})}\right\}.$$

With $W_t = 1/\hat{v}_t^2(\tilde{\boldsymbol{\theta}}_{(r)})$, $x_t = \dot{\boldsymbol{v}}_t(\tilde{\boldsymbol{\theta}}_{(r)})$, and $y_t = \dfrac{2X_t^2\hat{v}_t(\tilde{\boldsymbol{\theta}}_{(r)})}{\hat{v}_t(\tilde{\boldsymbol{\theta}}_{(r)}) + X_t^2} - \hat{v}_t(\tilde{\boldsymbol{\theta}}_{(r)})$, we compute (iteration $r+1$)

$$\tilde{\boldsymbol{\theta}}_{(r+1)} = \tilde{\boldsymbol{\theta}}_{(r)} + \left\{2\mathrm{E}\left[\frac{\epsilon^2}{(1+\epsilon^2)^2}\right]\right\}^{-1} \left\{\sum_t W_t x_t x_t'\right\}^{-1} \left\{\sum_t W_t x_t y_t\right\}.$$

8.4.2 Computation of the Bootstrap M-Estimators

The relevant function here is $\widehat{\boldsymbol{M}}^*_{n,H}(\boldsymbol{\theta})$ defined in (8.13) and the bootstrap estimator $\hat{\boldsymbol{\theta}}_{*n}$ can be computed from the current one $\check{\boldsymbol{\theta}}_*$, say, using the updating equation

$$\tilde{\boldsymbol{\theta}}_* = \check{\boldsymbol{\theta}}_* - \{2/\dot{\alpha}(1)\}\Big[\sum_{t=1}^{n} w_{nt}\Big\{\dot{\hat{\boldsymbol{v}}}_t(\check{\boldsymbol{\theta}}_*)\dot{\hat{\boldsymbol{v}}}_t'(\check{\boldsymbol{\theta}}_*)/\hat{v}_t^2(\check{\boldsymbol{\theta}}_*)\Big\}\Big]^{-1}$$
$$\times \sum_{t=1}^{n} w_{nt}\Big\{1 - H\{X_t/\hat{v}_t^{1/2}(\check{\boldsymbol{\theta}}_*)\}\Big\}\{\dot{\hat{\boldsymbol{v}}}_t(\check{\boldsymbol{\theta}}_*)/\hat{v}_t(\check{\boldsymbol{\theta}}_*)\}, \qquad (8.19)$$

where the M-estimator $\hat{\boldsymbol{\theta}}_n$ obtained via the iteration process (8.18) is chosen as the initial value.

We remark that the weighted bootstrap is computationally more friendly and easy-to-implement than the commonly-applied residual bootstrap (see, e.g., Jeong [5]) for GARCH models, since it avoids the computation of residuals at each iteration. In particular, one simply needs to generate weights once to compute a bootstrap estimator.

8.5 Monte Carlo Comparison of Performance

To compare the finite-sample performance of various M-estimators via their biases and mean squared errors (MSEs), we simulate n observations from GARCH models with specific choices of parameters and error distributions and compute the resulting M-estimators based on various score functions. This procedure is replicated R times to enable the estimation of the bias and MSE. For instance, with $p = 1 = q$, let $\hat{\boldsymbol{\theta}}_n = (\hat{\omega}_r, \hat{\alpha}_r, \hat{\beta}_r)'$ be the M-estimator of $\boldsymbol{\theta}_0 = (\omega_0, \alpha_{01}, \beta_{01})'$ based on the score function H at the r-th replication, $1 \le r \le R$. However, $(\hat{\omega}_r, \hat{\alpha}_r, \hat{\beta}_r)$ is a consistent estimator of $(c_H\omega_0, c_H\alpha_0, \beta_0)$, where c_H depends on the score function and the underlying error distribution (which are known in a simulation scenario). Therefore, we compare the performance at a specified error distribution across various score functions in terms of the *adjusted bias* and *adjusted MSE* defined by

$$\mathrm{E}(\hat{\omega}/c_H - \omega_0), \quad \mathrm{E}(\hat{\alpha}/c_H - \alpha_0), \quad \mathrm{E}(\hat{\beta} - \beta_0)$$

and

$$\mathrm{E}(\hat{\omega}/c_H - \omega_0)^2, \;\; \mathrm{E}(\hat{\alpha}/c_H - \alpha_0)^2, \;\; \mathrm{E}(\hat{\beta} - \beta_0)^2.$$

We consider R replicates of

$$(\hat{\omega}_r/c_H - \omega_0, \hat{\alpha}_r/c_H - \alpha_0, \hat{\beta}_r - \beta_0)'$$

and use the averaged quantities

$$R^{-1}\sum_{r=1}^{R}\{\hat{\omega}_r/c_H - \omega_0\}, \;\; R^{-1}\sum_{r=1}^{R}\{\hat{\alpha}_r/c_H - \alpha_0\}, \;\; R^{-1}\sum_{r=1}^{R}\{\hat{\beta}_r - \beta_0\} \qquad (8.20)$$

to estimate the *adjusted biases* and, for the *adjusted MSEs*,

$$R^{-1}\sum_{r=1}^{R}\{\hat{\omega}_r/c_H - \omega_0\}^2, \;\; R^{-1}\sum_{r=1}^{R}\{\hat{\alpha}_r/c_H - \alpha_0\}^2, \;\; R^{-1}\sum_{r=1}^{R}\{\hat{\beta}_r - \beta_0\}^2.$$

We consider the GARCH(1,1) model in Sect. 8.5.1 and higher-order GARCH(2,1) and GARCH(1,2) models in Sect. 8.5.2 and 8.5.4, respectively. Section 8.5.3 considers a case of misspecified GARCH orders.

8.5.1 GARCH(1,1) Models

In Tables 8.2 and 8.3, we report the adjusted biases and MSEs of the Huber and μ-type M-estimators to guide our choice of the tuning parameters k and μ. The underlying data-generating process (DGP) is the GARCH(1,1) model with $\boldsymbol{\theta}_0 = (1.65 \times 10^{-5}, 0.0701, 0.901)'$, under three types of innovation distributions: normal, double exponential, and logistic. The sample size is $n = 1000$, and we used $R = 150$ replications.

Results from Tables 8.2 and 8.3 reveal that the adjusted bias and MSE of Huber's k-estimator and the μ-estimator do not vary much with k and μ. Therefore, $k = 1.5$ and $\mu = 3$ are chosen for subsequent computations. Notice also that the minimum bias and MSE are obtained for the μ-estimator with $\mu = 3$ in most cases.

8.5.2 GARCH(2,1) Models

In this section, we consider GARCH(2,1) models with five types of innovation distributions: the normal, double exponential, logistic, and $t(d)$ with $d = 2.2, 3$, where the symbol $t(d)$ is the t-distribution with d degrees of freedom. The sample size is still $n = 1000$, and $R = 1000$ replications were generated from the GARCH(2,1)

Table 8.2 The adjusted bias and MSE of Huber estimators for various values of k under a GARCH(1,1) model with various error distributions (normal, double exponential, logistic); sample size $n = 1000$; $R = 150$ replications

	Adjusted bias			Adjusted MSE		
	ω	α	β	ω	α	β
Normal						
$k = 1$	1.03×10^{-5}	-2.44×10^{-3}	-1.96×10^{-2}	2.62×10^{-10}	4.20×10^{-4}	1.54×10^{-3}
$k = 1.5$	1.22×10^{-5}	2.47×10^{-3}	-1.98×10^{-2}	3.33×10^{-10}	4.55×10^{-4}	1.58×10^{-3}
$k = 2.5$	1.14×10^{-5}	-4.33×10^{-4}	-2.02×10^{-2}	3.10×10^{-10}	3.71×10^{-4}	1.58×10^{-3}
Double exponential						
$k = 1$	7.24×10^{-6}	1.29×10^{-3}	-1.57×10^{-2}	1.87×10^{-10}	4.65×10^{-4}	1.58×10^{-3}
$k = 1.5$	7.32×10^{-6}	1.67×10^{-3}	-1.63×10^{-2}	2.00×10^{-10}	4.82×10^{-4}	1.68×10^{-3}
$k = 2.5$	8.27×10^{-6}	2.94×10^{-3}	-1.92×10^{-2}	2.79×10^{-10}	5.60×10^{-4}	2.22×10^{-3}
Logistic						
$k = 1$	9.87×10^{-6}	2.15×10^{-3}	-2.03×10^{-2}	3.18×10^{-10}	5.25×10^{-4}	2.28×10^{-3}
$k = 1.5$	1.00×10^{-5}	2.04×10^{-3}	-2.04×10^{-2}	3.11×10^{-10}	4.89×10^{-4}	2.22×10^{-3}
$k = 2.5$	1.06×10^{-5}	2.18×10^{-3}	-2.16×10^{-2}	3.18×10^{-10}	4.84×10^{-4}	2.17×10^{-3}

Table 8.3 The adjusted bias and MSE of μ-estimators for various values of μ under a GARCH(1,1) model with various error distributions (normal, double exponential, logistic); sample size $n = 1000$; $R = 150$ replications

	Adjusted bias			Adjusted MSE		
	ω	α	β	ω	α	β
Normal						
$\mu = 2$	1.17×10^{-5}	2.97×10^{-3}	-2.13×10^{-2}	4.05×10^{-10}	6.73×10^{-4}	2.16×10^{-3}
$\mu = 2.5$	1.14×10^{-5}	1.80×10^{-3}	-2.12×10^{-2}	3.77×10^{-10}	5.71×10^{-4}	2.04×10^{-3}
$\mu = 3$	1.14×10^{-5}	1.36×10^{-3}	-2.11×10^{-2}	3.68×10^{-10}	5.21×10^{-4}	1.97×10^{-3}
Double exponential						
$\mu = 2$	7.39×10^{-6}	2.23×10^{-3}	-1.49×10^{-2}	2.74×10^{-10}	7.20×10^{-4}	2.21×10^{-3}
$\mu = 2.5$	7.36×10^{-6}	1.50×10^{-3}	-1.52×10^{-2}	2.68×10^{-10}	6.56×10^{-4}	2.16×10^{-3}
$\mu = 3$	7.40×10^{-6}	1.25×10^{-3}	-1.53×10^{-2}	2.62×10^{-10}	6.17×10^{-4}	2.09×10^{-3}
Logistic						
$\mu = 2$	7.73×10^{-6}	2.22×10^{-3}	-1.37×10^{-2}	2.45×10^{-10}	6.79×10^{-4}	1.99×10^{-3}
$\mu = 2.5$	7.66×10^{-6}	9.77×10^{-4}	-1.41×10^{-2}	2.48×10^{-10}	5.88×10^{-4}	1.97×10^{-3}
$\mu = 3$	7.72×10^{-6}	5.99×10^{-4}	-1.42×10^{-2}	2.54×10^{-10}	5.44×10^{-4}	1.94×10^{-3}

model with parameter

$$\boldsymbol{\theta}_0 = (4.46 \times 10^{-6}, 0.0525, 0.108, 0.832)',$$

and this choice is motivated by the QMLE computed from the FTSE 100 dataset analyzed in Sect. 8.7.1 using the R package fGarch.

The adjusted biases and MSEs of various M-estimators are reported in Table 8.4. It turns out that the biases and MSEs of all M-estimators are quite close to those of the QMLE under normal errors. However, the QMLE produces biases and MSEs that are sizably larger than those of the other M-estimators under heavier tail distributions. Under the $t(3)$ and $t(2.2)$ distributions with infinite fourth-order moments, the advantage of the M-estimators over the QMLE becomes more prominent. Also, under the $t(2.2)$ the distribution, the LAD and Huber estimators perform poorly compared with the μ- and Cauchy estimators since the former two yield significantly larger MSE than the latter two. This provides some evidence to support the following:

(i) under normal error distributions, all M-estimators have similar performance;
(ii) the better performance of some M-estimators under heavy-tail error distributions does not come at the cost of a loss of efficiency under normal error distribution; and
(iii) the μ- and Cauchy estimators are less sensitive to heavy-tailed errors than the LAD and Huber estimators.

8.5.3 A Misspecified GARCH Case

It is of interest to check whether the M-estimators remain consistent when the order of a GARCH model is misspecified. In particular, we consider overfitting a GARCH(p_0, q_0) with a higher-order GARCH(p, q) model when at least one of $p > p_0$ or $q > q_0$ holds. In this case, we are essentially fitting a GARCH model with some component(s) of the parameter $\boldsymbol{\theta}$ equal to zero (hence lying on the boundary of the parameter space, a case which is not covered by the consistency results available so far). However, a numerical exploration of a GARCH(1,1) misspecified as GARCH(2,1) indicates that consistency can be expected to hold under such overfitting as provided below.

Various M-estimators of a GARCH(2,1) were computed when the data are generated from GARCH(1,1) with parameter value $\boldsymbol{\theta}_0 = (1.65 \times 10^{-5}, 0.0701, 0.901)'$ and various error distributions (sample size $n = 1000$ and $R = 1000$ replications). The adjusted bias and MSE of the M-estimators are shown in Table 8.5 by wrongly fitting a GARCH(2,1) with parameter value $(1.65 \times 10^{-5}, 0.0701, 0, 0.901)'$. For all distributions considered, the M-estimators of the spurious parameter α_2 are close to zero, and the bias and the MSE are quite small, indicating good performance of the M-estimators despite the misspecification. As in Table 8.4, however, the QMLE

Table 8.4 The adjusted bias and MSE of M-estimators for GARCH(2,1) models under various error distributions (normal, double exponential, logistic, $t(3)$, $t(2.2)$); sample size $n = 1000$; $R = 1000$ replications

	Adjusted bias				Adjusted MSE			
	ω	α_1	α_2	β	ω	α_1	α_2	β
Normal								
QMLE	3.55×10^{-6}	1.88×10^{-3}	3.05×10^{-3}	-2.02×10^{-2}	2.18×10^{-11}	1.53×10^{-3}	2.08×10^{-3}	1.36×10^{-3}
LAD	3.35×10^{-6}	3.55×10^{-3}	1.80×10^{-4}	-1.76×10^{-2}	2.08×10^{-11}	1.74×10^{-3}	2.36×10^{-3}	1.32×10^{-3}
Huber's	3.53×10^{-6}	5.54×10^{-3}	4.37×10^{-3}	-1.71×10^{-2}	2.16×10^{-11}	1.84×10^{-3}	2.53×10^{-3}	1.27×10^{-3}
μ-estimator	2.84×10^{-6}	2.48×10^{-3}	1.16×10^{-3}	-1.60×10^{-2}	1.91×10^{-11}	2.18×10^{-3}	3.06×10^{-3}	1.65×10^{-3}
Cauchy	2.66×10^{-6}	1.60×10^{-3}	1.57×10^{-3}	-1.55×10^{-2}	2.03×10^{-11}	2.51×10^{-3}	3.58×10^{-3}	1.94×10^{-3}
Double exponential								
QMLE	2.51×10^{-6}	1.42×10^{-2}	-1.23×10^{-2}	-1.77×10^{-2}	1.49×10^{-11}	2.59×10^{-3}	2.59×10^{-3}	1.35×10^{-3}
LAD	1.74×10^{-6}	1.14×10^{-2}	-1.09×10^{-2}	-1.31×10^{-2}	6.60×10^{-12}	1.45×10^{-3}	1.84×10^{-3}	8.53×10^{-4}
Huber's	1.73×10^{-6}	1.21×10^{-2}	-1.21×10^{-2}	-1.28×10^{-2}	6.73×10^{-12}	1.49×10^{-3}	1.92×10^{-3}	8.93×10^{-4}
μ-estimator	1.44×10^{-6}	1.25×10^{-2}	-7.18×10^{-3}	-1.12×10^{-2}	5.64×10^{-12}	1.80×10^{-3}	2.46×10^{-3}	8.97×10^{-4}
Cauchy	1.37×10^{-6}	1.36×10^{-2}	-5.67×10^{-3}	-1.12×10^{-2}	6.61×10^{-12}	2.43×10^{-3}	3.28×10^{-3}	1.03×10^{-3}
Logistic								
QMLE	3.83×10^{-6}	1.38×10^{-2}	-1.73×10^{-2}	-1.75×10^{-2}	2.64×10^{-11}	3.78×10^{-3}	3.01×10^{-3}	1.57×10^{-3}
LAD	2.97×10^{-6}	8.27×10^{-3}	-1.43×10^{-2}	-1.20×10^{-2}	1.55×10^{-11}	2.01×10^{-3}	2.16×10^{-3}	1.11×10^{-3}
Huber's	3.03×10^{-6}	8.42×10^{-3}	-1.23×10^{-2}	-1.25×10^{-2}	1.64×10^{-11}	2.01×10^{-3}	2.03×10^{-3}	1.12×10^{-3}
μ-estimator	2.50×10^{-6}	6.28×10^{-3}	-1.25×10^{-2}	-8.64×10^{-3}	1.33×10^{-11}	2.19×10^{-3}	2.98×10^{-3}	1.23×10^{-3}
Cauchy	2.41×10^{-6}	6.46×10^{-3}	-1.10×10^{-2}	-8.62×10^{-3}	1.42×10^{-11}	2.50×10^{-3}	3.49×10^{-3}	1.46×10^{-3}
$t(3)$								
QMLE	1.67×10^{-6}	2.89×10^{-2}	-2.20×10^{-2}	-3.48×10^{-2}	2.74×10^{-11}	1.37×10^{-2}	1.56×10^{-2}	8.02×10^{-3}
LAD	1.00×10^{-6}	7.28×10^{-3}	-6.13×10^{-3}	-1.04×10^{-2}	5.62×10^{-12}	3.01×10^{-3}	4.58×10^{-3}	2.02×10^{-3}
Huber's	9.74×10^{-7}	8.20×10^{-3}	-8.00×10^{-3}	-1.05×10^{-2}	5.50×10^{-12}	2.99×10^{-3}	4.53×10^{-3}	2.01×10^{-3}
μ-estimator	6.62×10^{-7}	8.42×10^{-3}	-8.91×10^{-3}	-5.33×10^{-3}	3.93×10^{-12}	2.30×10^{-3}	3.59×10^{-3}	1.63×10^{-3}
Cauchy	5.89×10^{-7}	9.44×10^{-3}	-9.33×10^{-3}	-5.20×10^{-3}	4.33×10^{-12}	2.51×10^{-3}	3.91×10^{-3}	1.85×10^{-3}
$t(2.2)$								
QMLE	-4.35×10^{-7}	9.90×10^{-2}	-4.39×10^{-2}	-1.54×10^{-1}	1.90×10^{-11}	1.34×10^{-1}	1.48×10^{-1}	8.10×10^{-2}
LAD	1.13×10^{-6}	3.16×10^{-2}	-8.87×10^{-5}	-3.48×10^{-2}	1.35×10^{-11}	3.30×10^{-2}	4.54×10^{-2}	1.38×10^{-2}
Huber's	1.38×10^{-6}	5.30×10^{-2}	-1.08×10^{-2}	-4.40×10^{-2}	1.53×10^{-11}	4.43×10^{-2}	5.52×10^{-2}	1.58×10^{-2}
μ-estimator	4.55×10^{-7}	1.60×10^{-2}	-4.41×10^{-3}	-1.30×10^{-2}	5.51×10^{-12}	5.75×10^{-3}	9.33×10^{-3}	5.38×10^{-3}
Cauchy	4.69×10^{-7}	2.04×10^{-2}	-5.37×10^{-3}	-1.47×10^{-2}	6.74×10^{-12}	6.13×10^{-3}	1.06×10^{-2}	6.52×10^{-3}

appears to be sensitive to the heavy-tailed distributions while other M-estimators are more robust.

8.5.4 GARCH(1,2) Models

Simulations for the GARCH(1,2) (with parameter $\boldsymbol{\theta}_0 = (0.1, 0.1, 0.2, 0.6)'$, $R = 1000$, and $n = 1000$) were conducted in the same way as for GARCH(2,1) in Sect. 8.5.2. The results are shown in Table 8.6; we do not report the results for the QMLE under the $t(3)$ and $t(2.2)$ error distributions, since the algorithm did not converge for most replications, a failure of the QMLE.

Inspection of Table 8.6 reveals that under the normal error distribution, the LAD and Huber estimators produce MSEs that are close to the QMLE ones while the μ- and Cauchy estimators yield larger MSEs for the estimation of ω and α. For the double exponential and logistic distributions, there is no significant difference between the various estimators. The clear difference emerges under heavy-tailed distributions though; the μ- and Cauchy estimators produce smaller MSEs than the LAD and Huber estimators of ω and α under the $t(3)$ and $t(2.2)$ distributions, respectively.

8.6 Performance of the Bootstrap Confidence Intervals

The performance of bootstrap and classical confidence intervals (based on the QMLE) can be assessed and compared in terms of coverage probabilities. We generated $R = 500$ series of length $n = 1000$ from the GARCH(1, 1) model with parameter $\boldsymbol{\theta}_0 = (0.1, 0.1, 0.8)'$, under the normal and $t(3)$ errors. For each simulated series, we computed $B = 2000$ bootstrap estimators based on the schemes in Sect. 8.3 and constructed the bootstrap and asymptotic confidence intervals (8.16) and (8.17), respectively. The coverage probabilities are computed as the proportions of the confidence intervals covering the actual parameter value. In Table 8.7, we report these coverage probabilities (in percentage) for nominal confidence levels 90% and 95%.

Under the normal distribution, the coverage probabilities of the bootstrap approximation are generally close to the nominal levels. Also, the bootstrap approximation works better for the QMLE, LAD, and Huber estimators than for the μ- and Cauchy ones. However, under the $t(3)$ distribution, the bootstrap approximation works poorly for the QMLE while the coverage probabilities are reasonably good for all other M-estimators. For both distributions, Scheme U outperforms Schemes M and E. Except for the normal error case, thus, in terms of coverage probabilities, the classical confidence intervals based on the asymptotics of the QMLE are outperformed by the bootstrap confidence intervals based on the bootstrap Scheme U and is recommended in the analysis of the financial data.

Table 8.5 The adjusted bias and MSE of the M-estimators under a GARCH(1,1) model misspecified as GARCH(2,1) under various error distributions (normal, double exponential, logistic, $t(3)$); sample size $n = 1000$; $R = 1000$ replications

	Adjusted bias				Adjusted MSE			
	ω	α_1	$\alpha_2(=0)$	β	ω	α_1	$\alpha_2(=0)$	β
Normal								
QMLE	1.11×10^{-5}	-2.00×10^{-3}	5.97×10^{-3}	-2.38×10^{-2}	3.94×10^{-10}	1.55×10^{-3}	1.87×10^{-3}	2.64×10^{-3}
LAD	1.09×10^{-5}	-1.73×10^{-3}	5.65×10^{-3}	-2.43×10^{-2}	4.53×10^{-10}	1.73×10^{-3}	2.12×10^{-3}	3.09×10^{-3}
Huber's	1.22×10^{-5}	1.25×10^{-3}	6.08×10^{-3}	-2.43×10^{-2}	5.18×10^{-10}	1.82×10^{-3}	2.28×10^{-3}	3.13×10^{-3}
μ-estimator	1.11×10^{-5}	-5.36×10^{-4}	5.75×10^{-3}	-2.49×10^{-2}	5.27×10^{-10}	2.42×10^{-3}	2.99×10^{-3}	3.67×10^{-3}
Cauchy	1.13×10^{-5}	-5.85×10^{-4}	6.31×10^{-3}	-2.61×10^{-2}	6.26×10^{-10}	2.83×10^{-3}	3.57×10^{-3}	4.41×10^{-3}
Double exponential								
QMLE	9.70×10^{-6}	-1.07×10^{-3}	7.12×10^{-3}	-2.45×10^{-2}	4.19×10^{-10}	2.82×10^{-3}	3.33×10^{-3}	3.78×10^{-3}
LAD	8.11×10^{-6}	6.07×10^{-4}	4.72×10^{-3}	-1.89×10^{-2}	2.91×10^{-10}	2.24×10^{-3}	2.60×10^{-3}	2.51×10^{-3}
Huber's	7.84×10^{-6}	-7.00×10^{-4}	4.79×10^{-3}	-1.94×10^{-2}	2.92×10^{-10}	2.20×10^{-3}	2.54×10^{-3}	2.58×10^{-3}
μ-estimator	7.21×10^{-6}	2.45×10^{-3}	3.15×10^{-3}	-1.69×10^{-2}	2.85×10^{-10}	2.59×10^{-3}	3.02×10^{-3}	2.59×10^{-3}
Cauchy	7.49×10^{-6}	3.86×10^{-3}	3.29×10^{-3}	-1.79×10^{-2}	3.48×10^{-10}	3.10×10^{-3}	3.65×10^{-3}	3.20×10^{-3}
Logistic								
QMLE	1.24×10^{-5}	-1.95×10^{-3}	9.70×10^{-3}	-2.68×10^{-2}	5.24×10^{-10}	2.14×10^{-3}	2.61×10^{-3}	3.28×10^{-3}
LAD	1.03×10^{-5}	-2.81×10^{-3}	8.40×10^{-3}	-2.30×10^{-2}	3.88×10^{-10}	1.82×10^{-3}	2.23×10^{-3}	2.63×10^{-3}
Huber's	1.00×10^{-5}	-3.27×10^{-3}	8.11×10^{-3}	-2.28×10^{-2}	3.83×10^{-10}	1.78×10^{-3}	2.14×10^{-3}	2.62×10^{-3}
μ-estimator	9.47×10^{-6}	-2.29×10^{-3}	8.31×10^{-3}	-2.21×10^{-2}	3.88×10^{-10}	2.15×10^{-3}	2.69×10^{-3}	2.86×10^{-3}
Cauchy	9.74×10^{-6}	-8.90×10^{-4}	8.56×10^{-3}	-2.26×10^{-2}	4.34×10^{-10}	2.53×10^{-3}	3.21×10^{-3}	3.23×10^{-3}
***t*(3)**								
QMLE	1.08×10^{-5}	1.64×10^{-2}	1.14×10^{-2}	-5.47×10^{-2}	1.15×10^{-9}	1.93×10^{-2}	2.67×10^{-2}	1.97×10^{-2}
LAD	4.50×10^{-6}	1.05×10^{-3}	2.96×10^{-3}	-2.08×10^{-2}	1.85×10^{-10}	3.01×10^{-3}	3.41×10^{-3}	3.39×10^{-3}
Huber's	5.46×10^{-6}	4.83×10^{-3}	2.64×10^{-3}	-2.03×10^{-2}	2.19×10^{-10}	3.33×10^{-3}	3.80×10^{-3}	3.50×10^{-3}
μ-estimator	4.47×10^{-6}	5.91×10^{-3}	4.41×10^{-4}	-1.51×10^{-2}	1.45×10^{-10}	2.55×10^{-3}	2.84×10^{-3}	2.25×10^{-3}
Cauchy	3.65×10^{-6}	3.85×10^{-3}	4.77×10^{-5}	-1.54×10^{-2}	1.45×10^{-10}	2.51×10^{-3}	2.86×10^{-3}	2.56×10^{-3}

Table 8.6 The adjusted bias and MSE of M-estimators for GARCH(1,2) models under various error distributions (normal, double exponential, logistic, $t(3)$, $t(2.2)$); sample size $n = 1000$; $R = 1000$ replications

	Adjusted bias				Adjusted MSE			
	ω	α	β_1	β_2	ω	α	β_1	β_2
Normal								
QMLE	5.53×10^{-2}	1.10×10^{-3}	9.65×10^{-2}	-1.52×10^{-1}	2.66×10^{-2}	1.17×10^{-3}	1.45×10^{-1}	1.38×10^{-1}
LAD	5.93×10^{-2}	7.15×10^{-4}	9.01×10^{-2}	-1.50×10^{-1}	3.21×10^{-2}	1.31×10^{-3}	1.55×10^{-1}	1.45×10^{-1}
Huber	6.49×10^{-2}	4.64×10^{-3}	9.77×10^{-2}	-1.57×10^{-1}	3.72×10^{-2}	1.37×10^{-3}	1.56×10^{-1}	1.47×10^{-1}
μ-estimator	7.45×10^{-2}	8.93×10^{-4}	1.11×10^{-1}	-1.86×10^{-1}	7.41×10^{-2}	1.84×10^{-3}	2.16×10^{-1}	2.01×10^{-1}
Cauchy	7.51×10^{-2}	1.25×10^{-3}	1.29×10^{-1}	-2.06×10^{-1}	6.30×10^{-2}	2.17×10^{-3}	2.43×10^{-1}	2.31×10^{-1}
Double exponential								
QMLE	5.48×10^{-2}	2.93×10^{-3}	1.01×10^{-1}	-1.63×10^{-1}	3.15×10^{-2}	1.79×10^{-3}	1.62×10^{-1}	1.57×10^{-1}
LAD	3.73×10^{-2}	-1.93×10^{-3}	8.76×10^{-2}	-1.27×10^{-1}	1.20×10^{-2}	1.61×10^{-3}	1.46×10^{-1}	1.35×10^{-1}
Huber	3.83×10^{-2}	-1.22×10^{-3}	9.51×10^{-2}	-1.36×10^{-1}	1.21×10^{-2}	1.65×10^{-3}	1.53×10^{-1}	1.44×10^{-1}
μ-estimator	4.05×10^{-2}	1.15×10^{-3}	1.13×10^{-1}	-1.52×10^{-1}	1.72×10^{-2}	2.05×10^{-3}	1.73×10^{-1}	1.60×10^{-1}
Cauchy	4.74×10^{-2}	3.26×10^{-3}	1.18×10^{-1}	-1.66×10^{-1}	2.55×10^{-2}	2.48×10^{-3}	1.85×10^{-1}	1.72×10^{-1}
Logistic								
QMLE	5.77×10^{-2}	2.76×10^{-3}	1.06×10^{-1}	-1.61×10^{-1}	3.02×10^{-2}	1.49×10^{-3}	1.67×10^{-1}	1.59×10^{-1}
LAD	4.50×10^{-2}	-5.78×10^{-5}	7.27×10^{-2}	-1.18×10^{-1}	1.58×10^{-2}	1.37×10^{-3}	1.30×10^{-1}	1.18×10^{-1}
Huber	4.50×10^{-2}	-2.33×10^{-4}	8.85×10^{-2}	-1.34×10^{-1}	1.58×10^{-2}	1.36×10^{-3}	1.53×10^{-1}	1.39×10^{-1}
μ-estimator	4.52×10^{-2}	1.32×10^{-3}	9.39×10^{-2}	-1.40×10^{-1}	1.80×10^{-2}	1.72×10^{-3}	1.58×10^{-1}	1.44×10^{-1}
Cauchy	5.15×10^{-2}	2.91×10^{-3}	1.05×10^{-1}	-1.57×10^{-1}	2.98×10^{-2}	2.08×10^{-3}	1.85×10^{-1}	1.70×10^{-1}
$\boldsymbol{t(3)}$								
QMLE	–	–	–	–	–	–	–	–
LAD	2.93×10^{-2}	2.43×10^{-3}	1.08×10^{-1}	-1.40×10^{-1}	1.13×10^{-2}	2.49×10^{-3}	1.82×10^{-1}	1.59×10^{-1}
Huber	2.87×10^{-2}	1.50×10^{-3}	9.13×10^{-2}	-1.26×10^{-1}	1.18×10^{-2}	2.30×10^{-3}	1.60×10^{-1}	1.40×10^{-1}
μ-estimator	1.57×10^{-2}	8.75×10^{-5}	1.21×10^{-1}	-1.37×10^{-1}	5.59×10^{-3}	1.88×10^{-3}	1.63×10^{-1}	1.42×10^{-1}
Cauchy	1.50×10^{-2}	6.44×10^{-4}	1.38×10^{-1}	-1.54×10^{-1}	6.50×10^{-3}	2.15×10^{-3}	1.90×10^{-1}	1.65×10^{-1}
$\boldsymbol{t(2.2)}$								
QMLE	–	–	–	–	–	–	–	–
LAD	3.53×10^{-2}	2.57×10^{-2}	1.24×10^{-1}	-1.85×10^{-1}	1.30×10^{-2}	1.41×10^{-2}	2.41×10^{-1}	2.21×10^{-1}
Huber	4.86×10^{-2}	3.99×10^{-2}	7.81×10^{-2}	-1.66×10^{-1}	1.44×10^{-2}	1.63×10^{-2}	1.81×10^{-1}	1.79×10^{-1}
μ-estimator	1.72×10^{-2}	5.18×10^{-3}	1.51×10^{-1}	-1.78×10^{-1}	1.73×10^{-2}	4.27×10^{-3}	2.42×10^{-1}	2.12×10^{-1}
Cauchy	2.15×10^{-2}	9.68×10^{-3}	1.50×10^{-1}	-1.85×10^{-1}	2.05×10^{-2}	4.90×10^{-3}	2.34×10^{-1}	2.14×10^{-1}

Table 8.7 The coverage probabilities (in percentage) of the bootstrap schemes M, E, and U and asymptotic normal approximations for the M-estimators QMLE, LAD, Huber's, μ-, and Cauchy; the error distributions are normal and $t(3)$

			90% nominal level			95% nominal level		
			ω	α	β	ω	α	β
Normal	QMLE	Scheme M	89.0	86.2	88.2	91.0	92.2	91.4
		Scheme E	87.2	83.8	86.8	90.2	88.4	91.2
		Scheme U	90.2	87.4	87.2	94.4	92.6	93.2
		Asymptotic	82.6	91.0	85.8	87.0	95.2	89.0
Normal	LAD	Scheme M	86.0	83.4	84.2	88.2	87.2	88.4
		Scheme E	88.0	87.2	87.2	91.0	91.2	90.2
		Scheme U	88.6	88.4	88.0	93.2	91.8	91.8
		Asymptotic	94.0	98.8	87.0	96.4	99.4	90.4
Normal	Huber's	Scheme M	88.8	85.4	86.6	91.2	89.8	91.2
		Scheme E	88.2	89.0	88.0	91.4	92.4	90.0
		Scheme U	89.6	90.4	88.4	93.6	93.6	91.8
		Asymptotic	87.6	95.4	86.2	90.6	96.6	90.4
Normal	μ-estimator	Scheme M	88.0	84.6	86.8	89.6	87.8	88.6
		Scheme E	87.4	84.8	86.6	89.4	88.4	88.4
		Scheme U	88.6	88.4	87.6	91.8	91.8	90.6
		Asymptotic	71.4	69.6	86.8	77.4	78.2	90.8
Normal	Cauchy	Scheme M	85.6	84.0	84.4	87.8	85.8	87.6
		Scheme E	81.4	82.2	80.2	82.8	86.2	84.2
		Scheme U	88.4	88.2	87.0	90.4	91.4	89.4
		Asymptotic	97.8	99.8	85.0	98.2	100.0	89.6
$t(3)$	QMLE	Scheme M	71.0	75.4	74.8	75.0	79.0	78.0
		Scheme E	67.6	72.4	66.8	73.4	76.2	72.4
		Scheme U	75.6	84.6	75.0	81.6	87.2	80.0
		Asymptotic	–	–	–	–	–	–
$t(3)$	LAD	Scheme M	84.4	80.6	83.0	85.4	83.8	87.8
		Scheme E	84.6	85.0	81.4	87.6	87.0	86.6
		Scheme U	81.6	86.2	79.2	87.4	89.2	84.8
		Asymptotic	98.0	99.8	88.8	99.6	100.0	91.2
$t(3)$	Huber's	Scheme M	83.0	80.6	81.8	85.6	83.2	86.6
		Scheme E	81.8	79.2	80.8	85.8	81.6	85.8
		Scheme U	86.2	88.0	86.0	90.2	91.4	90.2
		Asymptotic	96.8	99.0	88.4	97.8	99.6	92.8
$t(3)$	μ-estimator	Scheme M	82.4	84.8	83.8	86.2	88.4	88.2
		Scheme E	84.6	84.0	84.6	87.4	88.0	88.8
		Scheme U	82.6	83.6	80.4	88.8	88.2	86.4
		Asymptotic	86.6	91.8	80.8	90.6	95.6	86.4
$t(3)$	Cauchy	Scheme M	78.2	83.4	78.4	81.8	86.2	82.0
		Scheme E	83.4	85.6	82.6	85.4	89.0	87.2
		Scheme U	85.0	85.0	84.8	90.0	88.6	89.2
		Asymptotic	100.0	100.0	85.6	100.0	100.0	90.8

8.7 Real Data Analysis

In this section, we analyze two financial series of daily log-returns, the FTSE 100 Index data from January 2007 to December 2009 ($n = 783$) and the Electric Fuel Corporation (EFCX) data from January 2000 to December 2001 ($n = 498$). Based on exploratory data analysis, a GARCH(1,1) model has been selected for the EFCX. A GARCH(2,1) model was preferred for the FTSE 100 data for two reasons. First, when fitted by the GARCH(2,1) model (via the `fGarch` package in R), the parameter α_2, with p-value 0.019, is highly significant; second, the Akaike information criterion (AIC) for the GARCH(2,1) model is smaller than that for the GARCH(1,1) model.

8.7.1 The FTSE 100 Data

Table 8.8 shows the estimates given by `fGarch` and by our M-estimators when fitting a GARCH(2,1) model to the FTSE 100 data. The QMLE (based on (8.18)) and `fGarch` provide similar results for all components of the parameter. Also, the M-estimates of β do not vary much. For ω, α_1, and α_2, the M-estimates are quite different since c_H in (8.12) depends on the score function H used for the estimation.

For a GARCH(p, q) model, using (8.10) and the formulas for $\{c_j(\boldsymbol{\theta}); j \geq 0\}$ in Berkes et al. ([1]; Sect. 3), we have $\hat{v}_t(\boldsymbol{\theta}_{0H}) = c_H \hat{v}_t(\boldsymbol{\theta}_0)$. Since an M-estimator $\hat{\boldsymbol{\theta}}_n$ is an estimator of $\boldsymbol{\theta}_{0H}$, $\hat{v}_t(\hat{\boldsymbol{\theta}}_n)$ estimates $c_H v_t(\boldsymbol{\theta}_0)$, which is a scale-transformation of the conditional variance. To examine the behavior of the market volatility after eliminating the effect of any particular M-estimator used, we define the normalized volatilities as

$$\hat{u}_t(\hat{\boldsymbol{\theta}}_n) := \hat{v}_t(\hat{\boldsymbol{\theta}}_n) / \sum_{i=1}^{n} \hat{v}_i(\hat{\boldsymbol{\theta}}_n); \;\; 1 \leq t \leq n. \tag{8.21}$$

Figure 8.1 shows the plot of $\{\hat{u}_t(\hat{\boldsymbol{\theta}}_n); 1 \leq t \leq n\}$ based on various M-estimators against the squared returns. Notice that although the M-estimators in Table 8.8 are distinct, the plot of their normalized volatilities in Fig. 8.1 almost overlap. Also, large

Table 8.8 FTSE 100 data. The M-estimates (QMLE, LAD, Huber's, μ-, and Cauchy) of the GARCH(2,1) model using the FTSE 100 data; the QMLEs are obtained by using `fGarch` and (8.18)

	`fGarch`	QMLE	LAD	Huber's	μ-estimator	Cauchy
ω	4.46×10^{-6}	4.65×10^{-6}	3.13×10^{-6}	3.55×10^{-6}	1.02×10^{-5}	2.51×10^{-6}
α_1	5.25×10^{-2}	4.51×10^{-2}	2.46×10^{-2}	3.45×10^{-2}	4.95×10^{-2}	6.83×10^{-3}
α_2	0.11	9.00×10^{-2}	5.57×10^{-2}	6.42×10^{-2}	0.17	4.18×10^{-2}
β	0.83	0.85	0.84	0.86	0.81	0.80

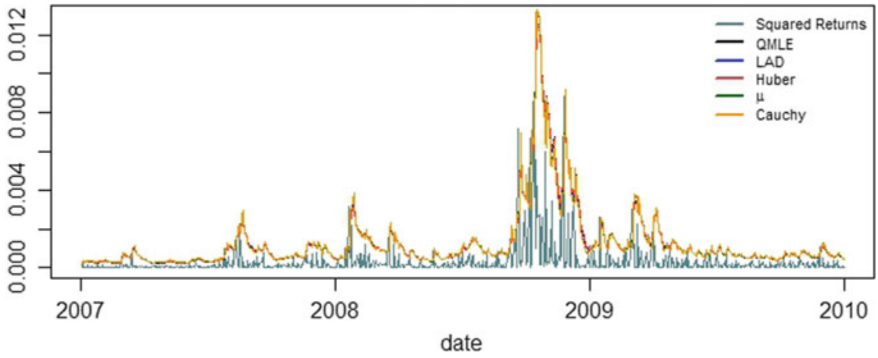


Fig. 8.1 FTSE 100 data. The plot of the squared returns and the estimated normalized conditional variances using various M-estimators for the FTSE 100 data

values of the normalized volatilities and large squared returns occur at the same time. In this sense, the volatilities are well-modeled by the resulting GARCH(2,1).

8.7.2 The Electric Fuel Corporation (EFCX) Data

Fitting a GARCH(1,1) model to the EFCX data, Muler and Yohai [9] note that the QMLE and LAD estimates of the parameter β are significantly different. In Table 8.9, we report estimates given by the `fGarch` and M-estimators. Note that in our previous analysis of the FTSE 100 data, `fGarch` estimates and the QMLE were quite close, while their differences, for this EFCX data, are much more prominent. It is also worth noting that while the LAD, Huber, $\mu-$, and Cauchy estimates of β are close to each other, they are all quite different from the corresponding value 0.84 of the QMLE. That difference might be related to the infinite fourth moment of the underlying innovation distribution and the non-robustness of the QMLE.

To determine whether the innovation distribution may have a finite fourth moment, we examine QQ-plots of the residuals $\{X_t/\hat{v}_t^{1/2}(\hat{\boldsymbol{\theta}}_n);\, 1 \le t \le n\}$ based on the μ-estimator $\hat{\boldsymbol{\theta}}_n$ against $t(d)$ distributions for various degrees of freedom d. We consider

Table 8.9 EFCX data. The M-estimates (QMLE, LAD, Huber, μ-, and Cauchy) of the GARCH(1,1) model for the EFCX data; the QMLEs are obtained by using `fGarch` and (8.18)

	`fGarch`	QMLE	LAD	Huber's	μ-estimator	Cauchy
ω	1.89×10^{-4}	6.28×10^{-4}	6.43×10^{-4}	8.37×10^{-4}	1.42×10^{-3}	2.97×10^{-4}
α	4.54×10^{-2}	7.20×10^{-2}	8.87×10^{-2}	0.10	0.27	6.35×10^{-2}
β	0.92	0.84	0.66	0.67	0.61	0.60

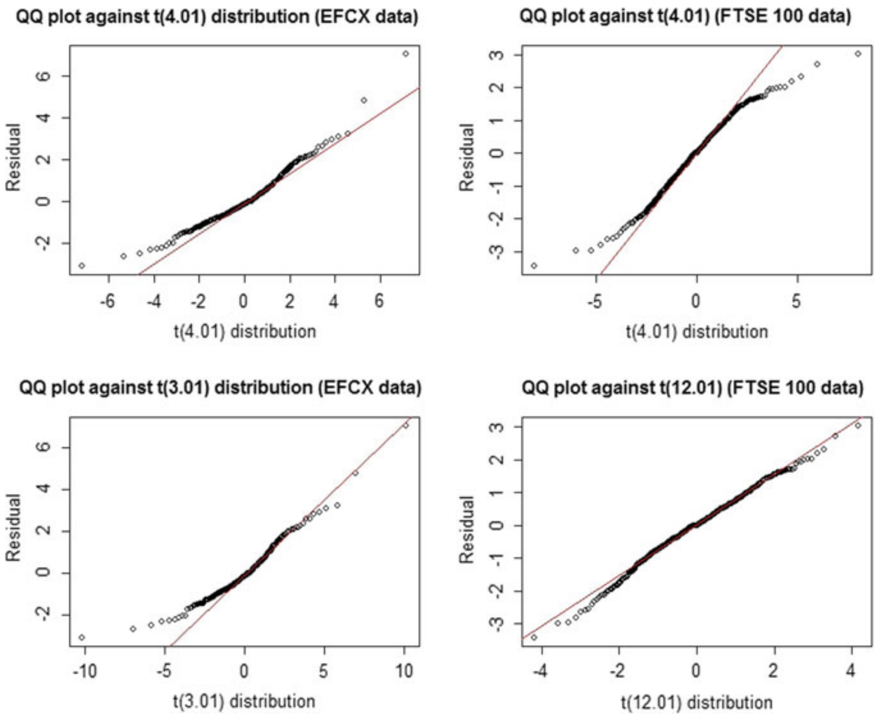


Fig. 8.2 EFCX and FTSE 100 data. The QQ-plot of the residuals against $t(d)$ distributions for the EFCX (left column, $d = 4.01$ and 3.01) and FTSE 100 (right column, $d = 4.01$ and 12.01) data

the μ-estimator, which requires the mildest moment assumptions on the innovation distribution.

The top-left panel of Fig. 8.2 shows the QQ-plot of the residuals against the $t(4.01)$ distribution for the EFCX data. The plot indicates a heavier-than-$t(4.01)$ upper tail, which implies that the fourth moment of the error term may not be finite. On the other hand, the QQ-plot against the $t(3.01)$ distribution (bottom-left panel of Fig. 8.2) yields a lighter-than-$t(3.01)$ lower tail—an indication that $\mathrm{E}|\epsilon|^3 < \infty$.

For the FTSE 100 data, the QQ-plot against the $t(4.01)$ distribution in the top-right panel of Fig. 8.2 shows that the residuals may have lighter-than-$t(4.01)$ tails. The fit in the QQ-plot against the $t(12.01)$ distribution (bottom-right panel of Fig. 8.2) looks quite good, from which we may conclude that $\mathrm{E}\epsilon^4 < \infty$ holds for the FTSE 100 data. This might explain why all M-estimators of β in Table 8.8 yield similar values.

8.8 Conclusion

We have considered a class of M-estimators and the weighted bootstrap approximation of their distributions for the GARCH models. Iteratively re-weighted algorithms for computing the M-estimators and their bootstrap replicates have been implemented. Both simulation and real data analysis demonstrate superior performance of the M-estimators for the GARCH(1,1), GARCH(2,1), and GARCH(1,2) models. Under heavy-tailed error distributions, we have shown that the M-estimators are more robust than the routinely applied QMLE. We also have demonstrated through simulations that the M-estimators work well when the true GARCH(1,1) model is misspecified as the GARCH(2,1) model. Simulation results indicate that under the finite sample size, the bootstrap approximation is better than the asymptotic normal approximation of the M-estimators.

References

1. BERKES, I., HORVATH L. AND KOKOSZKA, P. (2003). GARCH processes: structure and estimation. *Bernoulli* **9**, 201-228.
2. BERNDT, E., HALL, B., HALL, R. AND HAUSMAN, J. (1974). Estimation and inference in nonlinear structural models. *Annals of Economic and Social Measurement* **3**, 653-665.
3. CHATTERJEE, S. AND BOSE, A. (2005). Generalized bootstrap for estimating equations. *Annals of Statistics* **33**, 414-436.
4. FRANCQ, C. AND ZAKOIAN J. (2009). Testing the nullity of GARCH coefficients: correction of the standard tests and relative efficiency comparisons. *Journal of the American Statistical Association* **104**, 313-324.
5. JEONG, M. (2017). Residual-based GARCH bootstrap and second order asymptotic refinement. *Econometric Theory*, **33**, 779-790.
6. LEE, S. AND TANIGUCHI, M. (2005). Asymptotic theory for ARCH-SM models: LAN and residual empirical processes. *Statistica Sinica***15**, 215-234.
7. MUKHERJEE, K. (2008). M-estimation in GARCH models. *Econometric Theory* **24**, 1530-1553.
8. MUKHERJEE, K. (2020). Bootstrapping M-estimators in GARCH models. *Biometrika* **107**, 753-760.
9. MULER, N. AND YOHAI, V. (2008). Robust estimates for GARCH models. *Journal of Statistical Planning and Inference* **138**, 2918-2940.
10. TANIAI, H., USAMI, T., SUTO, N., AND TANIGUCHI, M. (2012). Asymptotics of realized volatility with non-Gaussian ARCH microstructure noise. *Journal of Financial Econometrics* **10**, 617-636.
11. TANIGUCHI, M., HIRUKAWA, J., AND TAMAKI, K. (2008). *Optimal Statistical Inference in Financial Engineering*. Chapman and Hall/CRC, New York.
12. TANIGUCHI, M., AMANO, T., OGATA, H., AND TANIAI, H. (2014). *Statistical Inference for Financial Engineering*. Springer-Verlag, Heidelberg.

Chapter 9
Rank Tests for Randomness Against Time-Varying MA Alternative

Junichi Hirukawa and Shunsuke Sakai

Abstract In this paper, we extend the idea of the problem of testing randomness against ARMA alternative to a class of locally stationary processes introduced by [1, 2]. We use the linear serial rank statistics and apply the notion of the contiguity [12] for the testing problem. Under the null hypothesis, the joint asymptotic normality of the proposed rank test statistics and log-likelihood ratio is established by making use of the local asymptotic normal property. Then, applying LeCam's third lemma, the asymptotic normality of test statistic under the alternative is shown automatically.

9.1 Introduction

The stationary process has been widely used for many statistical problems in time series analysis. Various properties of stationary model have been established and applied in many fields. Although these models play an important role in statistical analyses, the assumption of stationarity is severe in practice. Many empirical studies showed that actual time series data generally behave like non-stationary. Therefore, there is a natural need for a time series analysis method without the stationary assumption. In order to develop the asymptotic theory, [1, 2] proposed an important model of a non-stationary process that is referred to as locally stationary processes. The locally stationary processes have time-varying spectral density functions whose spectral structures change slowly in time.

In the asymptotic theory of statistical analyses, the locally asymptotic normality (LAN) (see, e.g., [12, 13]) is one of the most fundamental concepts and describes the optimal solution of virtually all asymptotic inference and testing problems. The LAN approach has been introduced in time series settings. The paper [14] established the

J. Hirukawa (✉)
Faculty of Science, Niigata University, 8050, Ikarashi Nino-cho, Nishi-ku, Niigata 950-2181, Japan
e-mail: hirukawa@math.sc.niigata-u.ac.jp

S. Sakai
Graduate School of Science and Technology, Niigata University, 8050, Ikarashi Nino-cho, Nishi-ku, Niigata 950-2181, Japan

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_9

LAN for AR models of finite order with a regression trend and applied it in the derivation of the local power of the Durbin–Watson test. The papers [10, 11] also proved the LAN property for ARMA and AR(∞) models, and constructed locally asymptotically minimax (LAM) adaptive estimators, as well as locally asymptotically max-min tests. For time series regression models with long memory disturbance, [7] showed LAN theorem and discussed an adaptive estimation.

ARMA models are widely used to describe time series data. Testing for ARMA model (or randomness) against other ARMA models is certainly a very important problem in time series analysis, because of its implication in the various identification and validation steps of time series model-building procedures. The paper [3] gave the fundamental contribution to the classical theory of rank-based inference. The paper [6] proposed the linear serial rank statistics for the problem of testing randomness against ARMA model. The paper [5] derived the asymptotic distribution of the log-likelihood ratio when an alternative ARMA model is contiguous to another null ARMA model. They also proposed a test based on the linear serial rank statistics and derived its asymptotic normality under both null and alternative hypotheses.

Based on the ideas of these previous studies, we consider the problem of testing randomness against locally stationary MA (time-varying MA) alternative models. We use the linear serial rank statistics and contiguity of LeCam's notion for the testing problem. In the contiguous case, the derivations of the limiting distributions of the test statistics are automatic by virtue of LeCam's third lemma. Then, our main purpose is establishing the asymptotic normality of the linear serial rank statistics under both randomness (null hypothesis) and time-varying MA models (alternative hypothesis).

The rest of this paper is organized as follows. In Sect. 9.2, we review two previous studies on which the methods and results of this paper are based. We introduce the linear serial rank statistics for the problem of testing randomness against ARMA alternative model (see [6]). We also state the LAN property for locally stationary processes (see [8]). In Sect. 9.3, we formulate the problem of testing randomness against time-varying MA alternative and show LAN property to this setting. That is, we derive the central sequence and the Fisher information matrix for hypotheses between randomness and time-varying MA alternative. Finally, in Sect. 9.4, we propose U-statistics which approximate the linear serial rank statistics and the central sequence, respectively. In addition, we show the asymptotic normality of the linear serial rank statistics under both null and alternative hypotheses.

9.2 Preceding Studies

In this section, we review two previous researches that form the basis for the results of this paper. In Sect. 9.2.1, we look back at the linear serial rank statistics for the problem of testing randomness against ARMA alternative model (see [6]). We also state the LAN property for locally stationary processes (see [8]) in Sect. 9.2.2.

9.2.1 Linear Serial Rank Tests for White Noise Against Stationary ARMA Alternatives

Nonparametric methods have been widely developed for the analysis of univariate and multivariate observations without the distributional assumption (e.g., Gaussian assumption). The nonparametric procedures are also powerful tools even in time series analysis. The paper [6] considered the study of testing for randomness against ARMA alternatives. They proposed the statistics of the form

$$S_T = \frac{1}{T-q}\sum_{t=1}^{T} c_T\left(R_t^{(T)}, R_{t-1}^{(T)}, \ldots, R_{t-q}^{(T)}\right),$$

where $R_t^{(T)}$ denotes the rank of the observation $X_t^{(T)}$ among the observed stretch $\boldsymbol{X}^{(T)} = \left(X_1^{(T)}, \ldots, X_T^{(T)}\right)^\top$ of length T. These statistics are the so-called *linear serial rank statistics* (of order q). The special cases of the score functions c_T include traditional test statistics, which are listed in [6], i.e.,

(i) the run statistic (with respect to median)

$$c_T(i_1, i_2) = \begin{cases} 1 & \text{if } (2i_1 - T - 1)(2i_2 - T - 1) < 0 \\ 0 & \text{if } (2i_1 - T - 1)(2i_2 - T - 1) \geq 0, \end{cases}$$

(ii) the turning point statistic

$$c_T(i_1, i_2, i_3) = \begin{cases} 1 & \text{if } i_1 > i_2 < i_3 \\ 1 & \text{if } i_1 < i_2 > i_3 \\ 0 & \text{elsewhere,} \end{cases}$$

(iii) Spearman's rank correlation coefficient of order q (up to additive and multiplicative constants)

$$c_T\left(i_1, i_2, \ldots, i_{q+1}\right) = \frac{i_1 i_{q+1}}{(T+1)^2}.$$

In what follows, we assume that there exists a function $J = J\left(v_{q+1}, \ldots, v_1\right)$, defined over $[0, 1]^{q+1}$, such that

$$0 < \int_{[0,1]^{q+1}} J^2\left(v_{q+1}, \ldots, v_1\right) dv_{q+1} \cdots dv_1 < \infty \tag{9.1}$$

and

$$\lim_{T\to\infty} E\left[\left\{J\left(U_{q+1}^{(T)},\ldots,U_1^{(T)}\right)-c_T\left(R_{q+1}^{(T)},\ldots,R_1^{(T)}\right)\right\}^2 \mid H_0^{(T)}\right]=0,$$

where $c_T(\ldots)$ is the score function. This assumption is satisfied when c_T is of the form

$$c_T\left(i_1,i_2,\ldots,i_{q+1}\right)=J\left(i_1/(T+1), i_2/(T+1),\ldots,i_{q+1}/(T+1)\right).$$

Such a function J is called a score-generating function (associated with the linear serial rank statistics S_T).

Let $a_1,\ldots,a_{q_1},b_1,\ldots,b_{q_2}$ be an arbitrary (q_1+q_2)-tuple of the real numbers, and consider stochastic difference equations

$$X_t^{(T)}-\frac{1}{\sqrt{T}}\sum_{i=1}^{q_1}a_iX_{t-i}^{(T)}=\varepsilon_t+\frac{1}{\sqrt{T}}\sum_{i=1}^{q_2}b_i\varepsilon_{t-i},\quad t\in\mathbb{Z},\ T\geq 1,$$

where $\{\varepsilon_t \mid t\in\mathbb{Z}\}$ is a sequence of i.i.d. random variables with mean zero, variance σ^2 and density function p. We assume that $\{\varepsilon_t \mid t\in\mathbb{Z}\}$ satisfies the following regularity conditions.

Assumption 9.1 ([6, Sect. 2])

(i) ε_t has finite moments up to the third order.
(ii) p is a.e. differentiable and its derivative p' satisfies $\int_{-\infty}^{\infty}\left|p'(x)\right|dx<\infty$.
(iii) p is absolutely continuous on finite intervals and satisfies $E\left\{\left|\frac{p'(\varepsilon_t)}{p(\varepsilon_t)}\right|^{2+\delta}\right\}<\infty$ for some $\delta>0$. This implies the finiteness of Fisher information $\mathcal{F}(p)$, i.e.,

$$0<\mathcal{F}(p):=\int\phi^2(x)\,p(x)\,dx=\int_{-\infty}^{\infty}\left\{\frac{p'(x)}{p(x)}\right\}^2p(x)\,dx<\infty,$$

where $\phi(x)=-p'(x)/p(x)$, $x\in\mathbb{R}$.
(iv) ϕ is a.e. differentiable and satisfies a Lipschitz condition $|\phi(x)-\phi(y)|<K|x-y|$ a.e., where K is a constant. Further, let P be distribution function of p, and $P^{-1}(u)=\inf\{x \mid P(x)\geq u\}$, $0<u<1$. Put

$$\phi\left(P^{-1}(u)\right)=-\frac{p'\left\{P^{-1}(u)\right\}}{p\left\{P^{-1}(u)\right\}},\quad 0<u<1.$$

Denote by $H_0^{(T)}$ the sequence of simple (null) hypotheses which satisfies $a_1=\cdots=a_{q_1}=b_1=\cdots=b_{q_2}=0$. Then, the process $\left\{X_t^{(T)}\right\}$ all coincide with the white noise process $\{\varepsilon_t\}$ which has a likelihood function

$$l_T^0\left(\boldsymbol{x}^{(T)}\right)=\prod_{t=1}^{T}p\left(x_t^{(T)}\right),$$

where $\boldsymbol{x}^{(T)}=\left(x_1^{(T)},\ldots,x_T^{(T)}\right)^\top$ is a value of an observed stretch $\boldsymbol{X}^{(T)}$. On the other hand, denote by $H_1^{(T)}$ the corresponding sequence of alternative hypotheses which satisfies that both a_{q_1} and b_{q_2} are different from zero. Then, the processes $\left\{X_t^{(T)}\right\}$ are ARMA(q_1, q_2) processes whose likelihood function is denoted by $l_T^1\left(\boldsymbol{x}^{(T)}\right)$. The exact expression of this likelihood function can be found in [6, Eq. (3.3)].

In what follows, we shall omit the superscript (T), i.e., write $\boldsymbol{x}$, x_t, $\boldsymbol{X}$ and X_t, in place of $\boldsymbol{x}^{(T)}$, $x_t^{(T)}$, $\boldsymbol{X}^{(T)}$ and $X_t^{(T)}$, respectively. Now, we consider the likelihood ratio

$$L_T(\boldsymbol{x})=\begin{cases} l_T^1(\boldsymbol{x})/l_T^0(\boldsymbol{x}) & \text{if } l_T^0(\boldsymbol{x})>0\\ 1 & l_T^1(\boldsymbol{x})=l_T^0(\boldsymbol{x})=0\\ \infty & l_T^1(\boldsymbol{x})>l_T^0(\boldsymbol{x})=0\end{cases}$$

between hypotheses $H_0^{(T)}$ and $H_1^{(T)}$. Then we have the following lemma.

Lemma 9.1 ([6, Proposition 3.1]) *Suppose that Assumption 9.1 holds. Under $H_0^{(T)}$, the log-likelihood ratio has the stochastic expansion*

$$\log L_T(\boldsymbol{X})=\mathcal{L}_T^0(\boldsymbol{X})-\frac{d^2}{2}+o_P(1),$$

where

$$\mathcal{L}_T^0(\boldsymbol{X})=\frac{1}{\sqrt{T}}\sum_{t=q+1}^{T}\phi(X_t)\sum_{i=1}^{q}d_iX_{t-i},$$

$$d_i=\begin{cases} a_i+b_i & 1\le i\le \min(q_1,q_2)\\ a_i & q_2<i\le q_1 \text{ if } q_2<q_1\\ b_i & q_1<i\le q_2 \text{ if } q_1<q_2,\end{cases}$$

$$q=\max(q_1,q_2),\quad d^2=\sum_{i=1}^{q}d_i^2\sigma^2\mathcal{F}(p).$$

Furthermore, $\mathcal{L}_T^0(\boldsymbol{X})$ is asymptotically normal, with mean zero and variance d^2.

From LeCam's first lemma (see, e.g., [4, Sect. 7.1.2]), we can see that this lemma implies that $H_1^{(T)}$ is contiguous to $H_0^{(T)}$.

Let

$$\begin{aligned} J^*\left(u_{q+1},\ldots,u_1\right) = J\left(u_{q+1},\ldots,u_1\right) \\ &- \sum_{k=1}^{q+1} \int_{[0,1]^q} J\left(\nu_q,\ldots,\nu_k,u_1,\nu_{k-1},\ldots,\nu_1\right) d\nu_1\cdots d\nu_q \\ &+ q\int_{[0,1]^{q+1}} J\left(\nu_{q+1},\ldots,\nu_1\right) d\nu_1\cdots d\nu_{q+1}, \end{aligned} \tag{9.2}$$

and

$$m_T = E\left(S_T \mid H_0^{(T)}\right) = \frac{1}{T(T-1)\cdots(T-q)} \sum_{1\le i_1\ne\cdots\ne i_{q+1}\le T} c_T\left(i_1,\ldots,i_{q+1}\right).$$

We also have the following lemma.

Lemma 9.2 ([6, Proposition 4.1]) *Suppose that Assumption 9.1 holds. Under $H_0^{(T)}$,*

$$\begin{pmatrix} \sqrt{T}\left(S_T - m_T\right) \\ \log L_T \end{pmatrix} \xrightarrow{\mathcal{L}} N\left(\boldsymbol{m},\boldsymbol{\Sigma}\right),$$

with

$$\boldsymbol{m} = \begin{pmatrix} 0 \\ -\frac{1}{2}\sum_{j=1}^q d_j^2\sigma^2\mathcal{F}(p) \end{pmatrix}$$

and

$$\boldsymbol{\Sigma} = \begin{pmatrix} V^2 & \sum_{j=1}^q d_jC_j \\ \sum_{j=1}^q d_jC_j & \sum_{j=1}^q d_j^2\sigma^2\mathcal{F}(p) \end{pmatrix},$$

where $\xrightarrow{\mathcal{L}}$ stands for the convergence in distribution,

$$\begin{aligned} V^2 := &\int_{[0,1]^{q+1}} \left\{J^*\left(\nu_{q+1},\ldots,\nu_1\right)\right\}^2 d\nu_1\cdots d\nu_{q+1} \\ &+ 2\sum_{l=1}^q \int_{[0,1]^{q+l+1}} J^*\left(\nu_{q+1},\ldots,\nu_1\right) J^*\left(\nu_{q+l+1},\ldots,\nu_{l+1}\right) d\nu_1\cdots d\nu_{q+l+1} \end{aligned}$$

and

$$C_j = \int_{[0,1]^{q+1}} J^* \left(\nu_{q+1}, \ldots, \nu_1\right) \sum_{l=0}^{q-j} \phi \left(P^{-1} \left(\nu_{q-l+1}\right)\right) P^{-1} \left(\nu_{q-l-j+1}\right) d\nu_1 \cdots d\nu_{q+1}.$$

The asymptotic normality of the rank test statistics under $H_1^{(T)}$ then follows from LeCam's third lemma (see, e.g., [4, Sect. 7.1.4]).

Corollary 9.1 ([6, Proposition 4.2]) *Suppose that Assumption 9.1 holds. Under* $H_1^{(T)}$,

$$\sqrt{T} \left(S_T - m_T\right) \xrightarrow{\mathcal{L}} N \left(\sum_{j=1}^{q} d_j C_j, V^2\right).$$

9.2.2 Locally Asymptotic Normality of Time-Varying AR Models

In dealing with non-stationary processes, one of the difficult problems to be solved is how to set up an adequate asymptotic theory. To meet this, [1, 2] introduced an important class of non-stationary processes and developed the statistical inference. We give the precise definition which is due to [1, 2].

Definition 9.1 A sequence of stochastic processes $X_{t,T}$ $(t = 1, \ldots, T;\ T \geq 1)$ is called locally stationary with transfer function A°, if there exists a representation

$$X_{t,T} = \int_{-\pi}^{\pi} \exp\left(i\lambda t\right) A_{t,T}^\circ \left(\lambda\right) d\zeta\left(\lambda\right),$$

such that

(i) $\zeta\left(\lambda\right)$ is a stochastic process on $[-\pi, \pi]$ with $\overline{\zeta\left(\lambda\right)} = \zeta\left(-\lambda\right)$ and

$$\text{cum}\left\{d\zeta\left(\lambda_1\right), \ldots, d\zeta\left(\lambda_k\right)\right\} = \eta\left(\sum_{j=1}^{k} \lambda_j\right) h_k\left(\lambda_1, \ldots, \lambda_{k-1}\right) d\lambda_1 \cdots d\lambda_{k-1},$$

where $\text{cum}\{\ldots\}$ denotes the cumulant of kth order, $h_1 = 0$, $h_2\left(\lambda\right) = \frac{1}{2\pi}$, $h_k\left(\lambda_1, \ldots, \lambda_{k-1}\right) = \frac{h_k}{(2\pi)^{k-1}}$ for all $k \geq 3$ and $\eta\left(\lambda\right) = \sum_{j=-\infty}^{\infty} \delta\left(\lambda + 2\pi j\right)$ is the 2π-periodic extension of the Dirac delta function;

(ii) there exists a constant K and a 2π-periodic function $A : [0, 1] \times \mathbb{R} \to \mathbb{C}$ with $A\left(u, -\lambda\right) = \overline{A\left(u, \lambda\right)}$,

$$\sup_{t,\lambda} \left| A_{t,T}^\circ\left(\lambda\right) - A\left(\frac{t}{T}, \lambda\right)\right| \leq KT^{-1}$$

for all T, and $A(u, \lambda)$ is continuous in u.

Let $X_{1,T}, \dots, X_{T,T}$ be realizations of a locally stationary process with transfer function A_θ°, where the corresponding A_θ is uniformly bounded from above and below and time-varying spectral density $f_\theta(u, \lambda) := |A_\theta(u, \lambda)|^2$ depends on a parameter vector $\theta = (\theta_0, \dots, \theta_q)^\top \in \Theta \subset \mathbb{R}^{q+1}$. Introducing the notations $\nabla_i = \frac{\partial}{\partial\theta_i}$, $\nabla = (\nabla_0, \dots, \nabla_q)^\top$, $\nabla_{i,j} = \frac{\partial}{\partial\theta_i}\frac{\partial}{\partial\theta_j}$, $\nabla^2 = (\nabla_{i,j})_{i,j=0,\dots,q}$, we make the following assumption.

Assumption 9.2 ([8, Assumption 1]) There exists a constant K with

$$\sup_{t,\lambda}\left|\nabla^s\left\{A_{\theta,t,T}^\circ(\lambda) - A_\theta\left(\frac{t}{T}, \lambda\right)\right\}\right| \le KT^{-1}$$

for $s = 0, 1, 2$. The components of $A_\theta(u, \lambda)$, $\nabla A_\theta(u, \lambda)$ and $\nabla^2 A_\theta(u, \lambda)$ are differentiable in u and λ with uniformly continuous derivatives $\frac{\partial}{\partial u}\frac{\partial}{\partial\lambda}$.

Writing $\varepsilon_t = \int_{-\pi}^{\pi}\exp(i\lambda t)\, d\zeta(\lambda)$, we assume.

Assumption 9.3 ([8, Assumption 2])

(i) ε_t's are i.i.d. random variables with mean zero, variance 1 and finite fourth order moment $E\left(\varepsilon_t^4\right)$. Furthermore, the distribution of ε_t is absolutely continuous with respect to the Lebesgue measure and has the probability density $p(z) > 0$ on $\mathbb{R}$.

(ii) p satisfies

$$\lim_{|z|\to\infty} p(z) = 0, \quad \text{and} \quad \lim_{|z|\to\infty} zp(z) = 0.$$

(iii) The continuous derivatives p' and p'' of p exist on $\mathbb{R}$, and p'' satisfies the Lipschitz condition.

(iv) Writing $\varphi(z) = \frac{p'(z)}{p(z)}$,

$$\mathcal{F}(p) = \int \varphi^2(z)\, p(z)\, dz < \infty,$$
$$E\left\{\varepsilon_t\varphi^2(\varepsilon_t)\right\} < \infty, \quad E\left\{\varepsilon_t^2\varphi^2(\varepsilon_t)\right\} < \infty, \quad E\left\{\varphi^4(\varepsilon_t)\right\} < \infty$$

and

$$\int p''(z)\, dz = 0, \quad \lim_{||\to\infty} p'(z)\, z^2 = 0.$$

Note that $\phi(x) = -\varphi(x)$ in Assumption 9.1.

This assumption partially overlaps with Assumption 9.1, but we leave them here to describe their respective preceding studies.

The paper [1] developed the asymptotic theory for Gaussian locally stationary processes, including LAN property. On the other hand, [8] derived the LAN result for non-Gaussian locally stationary processes. The paper [8] assumed the following type representations for locally stationary processes.

Assumption 9.4 ([8, Assumption 3])

(i) The sequence of the processes $\{X_{t,T}\}$ has the MA(∞) and AR (∞) representations

$$X_{t,T} = \sum_{j=0}^{\infty} \beta^{\circ}_{\theta,t,T}(j)\varepsilon_{t-j} \text{ and } \beta^{\circ}_{\theta,t,T}(0)\varepsilon_t = \sum_{k=0}^{\infty} \alpha^{\circ}_{\theta,t,T}(k)X_{t-k,T}, \tag{9.3}$$

where $\beta^{\circ}_{\theta,t,T}(j), \alpha^{\circ}_{\theta,t,T}(k) \in \mathbb{R}$, $\alpha^{\circ}_{\theta,t,T}(0) \equiv 1$ and $\beta^{\circ}_{\theta,t,T}(j) = \beta^{\circ}_{\theta,0,T}(j) = \beta^{\circ}_{\theta}(j)$ for $t \leq 0$.

(ii) Every $\beta^{\circ}_{\theta,t,T}(j)$ is continuously three times differentiable with respect to θ and the derivatives satisfy

$$\sup_{t,T} \left\{ \sum_{j=0}^{\infty} (1+j) \left| \nabla_{i_1} \cdots \nabla_{i_s} \beta^{\circ}_{\theta,t,T}(j) \right| \right\} < \infty \quad \text{for } s = 0, 1, 2, 3.$$

(iii) Every $\alpha^{\circ}_{\theta,t,T}(k)$ is continuously three times differentiable with respect to θ and the derivatives satisfy

$$\sup_{t,T} \left\{ \sum_{k=0}^{\infty} (1+k) \left| \nabla_{i_1} \cdots \nabla_{i_s} \alpha^{\circ}_{\theta,t,T}(k) \right| \right\} < \infty \quad \text{for } s = 0, 1, 2, 3.$$

(iv) Writing $f^{\circ}_{\theta,t,T}(\lambda) = \left| A^{\circ}_{\theta,t,T}(\lambda) \right|^2$,

$$\beta^{\circ}_{\theta,t,T}(0) = \exp\left\{ \frac{1}{4\pi} \int_{-\pi}^{\pi} f^{\circ}_{\theta,t,T}(\lambda)\, d\lambda \right\}.$$

Let $\mathcal{P}_{\theta,T}$ and $\mathcal{P}_{\varepsilon}$ be the probability distributions of $(\varepsilon_s,\ s \leq 0, X_{1,T}, \ldots, X_{T,T})$ and $(\varepsilon_s,\ s \leq 0)$, respectively. Denote by $H(p;\theta)$ the hypothesis under which the underlying parameter is $\theta \in \Theta$ and the probability density of ε_t is p. We define

$$\theta_T = \theta + \frac{1}{\sqrt{T}} h, \quad h = (h_1, \ldots, h_q)^{\top} \in \mathcal{H} \subset \mathbb{R}^q.$$

For two hypothetical values θ and θ_T, we define the log-likelihood ratio as

$$\Lambda_T(\theta, \theta_T) \equiv \log \frac{d\mathcal{P}_{\theta_T,T}}{d\mathcal{P}_{\theta,T}}.$$

Then, we have the following LAN result.

Lemma 9.3 ([8, Theorem 1]) *Suppose that Assumptions 9.2–9.4 hold. Then the sequence of experiments*

$$\mathcal{E}_T = \left\{\mathbb{R}^{\mathbb{Z}}, \mathcal{B}^{\mathbb{Z}}, \left\{\mathcal{P}_{\theta,T} \mid \theta \in \Theta \subset \mathbb{R}^q\right\}\right\}, \quad T \in \mathbb{N},$$

where $\mathcal{B}^{\mathbb{Z}}$ denotes the Borel σ-field on $\mathbb{R}^{\mathbb{Z}}$, is LAN and equicontinuous on compact subset C of $\mathcal{H}$. That is,

(i) For all $\theta \in \Theta$, the log-likelihood ratio $\Lambda_T(\theta, \theta_T)$ admits, under $H(p;\theta)$, as $T \to \infty$, the asymptotic representation

$$\Lambda_T(\theta, \theta_T) = h^\top \Delta_T(\theta) - \frac{1}{2} h^\top \Gamma(\theta) h + o_P(1),$$

where

$$\Delta_T(\theta) = \sum_{t=1}^{T} \left[\frac{\varphi(\varepsilon_t)}{\sqrt{T}\beta^\circ_{\theta,t,T}(0)} \sum_{k=1}^{t-1} \nabla \alpha^\circ_{\theta,t,T}(k) X_{t-k,T} - \frac{\nabla \beta^\circ_{\theta,t,T}(0)}{\sqrt{T}\beta^\circ_{\theta,t,T}(0)} \left\{1 + \varphi(\varepsilon_t)\varepsilon_t\right\} \right],$$

$$\begin{aligned}\Gamma(\theta) = \int_0^1 \Bigg[& \frac{\mathcal{F}(p)}{4\pi} \int_{-\pi}^{\pi} \frac{\{\nabla f_\theta(u,\lambda)\}\{\nabla f_\theta(u,\lambda)\}^\top}{|f_\theta(u,\lambda)|^2} d\lambda \\ & + \frac{1}{16\pi^2} \left[E\left\{\varepsilon_t^2 \varphi^2(\varepsilon_t)\right\} - 2\mathcal{F}(p) - 1 \right] \\ & \left\{ \int_{-\pi}^{\pi} \frac{\nabla f_\theta(u,\lambda)}{f_\theta(u,\lambda)} d\lambda \right\} \left\{ \int_{-\pi}^{\pi} \frac{\nabla f_\theta(u,\lambda)}{f_\theta(u,\lambda)} d\lambda \right\}^\top \Bigg] du\end{aligned}$$

and $f_\theta(u,\lambda)$ is the time-varying spectral density of the process (9.3).

(ii) Under $H(p;\theta)$,

$$\Delta_T(\theta) \xrightarrow{\mathcal{L}} N(0, \Gamma(\theta)).$$

(iii) For all $T \in \mathbb{N}$ and all $h \in \mathcal{H}$, the mapping $h \to \mathcal{P}_{\theta_T,T}$ is continuous with respect to the variational distance

$$\|\mathcal{P} - \mathcal{L}\| = \sup\left\{ |\mathcal{P}(A) - \mathcal{L}(A)| \ : \ A \in \mathcal{B}^{\mathbb{Z}} \right\}.$$

Hereafter, $\Delta_T(\theta)$ is referred to as the *central sequence*, and $\Gamma(\theta)$ the *Fisher information matrix* of the time series.

9.3 Local Asymptotic Normality for Time-Varying MA Processes

In this section, we describe the LAN property for time-varying MA(q) processes as a special case of Lemma 9.3.

9.3.1 Contiguous Hypotheses for Time-Varying MA Models

We formulate the contiguous hypotheses for time-varying MA(q) models. Let us consider the sequence of time-varying MA(q) models

$$X_{t,T}^{(T)} = \sum_{k=0}^{q} b_{\theta^{(k)}} \left(\frac{t}{T} \right) \varepsilon_{t-k}, \ t \in \mathbb{Z}, \tag{9.4}$$

where $\left\{ \varepsilon_t \mid t \in \mathbb{Z} \right\}$ is a sequence of i.i.d. random variables with mean zero, variance 1 and density function p. Let

$$b_{\theta^{(0)}} (u) = \sigma + \left(\theta^{(0)} - \sigma \right) \widetilde{b}_0 (u), \quad b_{\theta^{(k)}} (u) = \theta^{(k)} \widetilde{b}_k (u) \ \ (k = 1, \ldots, q),$$

where the functions $\widetilde{b}_k (u)$ $(k = 0, \ldots, q)$ are sufficiently smooth bounded functions in u. Therefore, the coefficient functions $b_{\theta^{(k)}} (u)$ $(k = 0, \ldots, q)$ are also sufficiently smooth in u. We suppose that Assumptions 9.2–9.4 are satisfied for $b_{\theta^{(k)}} (u)$ $(k = 0, \ldots, q)$.

Now, we consider the null and the alternative hypotheses as follows. Given

$$\theta = \left(\theta_0, \ldots, \theta_q \right)^\top \in \Theta \subset \mathbb{R}^{q+1}, \ h = \left(h_0, \ldots, h_q \right)^\top \in \mathcal{H} \subset \mathbb{R}^{q+1},$$

the null hypothesis $H(p; \theta)$ is

$$H(p; \theta) \ : \ \theta^{(0)} = \theta_0 = \sigma, \quad \theta^{(k)} = \theta_k = 0 \ (k = 1, \ldots, q)$$

and the alternative hypothesis $K(p; \theta_T)$ is

$$K(p; \theta_T) \ : \theta^{(0)} = \theta_T^{(0)} = \sigma + \frac{h_0}{\sqrt{T}}, \quad \theta^{(k)} = \theta_T^{(k)} = \frac{h_k}{\sqrt{T}} \ (k = 1, \ldots, q).$$

Then, we can write

$$\theta_T = \left(\theta_T^{(0)}, \ldots, \theta_T^{(q)} \right)^\top = \theta + \frac{1}{\sqrt{T}} h \ \in \ \mathbb{R}^{q+1}.$$

Under the two hypotheses, the coefficient functions $b_{\theta^{(k)}}(u)$ $(k = 0, \ldots, q)$ are given by

$$b_{\theta^{(0)}}(u) = \begin{cases} \sigma & \text{under } H(p;\theta) \\ \sigma + \frac{h_0}{\sqrt{T}}\widetilde{b}_0(u) & \text{under } K(p;\theta_T) \end{cases}$$

and

$$b_{\theta^{(k)}}(u) = \begin{cases} 0 & \text{under } H(p;\theta) \\ \frac{h_k}{\sqrt{T}}\widetilde{b}_k(u) & \text{under } K(p;\theta_T) \end{cases} \quad (k = 1, \ldots, q).$$

Furthermore, we note that the time-varying MA(q) process defined in (9.4) has time-varying spectral density

$$f_{\theta^{(\cdot)}}(u,\lambda) = \left|\sum_{k=0}^{q} b_{\theta^{(k)}}(u)\, e^{ik\lambda}\right|^2 = b_{\theta^{(0)}}(u)^2 \left|B_{\theta^{(\cdot)}}\left(u, e^{i\lambda}\right)\right|^2,$$

where

$$B_{\theta^{(\cdot)}}\left(u, e^{i\lambda}\right) := \sum_{k=0}^{q} \frac{b_{\theta^{(k)}}(u)}{b_{\theta^{(0)}}(u)} e^{ik\lambda}$$

and $\theta^{(\cdot)} = \left(\theta^{(0)}, \ldots, \theta^{(q)}\right)^{\top}$.

9.3.2 Fisher Information Matrix for Time-Varying MA Model

We derive the Fisher information matrix $\Gamma(\theta)$ of Lemma 9.3 for time-varying MA model in (9.4). We denote the partial derivative in $\theta^{(j)}$ by $\nabla_j = \frac{\partial}{\partial\theta^{(j)}}$ $(j = 0, \ldots, q)$. Noting that

$$\nabla_j b_{\theta^{(k)}}(u) = \begin{cases} \widetilde{b}_j(u) & (j = k) \\ 0 & (j \neq k), \end{cases}$$

we see that

$$\frac{\nabla_j f_{\theta^{(\cdot)}}(u,\lambda)}{f_{\theta^{(\cdot)}}(u,\lambda)} = \frac{\widetilde{b}_j(u)}{b_{\theta^{(0)}}(u)^2} \sum_{k=0}^{q} \sum_{l,m=0}^{\infty} b_{\theta^{(k)}}(u)\, g_{\theta^{(\cdot)},l}(u)\, g_{\theta^{(\cdot)},m}(u)$$
$$\left\{e^{i(k-j+l-m)\lambda} + e^{i(j-k+l-m)\lambda}\right\} \quad (j = 0, \ldots, q),$$

where

$$G_{\theta(\cdot)}\left(u, e^{i\lambda}\right) := \sum_{l=0}^{\infty} g_{\theta(\cdot),l}(u)\, e^{il\lambda} \tag{9.5}$$

(let $g_{\theta(\cdot),0} := 1$) satisfies

$$G_{\theta(\cdot)}\left(u, e^{i\lambda}\right) B_{\theta(\cdot)}\left(u, e^{i\lambda}\right) = \left(\sum_{l=0}^{\infty} g_{\theta(\cdot),l}(u)\, e^{il\lambda}\right)\left(\sum_{k=0}^{q} \frac{b_{\theta^{(k)}}(u)}{b_{\theta^{(0)}}(u)} e^{ik\lambda}\right) = 1. \tag{9.6}$$

Conventionally defining $g_{\theta(\cdot),l}(u) := 0$ for $l < 0$ in (9.5), we can derive the difference equations

$$\sum_{k=0}^{q} \frac{b_{\theta^{(k)}}(u)}{b_{\theta^{(0)}}(u)} g_{\theta(\cdot),t-k}(u) = \delta_t, \quad t \in \mathbb{Z} \tag{9.7}$$

from the relationship in (9.6), where δ_t is the dirac delta function. Combing the fact that $\int_{-\pi}^{\pi} e^{i\lambda} d\lambda = 2\pi\delta_n$ with (9.7), we obtain

$$\int_{-\pi}^{\pi} \frac{\nabla_j f_{\theta(\cdot)}(u,\lambda)}{f_{\theta(\cdot)}(u,\lambda)} d\lambda = \begin{cases} \frac{4\pi \widetilde{b}_0(u)}{b_{\theta(0)}(u)} & (j = 0) \\ 0 & (j = 1, \ldots, q) \end{cases}$$

and therefore,

$$\left(\int_{-\pi}^{\pi} \frac{\nabla_j f_{\theta(\cdot)}(u,\lambda)}{f_{\theta(\cdot)}(u,\lambda)} d\lambda\right)\left(\int_{-\pi}^{\pi} \frac{\nabla_k f_{\theta(\cdot)}(u,\lambda)}{f_{\theta(\cdot)}(u,\lambda)} d\lambda\right) = \begin{cases} \frac{16\pi^2 \widetilde{b}_0(u)^2}{b_{\theta(0)}(u)^2} & ((j,k) = (0,0)) \\ 0 & ((j,k) \neq (0,0)). \end{cases}$$

Similarly, we can evaluate

$$\int_{-\pi}^{\pi} \frac{\left\{\nabla_j f_{\theta(\cdot)}(u,\lambda)\right\}\left\{\nabla_k f_{\theta(\cdot)}(u,\lambda)\right\}}{\left|f_{\theta(\cdot)}(u,\lambda)\right|^2} d\lambda$$
$$= \frac{4\pi \widetilde{b}_j(u)\, \widetilde{b}_k(u)}{b_{\theta^{(0)}}(u)^2}\left(\delta_k\delta_j + \sum_{l=0}^{\infty} g_{\theta(\cdot),l}(u)\, g_{\theta(\cdot),l-j+k}(u)\right).$$

Recalling $b_{\theta^{(0)}}(u) = \sigma$ and $b_{\theta^{(k)}}(u) = 0\,(k = 1, \ldots, q)$, i.e., $g_{\theta(\cdot),l}(u) = \delta_l$ under the null hypothesis $H(p; \theta)$, we conclude that

$$-\frac{1}{2} h^\top \Gamma(\theta)\, h = -\frac{\mathcal{F}(p)}{2\sigma^2} \sum_{j=0}^{q} \int_0^1 \widetilde{d}_j(u)^2\, du, \tag{9.8}$$

where

$$\widetilde{d}_j(u) = \begin{cases} \sqrt{\frac{E\{\varepsilon_t^2\varphi^2(\varepsilon_t)\}-1}{\mathcal{F}(p)}}h_0\widetilde{b}_0(u) & (j=0) \\ h_j\widetilde{b}_j(u) & (j=1,\ldots,q). \end{cases} \tag{9.9}$$

9.3.3 Central Sequence for Time-Varying MA Model

We derive the central sequence $\Delta_T(\theta)$ in Lemma 9.3 for time-varying MA model in (9.4). Rewrite MA to AR representation, i.e.,

$$b_{\theta^{(0)}}\left(\frac{t}{T}\right)\varepsilon_t = \sum_{l=0}^{\infty} g_{\theta^{(\cdot)},l}\left(\frac{t}{T}\right)X_{t-l,T}. \tag{9.10}$$

Comparing (9.10) with (9.3), we see that

$$\beta^{\circ}_{\theta,t,T}(0) = b_{\theta^{(0)}}\left(\frac{t}{T}\right) \quad \text{and} \quad \alpha^{\circ}_{\theta,t,T}(l) = g_{\theta^{(\cdot)},l}\left(\frac{t}{T}\right).$$

Recalling the difference equations (9.7), the partial derivatives satisfy

$$\nabla_j g_{\theta^{(\cdot)},t}(u) = \begin{cases} \sum_{k=1}^{q}\frac{b_{\theta^{(k)}}(u)\widetilde{b}_0(u)}{b_{\theta^{(0)}}(u)^2}g_{\theta^{(\cdot)},t-k}(u) - \sum_{k=1}^{q}\frac{b_{\theta^{(k)}}(u)}{b_{\theta^{(0)}}(u)}\nabla_0 g_{\theta^{(\cdot)},t-k}(u) & (j=0) \\ -\frac{\widetilde{b}_j(u)}{b_{\theta^{(0)}}(u)}g_{\theta^{(\cdot)},t-j}(u) - \sum_{k=1}^{q}\frac{b_{\theta^{(k)}}(u)}{b_{\theta^{(0)}}(u)}\nabla_j g_{\theta^{(\cdot)},t-k}(u) & (j=1,\ldots,q). \end{cases}$$

Therefore, we obtain, under the null hypothesis $H(p;\theta)$,

$$h^{\top}\Delta_T(\theta) = \frac{-1}{\sqrt{T}\sigma}\sum_{t=q+1}^{T}\left[\varphi(\varepsilon_t)\sum_{j=1}^{q}\frac{h_j\widetilde{b}_j\left(\frac{t}{T}\right)}{\sigma}X_{t-j,T} + h_0\widetilde{b}_0\left(\frac{t}{T}\right)\{1+\varphi(\varepsilon_t)\varepsilon_t\}\right] + o_P(1). \tag{9.11}$$

9.4 Main Results

In this section, we derive the asymptotic properties of linear serial rank statistics for time-varying MA model in (9.4) under both null and alternative hypotheses, which is the main contribution of this paper.[1]

[1] The limiting distributions of test statistics under the alternative are important when we consider the power properties of the respective tests. Generally speaking, their derivations are considerably more difficult than those of the limiting distributions under the null hypotheses. Fortunately, in the

9.4.1 Linear Serial Rank Statistics for Time-Varying MA Model

Associated with the observed stretch $\boldsymbol{X}^{(T)} = \left(X_{1,T}^{(T)}, \ldots, X_{T,T}^{(T)}\right)$, we define the filtered series $\boldsymbol{Z}^{(T)} = \left(Z_{1,T}^{(T)}, \ldots, Z_{T,T}^{(T)}\right)$ as

$$Z_{t,T}^{(T)} = \widetilde{B}_{\theta^{(\cdot)}} \left(\frac{t}{T}, L\right)^{-1} X_{t,T}^{(T)},$$

where

$$\widetilde{B}_{\theta^{(\cdot)}} (u, L) := \sum_{k=0}^{q} b_{\theta^{(k)}} (u) L^k$$

and L is the lag operator. Under the null hypothesis $H_0^{(T)} := H(p; \theta)$, this filtered process becomes

$$Z_{t,T}^{(T)} = \sigma^{-1} X_{t,T}^{(T)}.$$

Then, we consider linear serial rank statistics, of order q, in the form

$$S^{(T)} := \frac{1}{T-q} \sum_{t=q+1}^{T} c_T \left(R_{t,T}^{(T)}, R_{t-1,T}^{(T)}, \ldots, R_{t-q,T}^{(T)}\right), \tag{9.12}$$

where $R_{t,T}^{(T)}$ denotes the rank of the filtered process $Z_{t,T}^{(T)}$ in the filtered series $\boldsymbol{Z}^{(T)} = \left(Z_{1,T}^{(T)}, \ldots, Z_{T,T}^{(T)}\right)$ of length T, and $c_T(\ldots)$ is some given score function. This test statistic is available even if the innovation variance σ^2 is unknown. We assume that the innovation process $\{\varepsilon_t \mid t \in \mathbb{Z}\}$ in (9.4) satisfies the following assumption, which is somewhat stronger assumption than that imposed for [6, Proposition 4.1] (see Assumption 9.1). Note that they assumed ε_t has finite moments up to the third order.

Assumption 9.5 ε_t has finite moments of all orders.

contiguous case, the difficulties are essentially diminished by LeCam's third lemma (see, e.g., [4, Sect. 7.1.4]).

9.4.2 Asymptotic Normality of Test Statistics for Time-Varying MA Model

We derive the asymptotic distribution under alternative hypothesis $K(p;\theta_T)$ of the linear serial rank statistics $S^{(T)}$ defined in (9.12). For this purpose, we first study the joint asymptotic normality of

$$\begin{pmatrix} \sqrt{T}\left(S^{(T)} - m^{(T)}\right) \\ \Lambda_T(\theta, \theta_T) \end{pmatrix}$$

under the null hypothesis $H_0^{(T)}$, where

$$m^{(T)} := E\left\{S^{(T)} \mid H_0^{(T)}\right\} = \frac{1}{T(T-1)\cdots(T-q)} \sum_{1\le i_1 \ne \cdots \ne i_{q+1} \le T} c_T\left(i_1, \ldots, i_{q+1}\right).$$

Then, the asymptotic normality of the rank statistics under $K(p;\theta_T)$ is automatically shown from LeCam's third lemma (see, e.g., [4, Sect. 7.1.4]).

Remember that under the null hypothesis $H_0^{(T)} := H(p;\theta)$, the observations satisfy $X_{t,T}^{(T)} = \sigma\varepsilon_t$. Let $U_t^{(T)} = P\left(X_{t,T}^{(T)}/\sigma\right)$ and $\boldsymbol{U}^{(T)} = \left(U_1^{(T)}, \ldots, U_T^{(T)}\right)^{\top}$.

The approximated statistics of $S^{(T)}$ and $m^{(T)}$ are proposed in [6, Sect. 4.1] and they showed that the asymptotic equivalence of $(T-q)^{1/2}\left(S^{(T)} - m^{(T)}\right)$ and $\left(\mathcal{S}^{(T)} - \mathcal{M}^{(T)}\right)$ (under $H_0^{(T)}$), where

$$\mathcal{S}^{(T)} := (T-q)^{-1/2} \sum_{t=q+1}^{T} J\left(P\left(X_{t,T}^{(T)}/\sigma\right), \ldots, P\left(X_{t-q,T}^{(T)}/\sigma\right)\right)$$

and

$$\mathcal{M}^{(T)} := \frac{(T-q)^{-1/2}}{T(T-1)\cdots(T-p)} \sum_{1\le t_1 \ne \cdots \ne t_{q+1} \le T} J\left(P\left(X_{t_1,T}^{(T)}/\sigma\right), \ldots, P\left(X_{t_{q+1},T}^{(T)}/\sigma\right)\right).$$

On the other hand, from the LAN result for time-varying MA model in (9.4) with the central sequence $\Delta_T(\theta)$ and the Fisher information matrix $\Gamma(\theta)$ given in (9.11) and (9.8), we obtain

$$\Lambda_T(\theta, \theta_T) = \mathcal{L}^{(T)} - \frac{\mathcal{F}(p)}{2\sigma^2} \sum_{j=0}^{q} \int_0^1 \widetilde{d}_j(u)^2\, du + o_P(1),$$

where

$$\mathcal{L}^{(T)} := \frac{-1}{\sqrt{T}\sigma} \sum_{t=q+1}^{T} \left[\varphi(\varepsilon_t) \sum_{j=1}^{q} h_j \widetilde{b}_j \left(\frac{t}{T}\right) \frac{X_{t-j,T}^{(T)}}{\sigma} + h_0 \widetilde{b}_0 \left(\frac{t}{T}\right) \{1 + \varphi(\varepsilon_t)\varepsilon_t\} \right]$$

and $\widetilde{d}_j(u)$ is defined in (9.9).

9.4.2.1 U-Statistics for the Score-Generating Function

We propose the U-statistics $\mathcal{U}_{\mathcal{S}}^{(T)}$ and $\mathcal{U}_{\mathcal{M}}^{(T)}$ such that $\mathcal{U}_{\mathcal{S}}^{(T)} - \mathcal{U}_{\mathcal{M}}^{(T)}$ is asymptotically equivalent to $T^{-1/2}\left(\mathcal{S}^{(T)} - \mathcal{M}^{(T)}\right)$. Define the $(q+1)$-dimensional random variables

$$\boldsymbol{Y}_t^{(T)} = \left(Y_{t,1}^{(T)}, \ldots, Y_{t,q+1}^{(T)}\right)^{\top} = \left(U_t^{(T)}, \ldots, U_{t-q}^{(T)}\right)^{\top}, \quad q+1 \le t \le T.$$

The $\boldsymbol{Y}_t^{(T)}$'s are identically distributed under $H_0^{(T)}$ (uniformly over $[0,1]^{q+1}$). Clearly, they are not independent—but, being q-dependent, and satisfy various mixing conditions. In [6, Sect. 4.2], the kernel

$$\begin{aligned}\Phi_{\mathcal{S}}\left(\boldsymbol{Y}_{t_1}^{(T)}, \ldots, \boldsymbol{Y}_{t_{q+1}}^{(T)}\right) &= \frac{1}{q+1} \sum_{j=1}^{q+1} J\left(\boldsymbol{Y}_{t_j}^{(T)}\right) \\ &= \frac{1}{q+1} \sum_{j=1}^{q+1} J\left(Y_{t_j,1}^{(T)}, \ldots, Y_{t_j,q+1}^{(T)}\right)\end{aligned}$$

provides a U-statistic which is asymptotically equivalent to $T^{-1/2}\mathcal{S}^{(T)}$;

$$\begin{aligned}\mathcal{U}_{\mathcal{S}}^{(T)} &= \binom{T-q}{q+1}^{-1} \sum_{q+1 \le t_1 < \cdots < t_{q+1} \le T} \Phi_{\mathcal{S}}\left(\boldsymbol{Y}_{t_1}^{(T)}, \ldots, \boldsymbol{Y}_{t_{q+1}}^{(T)}\right) \\ &= T^{-1/2}\mathcal{S}^{(T)} + o_P\left(T^{-1/2}\right).\end{aligned}$$

On the other hand, the kernel

$$\Phi_{\mathcal{M}}\left(\boldsymbol{Y}_{t_1}^{(T)}, \ldots, \boldsymbol{Y}_{t_{q+1}}^{(T)}\right) = \frac{1}{(q+1)!} \sum_{j} J\left(Y_{j_1,1}^{(T)}, \ldots, Y_{j_{q+1},1}^{(T)}\right),$$

where the summation $\sum_j$ extends over all possible $(q+1)!$ permutations $\left(j_1, \ldots, j_{q+1}\right)$ of $\left(t_1, \ldots, t_{q+1}\right)$, defines the corresponding U-statistic

$$\begin{aligned}\mathcal{U}_{\mathcal{M}}^{(T)} &= \binom{T-q}{q+1}^{-1} \sum_{q+1\le t_1<\cdots<t_{q+1}\le T} \Phi_{\mathcal{M}}\left(\boldsymbol{Y}_{t_1}^{(T)},\ldots,\boldsymbol{Y}_{t_{q+1}}^{(T)}\right)\\ &= \frac{1}{(T-q)\cdots(T-2q)} \sum_{q+1\le t_1\ne\cdots\ne t_{q+1}\le T} J\left(U_{t_1}^{(T)},\ldots,U_{t_{q+1}}^{(T)}\right)\\ &= T^{-1/2}\mathcal{M}^{(T)} + o_P\left(T^{-1/2}\right).\end{aligned}$$

Note that $E\left\{J^*\left(U_{q+1},\ldots,U_1\right)\right\}=0$, where J^* is defined in (9.2).

9.4.2.2 U-Statistics for the Central Sequence

We define the U-statistic $\mathcal{U}_{\mathcal{L}}^{(T)}$ associated with the central sequence $\Delta_T(\theta)$, which satisfies

$$\mathcal{U}_{\mathcal{L}}^{(T)} = T^{-1/2}h^{\top}\Delta_T(\theta) + o_{\mathcal{P}}\left(T^{-1/2}\right) = T^{-1/2}\mathcal{L}^{(T)} + o_P\left(T^{-1/2}\right).$$

Consider the function $W_{t,T}(\boldsymbol{y})$ (of $(q+1)$ arguments $\boldsymbol{y}=\left(y_1,\ldots,y_{q+1}\right)\in[0,1]^{q+1}$) where

$$\begin{aligned}W_{t,T}(\boldsymbol{y}) &= W_{t,T}\left(y_1,\ldots,y_{q+1}\right)\\ &= -\Bigg[\varphi\left(P^{-1}(y_1)\right)\sum_{j=1}^{q}\frac{h_j}{\sigma}\widetilde{b}_j\left(\frac{t}{T}\right)P^{-1}\left(y_{j+1}\right)\\ &\qquad + \frac{h_0}{\sigma}\widetilde{b}_0\left(\frac{t}{T}\right)\left\{1+\varphi\left(P^{-1}(y_1)\right)P^{-1}(y_1)\right\}\Bigg].\end{aligned}$$

Using this function $W_{t,T}$, we define a kernel of degree $(q+1)$ as

$$\Phi^{\mathcal{L}}_{(t_1,\ldots,t_{q+1}),T}\left(\boldsymbol{Y}_{t_1}^{(T)},\ldots,\boldsymbol{Y}_{t_{q+1}}^{(T)}\right) = \frac{1}{q+1}\sum_{j=1}^{q+1} W_{t_j,T}\left(\boldsymbol{Y}_{t_j}^{(T)}\right)$$

to provide a U-statistic which is asymptotically equivalent to $T^{-1/2}\mathcal{L}^{(T)}$;

$$\begin{aligned}\mathcal{U}_{\mathcal{L}}^{(T)} &= \binom{T-q}{q+1}^{-1} \sum_{q+1\le t_1<\cdots<t_{q+1}\le T} \Phi^{\mathcal{L}}_{(t_1,\ldots,t_{q+1}),T}\left(\boldsymbol{Y}_{t_1}^{(T)},\ldots,\boldsymbol{Y}_{t_{q+1}}^{(T)}\right)\\ &= T^{-1/2}\mathcal{L}^{(T)} + o_P\left(T^{-1/2}\right).\end{aligned}$$

9.4.2.3 Asymptotic Normality

From the above, $\mathcal{V}_{\alpha,\beta}^{(T)} = T^{-1/2}\left\{\alpha\left(\mathcal{S}^{(T)} - \mathcal{M}^{(T)}\right) + \beta\mathcal{L}^{(T)}\right\}$ is asymptotically equivalent (up to $o_{\mathcal{P}}\left(T^{-1/2}\right)$ terms) to a sequence of U-statistics

$$\mathcal{U}_{\alpha,\beta}^{(T)} = \binom{T-q}{q+1}^{-1} \sum_{q+1\le t_1<\cdots<t_{q+1}\le T} \Phi_{(t_1,\ldots,t_{q+1}),T}^{\alpha,\beta}\left(\boldsymbol{Y}_{t_1}^{(T)},\ldots,\boldsymbol{Y}_{t_{q+1}}^{(T)}\right)$$
$$= \frac{1}{(T-q)\cdots(T-2q)} \sum_{q+1\le t_1\neq\cdots\neq t_{q+1}\le T}$$
$$\int_{[0,1]^{(q+1)^2}} \Phi_{(t_1,\ldots,t_{q+1}),T}^{\alpha,\beta}\left(\boldsymbol{y}_{t_1},\ldots,\boldsymbol{y}_{t_{q+1}}\right)\prod_{j=1}^{q+1} d\delta\left(\boldsymbol{y}_{t_j} - \boldsymbol{Y}_{t_j}^{(T)}\right) \tag{9.13}$$

with kernel

$$\Phi_{(t_1,\ldots,t_{q+1}),T}^{\alpha,\beta} = \alpha\left(\Phi^{\mathcal{S}} - \Phi^{\mathcal{M}}\right) + \beta\Phi_{(t_1,\ldots,t_{q+1}),T}^{\mathcal{L}},$$

where α and β are arbitrary coefficients. Note that

$$T^{1/2}\left\{\mathcal{U}_{\alpha,\beta}^{(T)} - E\left(\mathcal{U}_{\alpha,\beta}^{(T)}\right)\right\}$$
$$= \alpha T^{1/2}\left\{S^{(T)} - m^{(T)}\right\} + \beta\left\{\Lambda_T(\theta,\theta_T) + \frac{\mathcal{F}(p)}{2\sigma^2}\sum_{j=0}^{q}\int_0^1 \widetilde{d}_j(u)^2\,du\right\} + o_P(1).$$

For every c $(1 \le c \le q+1)$, let

$$g_{(t_1,\ldots,t_{q+1}),T}^{\alpha,\beta,(c)}\left(\boldsymbol{y}_{t_1},\ldots,\boldsymbol{y}_{t_c}\right)$$
$$:= \int_{[0,1]^{(q-c+1)(q+1)}} \Phi_{(t_1,\ldots,t_{q+1}),T}^{\alpha,\beta}\left(\boldsymbol{y}_{t_1},\ldots,\boldsymbol{y}_{t_c},\boldsymbol{y}_{t_{c+1}},\ldots,\boldsymbol{y}_{t_{q+1}}\right)d\boldsymbol{y}_{t_{c+1}}\cdots d\boldsymbol{y}_{t_{q+1}},$$

so that $g_{(t_1,\ldots,t_{q+1}),T}^{\alpha,\beta,(q+1)} = \Phi_{(t_1,\ldots,t_{q+1}),T}^{\alpha,\beta}$,

$$\mathcal{U}_{\alpha,\beta}^{(T,c)} = \frac{1}{(T-q)\cdots(T-q-c+1)} \sum_{q+1\le t_1\neq\cdots\neq t_c\le T}$$
$$\int_{[0,1]^{(q+1)^2}} \Phi_{(t_1,\ldots,t_{q+1}),T}^{\alpha,\beta}\left(\boldsymbol{y}_{t_1},\ldots,\boldsymbol{y}_{t_{q+1}}\right)$$
$$\prod_{j=1}^{c}\left\{d\delta\left(\boldsymbol{y}_{t_j} - \boldsymbol{Y}_{t_j}^{(T)}\right) - d\boldsymbol{y}_{t_j}\right\} d\boldsymbol{y}_{t_{c+1}}\cdots d\boldsymbol{y}_{t_{q+1}}$$

$$= \frac{1}{(T-q)\cdots(T-q-c+1)} \sum_{q+1 \le t_1 \ne \cdots \ne t_c \le T}$$
$$\int_{[0,1]^{c(q+1)}} g^{\alpha,\beta,(c)}_{(t_1,\ldots,t_{q+1}),T}\left(\boldsymbol{y}_{t_1},\ldots,\boldsymbol{y}_{t_c}\right) \prod_{j=1}^{c} \left\{ d\delta\left(\boldsymbol{y}_{t_j} - \boldsymbol{Y}^{(T)}_{t_j}\right) - d\boldsymbol{y}_{t_j} \right\}$$

and

$$\mathcal{U}^{(T,0)}_{\alpha,\beta} := g^{\alpha,\beta,(0)}_{(t_1,\ldots,t_{q+1}),T} := \int_{[0,1]^{(q+1)^2}} \Phi^{\alpha,\beta}_{(t_1,\ldots,t_{q+1}),T}\left(\boldsymbol{y}_{t_1},\ldots,\boldsymbol{y}_{t_{q+1}}\right) d\boldsymbol{y}_{t_1}\cdots d\boldsymbol{y}_{t_{q+1}}$$
$$= E\left\{ \Phi^{\alpha,\beta}_{(t_1,\ldots,t_{q+1}),T}\left(\boldsymbol{Y}^{(T)}_{t_1},\ldots,\boldsymbol{Y}^{(T)}_{t_{q+1}}\right)\right\} = 0,$$

where $d\boldsymbol{y}_t$ stands for $dy_{t,1}\cdots dy_{t,q+1}$. We rewrite the U-statistic $\mathcal{U}^{(T)}_{\alpha,\beta}$ in (9.13) as

$$\mathcal{U}^{(T)}_{\alpha,\beta} = \mathcal{U}^{(T,0)}_{\alpha,\beta} + \sum_{c=1}^{q+1} \binom{q+1}{c} \mathcal{U}^{(T,c)}_{\alpha,\beta} \tag{9.14}$$

with

$$\mathcal{U}^{(T,1)}_{\alpha,\beta} = \frac{1}{(T-q)} \sum_{t=q+1}^{T} \int_{[0,1]^{(q+1)}} g^{\alpha,\beta,(1)}_{(t,t_2\ldots,t_{q+1}),T}\left(\boldsymbol{y}_t\right) \left\{ d\delta\left(\boldsymbol{y}_t - \boldsymbol{Y}^{(T)}_t\right) - d\boldsymbol{y}_t \right\}$$
$$= \frac{1}{(T-q)} \sum_{t=q+1}^{T} g^{\alpha,\beta,(1)}_{(t,t_2,\ldots,t_{q+1}),T}\left(\boldsymbol{Y}^{(T)}_t\right)$$

and

$$g^{\alpha,\beta,(1)}_{(t,t_2,\ldots,t_{q+1}),T}\left(\boldsymbol{Y}^{(T)}_t\right)$$
$$= \int_{[0,1]^{q(q+1)}} \Phi^{\alpha,\beta}_{(t,t_2,\ldots,t_{q+1}),T}\left(\boldsymbol{Y}^{(T)}_t, \boldsymbol{y}_{t_2},\ldots,\boldsymbol{y}_{t_{q+1}}\right) d\boldsymbol{y}_{t_2}\cdots d\boldsymbol{y}_{t_{q+1}}.$$

We are ready to state our main theorem.

Theorem 9.1 *Suppose that Assumptions 9.1–9.5 hold. Under the null hypothesis* $H^{(T)}_0 = H(p;\theta)$,

$$\begin{pmatrix} \sqrt{T}\left(S^{(T)} - m^{(T)}\right) \\ \Lambda_T(\theta,\theta_T) \end{pmatrix} \xrightarrow{\mathcal{L}} N\left(\widetilde{\boldsymbol{m}}, \widetilde{\boldsymbol{\Sigma}}\right),$$

with

$$\widetilde{\boldsymbol{m}} = \begin{pmatrix} 0 \\ -\frac{\mathcal{F}(p)}{2\sigma^2} \sum_{j=0}^{q} \int_0^1 \widetilde{d}_j(u)^2 \, du \end{pmatrix}$$

and

$$\widetilde{\boldsymbol{\Sigma}} = \begin{pmatrix} V^2 & -\sum_{j=0}^{q} \frac{C_j h_j}{\sigma} \int_0^1 \widetilde{b}_j(u) \, du \\ -\sum_{j=0}^{q} \frac{C_j h_j}{\sigma} \int_0^1 \widetilde{b}_j(u) \, du & \frac{\mathcal{F}(p)}{\sigma^2} \sum_{j=0}^{q} \int_0^1 \widetilde{d}_j(u)^2 \, du \end{pmatrix}, \tag{9.15}$$

where

$$\begin{aligned} V^2 := & \int_{[0,1]^{q+1}} \left\{ J^* \left(\nu_{q+1}, \ldots, \nu_1 \right) \right\}^2 d\nu_1 \cdots d\nu_{q+1} \\ & + 2 \sum_{l=1}^{q} \int_{[0,1]^{q+l+1}} J^* \left(\nu_{q+1}, \ldots, \nu_1 \right) J^* \left(\nu_{q+l+1}, \ldots, \nu_{l+1} \right) d\nu_1 \cdots d\nu_{q+l+1} \end{aligned} \tag{9.16}$$

and

$$C_j = \int_{[0,1]^{q+1}} J^* \left(\nu_{q+1}, \ldots, \nu_1 \right) \sum_{l=0}^{q-j} \varphi \left(P^{-1} \left(\nu_{q-l+1} \right) \right) P^{-1} \left(\nu_{q-l-j+1} \right) d\nu_1 \cdots d\nu_{q+1}. \tag{9.17}$$

Remark 9.1 Comparing Theorem 9.1 with Lemma 9.2, the asymptotic variance $\widetilde{\boldsymbol{\Sigma}}$ in (9.15) includes the affect of time-varying coefficients $\widetilde{b}_j(u)$. In addition, the term $\widetilde{b}_0(u)$ corresponding to $j = 0$ is also contained, which is the result of the time-varying (unconditional) variance $\{b_{\theta^{(0)}}(u)\}^2$ of innovation in the alternative time-varying MA process. This nontrivial effect has already appeared in the central sequence (9.11) and the Fisher information (9.8).

From LeCam's third lemma, we immediately obtain the following result.

Corollary 9.2 *Suppose that Assumptions 9.1–9.5 hold. Under $K(p; \theta_T)$,*

$$\sqrt{T} \left(S^{(T)} - m^{(T)} \right) \xrightarrow{\mathcal{L}} N \left(- \sum_{j=0}^{q} \frac{C_j h_j}{\sigma} \int_0^1 \widetilde{b}_j(u) \, du, V^2 \right).$$

Proof of Theorem 9.1 Since

$$\int_{[0,1]^{q(q+1)}} \Phi^{\mathcal{L}}_{(t, t_2, \ldots, t_{q+1}), T} \left(\boldsymbol{Y}_t^{(T)}, \boldsymbol{y}_{t_2}, \ldots, \boldsymbol{y}_{t_{q+1}} \right) d\boldsymbol{y}_{t_2} \cdots d\boldsymbol{y}_{t_{q+1}}$$

$$= \frac{1}{q+1}\left[W_{t,T}\left(\boldsymbol{Y}_t^{(T)}\right) + \sum_{j=2}^{q+1} E\left\{W_{t_j,T}\left(\boldsymbol{Y}_{t_j}^{(T)}\right)\right\}\right]$$

$$= -\frac{1}{q+1}\Bigg[\varphi\left(P^{-1}\left(U_t\right)\right)\sum_{j=1}^{q}\frac{h_j}{\sigma}\widetilde{b}_j\left(\frac{t}{T}\right)P^{-1}\left(U_{t-j}\right) + \frac{h_0}{\sigma}\widetilde{b}_0\left(\frac{t}{T}\right)\left\{1+\varphi\left(P^{-1}\left(U_t\right)\right)P^{-1}\left(U_t\right)\right\}\Bigg],$$

$$\int_{[0,1]^{q(q+1)}} \Phi^{\mathcal{S}}\left(\boldsymbol{Y}_t^{(T)}, \boldsymbol{y}_{t_2}, \ldots, \boldsymbol{y}_{t_{q+1}}\right) d\boldsymbol{y}_{t_2}\cdots d\boldsymbol{y}_{t_{q+1}}$$
$$= \frac{1}{q+1}\left\{J\left(\boldsymbol{Y}_t^{(T)}\right) + q\int_{[0,1]^{q+1}} J\left(\nu_{q+1}, \ldots, \nu_1\right) d\nu_1\cdots d\nu_{q+1}\right\}$$

and

$$\int_{[0,1]^{q(q+1)}} \Phi^{\mathcal{M}}\left(\boldsymbol{Y}_t^{(T)}, \boldsymbol{y}_{t_2}, \ldots, \boldsymbol{y}_{t_{q+1}}\right) d\boldsymbol{y}_{t_2}\cdots d\boldsymbol{y}_{t_{q+1}}$$
$$= \frac{1}{q+1}\sum_{k=1}^{q+1}\int_{[0,1]^q} J\left(\nu_q, \ldots, \nu_k, U_t, \nu_{k-1}, \ldots, \nu_1\right) d\nu_1\cdots d\nu_q,$$

we obtain

$$g^{\alpha,\beta,(1)}_{(t,t_2,\ldots,t_{q+1}),T}\left(\boldsymbol{Y}_t^{(T)}\right) = g^{\alpha,\beta,(1)}_{t,T}\left(\boldsymbol{Y}_t^{(T)}\right) = g^{\alpha,\beta,(1)}_{t,T}\left(U_t, \ldots, U_{t-q}\right)$$
$$= \frac{1}{q+1}\left\{\alpha J^*\left(U_t, \ldots, U_{t-q}\right) + \beta d_{t,T}\left(U_t, \ldots, U_{t-q}\right)\right\},$$

where

$$d_{t,T}\left(U_t, \ldots, U_{t-q}\right) := -\Bigg[\varphi\left(P^{-1}\left(U_t\right)\right)\sum_{j=1}^{q}\frac{h_j}{\sigma}\widetilde{b}_j\left(\frac{t}{T}\right)P^{-1}\left(U_{t-j}\right) + \frac{h_0}{\sigma}\widetilde{b}_0\left(\frac{t}{T}\right)\left\{1+\varphi\left(P^{-1}\left(U_t\right)\right)P^{-1}\left(U_t\right)\right\}\Bigg]$$
$$= -\left[\varphi\left(P^{-1}\left(U_t\right)\right)\sum_{j=0}^{q}\frac{h_j}{\sigma}\widetilde{b}_j\left(\frac{t}{T}\right)P^{-1}\left(U_{t-j}\right) + \frac{h_0}{\sigma}\widetilde{b}_0\left(\frac{t}{T}\right)\right].$$

Clearly, $E\left\{g^{\alpha,\beta,(1)}_{t,T}\left(\boldsymbol{Y}_t^{(T)}\right)\right\} = 0$. The variance of $g^{\alpha,\beta,(1)}_{t,T}\left(\boldsymbol{Y}_t^{(T)}\right)$ is

$$
\begin{aligned}
&E\left[\left\{g_{t,T}^{\alpha,\beta,(1)}\left(\boldsymbol{Y}_t^{(T)}\right)\right\}^2\right]\\
&=\frac{1}{(q+1)^2}\Bigg[\alpha^2\int_{[0,1]^{q+1}}\left\{J^*\left(\nu_{q+1},\ldots,\nu_1\right)\right\}^2 d\nu_1\cdots d\nu_{q+1}\\
&-2\alpha\beta\sum_{j=0}^{q}\frac{h_j}{\sigma}\widetilde{b}_j\left(\frac{t}{T}\right)\int_{[0,1]^{q+1}}J^*\left(\nu_{q+1},\ldots,\nu_1\right)\varphi\left(P^{-1}\left(\nu_{q+1}\right)\right)P^{-1}\left(\nu_{q-j+1}\right)\\
&\qquad d\nu_1\cdots d\nu_{q+1}\\
&+\beta^2\Bigg[\mathcal{F}(p)\sum_{j=1}^{q}\frac{h_j^2}{\sigma^2}\left\{\widetilde{b}_j\left(\frac{t}{T}\right)\right\}^2+\frac{h_0^2}{\sigma^2}\left\{\widetilde{b}_0\left(\frac{t}{T}\right)\right\}^2\left[E\left\{\varepsilon_t^2\varphi^2\left(\varepsilon_t\right)\right\}-1\right]\Bigg]\Bigg].
\end{aligned}
$$

Further, the covariances between $g_{t,T}^{\alpha,\beta,(1)}\left(\boldsymbol{Y}_t^{(T)}\right)$ and $g_{t+l,T}^{\alpha,\beta,(1)}\left(\boldsymbol{Y}_{t+l}^{(T)}\right)$ $(1\le l\le q)$ become

$$
\begin{aligned}
&E\left[g_{t,T}^{\alpha,\beta,(1)}\left(\boldsymbol{Y}_t^{(T)}\right)g_{t+l,T}^{\alpha,\beta,(1)}\left(\boldsymbol{Y}_{t+l}^{(T)}\right)\right]\\
&=\frac{1}{(q+1)^2}\Bigg[\alpha^2\int_{[0,1]^{q+l+1}}J^*\left(\nu_{q+1},\ldots,\nu_1\right)J^*\left(\nu_{q+l+1},\ldots,\nu_{l+1}\right)d\nu_1\cdots d\nu_{q+l+1}\\
&-\alpha\beta\sum_{j=0}^{q-l}\frac{h_j}{\sigma}\widetilde{b}_j\left(\frac{t}{T}\right)\int_{[0,1]^{q+1}}J^*\left(\nu_{q+1},\ldots,\nu_1\right)\varphi\left(P^{-1}\left(\nu_{q-l+1}\right)\right)P^{-1}\left(\nu_{q-l-j+1}\right)\\
&\qquad d\nu_1\cdots d\nu_{q+1}\Bigg].
\end{aligned}
$$

On the account of Assumptions 9.1(iii), 9.5 and (9.1),

$$
\begin{aligned}
\sigma_{\alpha,\beta}^2&:=\frac{1}{T}E\left(\sum_{t=q+1}^{T}g_{t,T}^{\alpha,\beta,(1)}\left(\boldsymbol{Y}_t^{(T)}\right)\sum_{s=q+1}^{T}g_{s,T}^{\alpha,\beta,(1)}\left(\boldsymbol{Y}_s^{(T)}\right)\right)\\
&=\frac{1}{T}\sum_{t=q+1}^{T}\left[E\left[\left\{g_{t,T}^{\alpha,\beta,(1)}\left(\boldsymbol{Y}_t^{(T)}\right)\right\}^2\right]+2\sum_{l=1}^{q}E\left[g_{t,T}^{\alpha,\beta,(1)}\left(\boldsymbol{Y}_t^{(T)}\right)g_{t+l,T}^{\alpha,\beta,(1)}\left(\boldsymbol{Y}_{t+l}^{(T)}\right)\right]\right]\\
&=\frac{1}{T(q+1)^2}\sum_{t=q+1}^{T}\left[\alpha^2V^2-2\alpha\beta\sum_{j=0}^{q}\frac{C_jh_j}{\sigma}\widetilde{b}_j\left(\frac{t}{T}\right)+\beta^2\frac{\mathcal{F}(p)}{\sigma^2}\sum_{j=0}^{q}\left\{\widetilde{d}_j\left(\frac{t}{T}\right)\right\}^2\right]\\
&=\frac{1}{(q+1)^2}\Bigg(\alpha^2V^2-2\alpha\beta\sum_{j=0}^{q}\frac{C_jh_j}{\sigma}\int_0^1\widetilde{b}_j(u)\,du\\
&\qquad+\beta^2\frac{\mathcal{F}(p)}{\sigma^2}\sum_{j=0}^{q}\int_0^1\widetilde{d}_j(u)^2\,du\Bigg)+o(1),
\end{aligned}
$$

where $\widetilde{d}_j(u)$, V^2 and C_j are defined in (9.9), (9.16) and (9.17), respectively, and

$$T^{1/2}\left\{\mathcal{U}_{\alpha,\beta}^{(T,1)} - E\left(\mathcal{U}_{\alpha,\beta}^{(T,1)}\right)\right\} \tag{9.18}$$

is asymptotically normal with mean zero and variance $\sigma_{\alpha,\beta}^2$.

Since $\mathcal{U}_{\alpha,\beta}^{(T,c)}$ $(c = 2, \ldots, q+1)$ in (9.14) satisfy

$$T^{1/2}\left\{\mathcal{U}_{\alpha,\beta}^{(T,c)} - E\left(\mathcal{U}_{\alpha,\beta}^{(T,c)}\right)\right\} = o_P(1)$$

(see [9, 15]), this complete the proof. □

9.5 Concluding Remarks

As mentioned before, Assumption 9.5 was somewhat stronger assumption than that imposed for [6, Proposition 4.1] (see Assumption 9.1), which is needed for the asymptotic normality of (9.18). We can expect to relax this assumption if we describe the proof in more detail, but we impose it for the sake of simplicity.

The extensions to the rank tests for randomness against time-varying ARMA alternative (see [6]) and those for a stationary ARMA model against time-varying ARMA alternative (see [5]) are also expected. In addition, the discussion about the asymptotic efficiency of linear rank statistics will develop the results. If we can add numerical experiments, it will be easier to see the results visually.

References

1. DAHLHAUS, R. (1996a). Maximum likelihood estimation and model selection for locally stationary processes. *Journal of Nonparametric Statistics* **6** 171–191.
2. DAHLHAUS, R. (1996b). On the Kullback–Leibler information divergence of locally stationary processes. *Stochastic Processes and their Applications* **62** 139–168.
3. HÁJEK, J. (1962). Asymptotically most powerful rank-order tests. *The Annals of Mathematical Statistics* **33** 1124–1147.
4. HÁJEK, J., ŠIDÁK, Z. AND SEN, P. K. (1999). *Theory of Rank Tests*. Elsevier Science, San Diego.
5. HALLIN, M. AND PURI, M. L. (1988). Optimal rank-based procedures for time series analysis: testing an ARMA model against other ARMA models. *The Annals of Statistics* **16** 402–432.
6. HALLIN, M., INGENBLEEK, J.-F. AND PURI, M. L. (1985). Linear serial rank tests for randomness against ARMA alternatives. *The Annals of Statistics* **13** 1156–1181.
7. HALLIN, M., TANIGUCHI, M., SERROUKH, A. AND CHOY, K. (1999). Local asymptotic normality for regression models with long-memory disturbance. *The Annals of Statistics* **27** 2054–2080.
8. HIRUKAWA, J. AND TANIGUCHI, M. (2006). LAN theorem for non-Gaussian locally stationary processes and its applications. *Journal of Statistical Planning and Inference* **136** 640–688.

9. Korolyuk, V. and Borovskich, Y. (2013). *Theory of U-Statistics*. Springer Netherlands.
10. Kreiss, J.- P. (1987). On adaptive estimation in stationary ARMA processes. *The Annals of Statistics* **15** 112–133.
11. Kreiss, J.- P. (1990). Local asymptotic normality for autoregression with infinite order. *Journal of Statistical Planning and Inference* **26** 185–219.
12. LeCam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer, UK.
13. Strasser, H. (1985). *Mathematical Theory of Statistics*. De Gruyter, Berlin.
14. Swensen, A. R. (1985). The asymptotic distribution of the likelihood ratio for autoregressive time series with a regression trend. *Journal of Multivariate Analysis* **16** 54–70.
15. Yoshihara, K. (1976). Limiting behavior of U-statistics for stationary, absolutely regular processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **35** 237–252.

Chapter 10
Asymptotic Expansions for Several GEL-Based Test Statistics and Hybrid Bartlett-Type Correction with Bootstrap

Yoshihide Kakizawa

Abstract This paper mainly discusses two issues about asymptotic expansions for the distributions of χ^2-type test statistics. First, it is shown that the generalized empirical likelihood ratio, Wald-type, and score-type test statistics for a subvector hypothesis in the possibly over-identified moment restrictions are, in general, not Bartlett-correctable, except for the empirical likelihood ratio test statistic. Second, starting with the classical likelihood or the modern generalized empirical likelihood, the Bartlett-type corrected test statistics, with the bootstrap procedure, are proposed to achieve a higher-order accurate testing inference for the nonparametric setup as well as the parametric setup.

10.1 Introduction

Higher-order accurate testing inference is a long-standing issue in the literature. This is one of the research questions that the author's teachers Professor Masanobu Taniguchi and Professor Minoru Siotani have continuously engaged. We refer the readers to, e.g., influential books [70, 71] on multivariate analysis and time series analysis.

Given a parametric model with suitable regularity conditions, the likelihood ratio (LR) test statistic $\mathrm{LR}^{(N)}$ is, under the null hypothesis, asymptotically chi-squared distributed, whose expectation is given in the form of

$$E^{(N)}[\mathrm{LR}^{(N)}] = f\left(1 - \frac{b}{N}\right) + o(N^{-1}),$$

where N is the sample size, f is the degrees of freedom of the test, and b is some constant. Hereafter, the symbol χ^2_ν stands for the chi-squared random variable with ν degrees of freedom. A simple mean adjustment of $\mathrm{LR}^{(N)}$ through multiplication by a constant $1 + b/N$ (equivalently, division by a constant $1 - b/N$) implies not only

Y. Kakizawa (✉)
Hokkaido University, Kita-ku, Sapporo, Hokkaido 060-0809, Japan
e-mail: kakizawa@econ.hokudai.ac.jp

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_10

$$E^{(N)}\Big[\Big(1+\frac{b}{N}\Big)\mathrm{LR}^{(N)}\Big]=f+o(N^{-1})$$

but also

$$P^{(N)}\Big[\Big(1+\frac{b}{N}\Big)\mathrm{LR}^{(N)}\le x\Big]=\Pr[\chi_f^2\le x]+o(N^{-1})\,.$$

The discovery of this kind of a phenomenon dates back to Bartlett [4] in his test for homogeneity of variances among two or more groups. Nowadays, such a mysterious statement (not merely the former meaning) for getting an accurate χ^2 approximation is widely known as the Bartlett-correctability (B-correctability) after Lawley [53] (see also Hayakawa [37] and Jensen [40]). Even in the case where the Bartlett factor b is unknown, the same technique can be applied by substituting a suitable estimator $b^{(N)}$ for b.

The B-correctability is a unique feature of $\mathrm{LR}^{(N)}$, and it does not hold for other test statistics, in general. As Cox [21] posed (see also Barndorff-Nielsen and Cox [3, p. 132]), it was an open issue (at that time) whether there is an effective general way of improving the large sample χ^2 approximations to the test statistics other than $\mathrm{LR}^{(N)}$. Possible solutions are "Bartlett-type corrections", which were developed in the three papers by CM, CF, and T (in alphabetical order), i.e., Chandra and Mukerjee [14], Cordeiro and Ferrari [20], and Taniguchi [71] from the likelihood inference of iid/non-identical models and Gaussian stationary time series models. The CM/T approaches were originally designed for a simple hypothesis about a scalar parameter in the absence of the nuisance parameter. Cribari-Neto and Cordeiro [23] gave an extensive historical review on Bartlett/Bartlett-type corrections, until 1996. We emphasize that the CM/T methods have been expanded to a subvector hypothesis even in the presence of the nuisance parameter (see Kakizawa [45–49]), parallel to the CF method.

In the above-mentioned works, Fisher's [26] likelihood was a starting point. Empirical likelihood (EL) method enables us to make a likelihood-type inference even in the nonparametric setup. Building on an earlier suggestion of Thomas and Grunkemeier [74], Owen [57] (see also his monograph [60]) developed the EL for univariate means, including certain univariate M-functionals. Hall and La Scala [31] and Owen [58] made an extension to smooth functions of mean vectors or M-functionals. Owen [59] discussed a triangular array version of the EL, where the one-way ANOVA and heteroscedastic regression models were considered in details. See also Kolaczyk [52] for a further application to generalized linear models. As a desirable large sample property, the EL ratio (ELR) was shown to be asymptotically χ^2 distributed, in parallel to Wilk's [76] theorem on the LR for the parametric models. The B-correctability of the EL was first established by DiCiccio et al. [24] for smooth functions of mean vectors. Zhang [77] examined the B-correctability of Owen's [57] EL for univariate M-functionals. Chen [15, 16] and Bravo [9] proved the B-correctability of Owen's [59] EL for linear regression models in the absence/presence of the nuisance parameter.

The EL method has been increasingly applied since Qin and Lawless [62, 63]. Their main focus was a general situation where the number of estimating equations (or moment restrictions)[1] is larger than the number of parameters to be inferenced. Chen and Cui [17] established the B-correctability of the ELR test statistic for testing a subvector hypothesis in the case of the just-identified moment restrictions (see Bravo [12] for testing a simple full vector parameter hypothesis). Even in the over-identified case, Chen and Cui [18] showed that the ELR test statistic for testing the full parameter vector is still B-correctable.

Continuous updating (CU) method and exponential tilting (ET) method, studied by Hansen et al. [33], Kitamura and Stutzer [50], respectively, are famous alternatives to the EL method. Generalized empirical likelihood (GEL), developed by Newey and Smith [56], encompasses an empirical counterpart of the Cressie and Read [22] class of power divergences, which includes the EL, CU, and ET as special cases. Here, we use the terminology of "the empirical Cressie–Read (ECR)" as the GEL dual problem rather than the minimum divergence problem; see Ragusa [64]. Several GEL-based test statistics have the same limiting χ^2 distribution.

The contribution of this paper is twofold: first, to study a valid N^{-1}-asymptotic expansion of several GEL-based test statistics for testing a subvector hypothesis in the possibly over-identified moment restrictions. After reviewing the GEL framework in Sect. 10.2, together with the required notations and assumptions, we show, in Sect. 10.3, that the ELR test statistic is, in general, the only member within the ECR-based test statistics that is B-correctable.[2] The finding is a nontrivial extension of Baggerly [2] to the GEL framework. See also Jing and Wood [41], Corcoran [19], Bravo [12], Camponovo and Otsu [13]. Second, in order to achieve the N^{-1}-accurate testing inference for the modern nonparametric setup as well as the classical parametric setup, we suggest, in Sect. 10.4, a hybrid Bartlett-type correction with the bootstrap procedure,[3] which is of independent interest.[4]

In the remainder of this paper, the symbol $^\top$ is the transpose of any matrix. We denote by $\mathbf{I}_\nu$, $\mathbf{0}_\nu$, and $\mathbf{O}_{\nu_1,\nu_2}$ the $\nu \times \nu$ identity matrix, the $\nu \times 1$ zero vector, and the $\nu_1 \times \nu_2$ zero matrix, respectively. Let $||\mathbf{v}|| = (\mathbf{v}^\top \mathbf{v})^{1/2}$ be the Euclidean norm of any

---

[1] The idea of using "estimating equations" has a long history in the literature, at least since Karl Pearson's introduction of the so-called method of moments. It seems that a motivation behind Qin and Lawless [62] is Godambe's optimal estimating equations theory. In the econometric literature, Hansen's [32] generalized method of moments is a basic inferential technique. We do not mention these two big methodologies anymore, to save space.

[2] The finding was first announced by the author at The Mathematical Society of Japan (Spring Meeting 2012).

[3] The bootstrap method was introduced by Efron [25], inspired by an earlier work on the jackknife. The methodology of the bootstrap is a computer-intensive technique that provides a basis for every field from modern statistical science. It is well recognized that its theoretical validity, as well as a deep understanding of the bootstrap procedure, is closely related to higher-order asymptotic statistical theory (e.g., Hall [30]); the details are omitted here.

[4] A hybrid method of "the bootstrap-based Bartlett-type adjustment" from the ordinary parametric likelihood/GEL testing inference was reported by the author at Nara-Symposium (September 2014) that Professor Taniguchi organized, and also at the Japanese Joint Statistical Meeting (September 2014).

finite-dimensional vector $\mathbf{v}$. For any $\nu \times \nu$ symmetric matrix $\mathbf{M} = [m_{ii'}]_{i,i' \in \{1,\ldots,\nu\}}$, let $\rho_1^{\uparrow}(\mathbf{M}) \le \cdots \le \rho_\nu^{\uparrow}(\mathbf{M})$ be the eigenvalues of $\mathbf{M}$, arranged in increasing order, and the spectral radius of $\mathbf{M}$ is defined by $\rho(\mathbf{M}) = \max\{|\rho_1^{\uparrow}(\mathbf{M})|, |\rho_\nu^{\uparrow}(\mathbf{M})|\}$. We also define the spectral norm of $\mathbf{M}$ (here, $\mathbf{M}$ is not necessary a symmetric matrix) as $||\mathbf{M}|| = \max_{||\mathbf{z}||=1} ||\mathbf{M}\mathbf{z}||$, which is equal to $\{\rho_\nu^{\uparrow}(\mathbf{M}^\top\mathbf{M})\}^{1/2}(= \rho(\mathbf{M})$ if $\mathbf{M} = \mathbf{M}^\top)$. It is easy to see that $||\mathbf{M}|| \le (\sum_{i,i'=1}^{\nu} m_{ii'}^2)^{1/2} \le \sum_{i,i'=1}^{\nu} |m_{ii'}|$ and that, for any nonsingular matrix $\mathbf{M}$, $||\mathbf{M}^{-1}|| = 1/\{\rho_1^{\uparrow}(\mathbf{M}^\top\mathbf{M})\}^{1/2}$. For ease of reference, a variant of Weyl's theorem (e.g., Horn and Johnson [38, p. 198]) states that

$$\max_{j \in \{1,\ldots,\nu\}} |\rho_j^{\uparrow}(\mathbf{M}+\mathbf{D}) - \rho_j^{\uparrow}(\mathbf{M})| \le \rho(\mathbf{D}) = ||\mathbf{D}|| \tag{10.1}$$

for any $\nu \times \nu$ symmetric matrices $\mathbf{M}$ and $\mathbf{D}$. Also, we recall the basic formula for the inverse of the partitioned (nonsingular) matrix

$$\mathbf{M} = \begin{pmatrix} \mathbf{M}_{(11)} & \mathbf{M}_{(12)} \\ \mathbf{M}_{(21)} & \mathbf{M}_{(22)} \end{pmatrix}$$

(we assume the nonsingularity of $\mathbf{M}_{(22)}$):

$$\mathbf{M}^{-1} = \begin{pmatrix} \mathbf{I}_{\nu_1} \\ -\mathbf{M}_{(22)}^{-1}\mathbf{M}_{(21)} \end{pmatrix} \mathbf{M}_{(11\cdot 2)}^{-1}(\mathbf{I}_{\nu_1} \ - \mathbf{M}_{(12)}\mathbf{M}_{(22)}^{-1}) + \begin{pmatrix} \mathbf{O}_{\nu_1,\nu_1} & \mathbf{O}_{\nu_1,\nu_2} \\ \mathbf{O}_{\nu_2,\nu_1} & \mathbf{M}_{(22)}^{-1} \end{pmatrix},$$

where $\mathbf{M}_{(11\cdot 2)} = \mathbf{M}_{(11)} - \mathbf{M}_{(12)}\mathbf{M}_{(22)}^{-1}\mathbf{M}_{(21)}$ is called the Schur complement of $\mathbf{M}_{(22)}$ in $\mathbf{M}$. Then,

$$\begin{pmatrix} \mathbf{M}_{(11)} & \mathbf{M}_{(12)} \\ \mathbf{M}_{(21)} & \mathbf{M}_{(22)} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{z}_{(1)} \\ \mathbf{0}_{\nu_2} \end{pmatrix} = \begin{pmatrix} \mathbf{I}_{\nu_1} \\ -\mathbf{M}_{(22)}^{-1}\mathbf{M}_{(21)} \end{pmatrix} \mathbf{M}_{(11\cdot 2)}^{-1}\mathbf{z}_{(1)}$$

and

$$(\mathbf{z}_{(1)}^\top \ \mathbf{0}_{\nu_2}^\top) \begin{pmatrix} \mathbf{M}_{(11)} & \mathbf{M}_{(12)} \\ \mathbf{M}_{(21)} & \mathbf{M}_{(22)} \end{pmatrix}^{-1} \begin{pmatrix} \mathbf{z}_{(1)} \\ \mathbf{0}_{\nu_2} \end{pmatrix} = \mathbf{z}_{(1)}^\top \mathbf{M}_{(11\cdot 2)}^{-1}\mathbf{z}_{(1)}\,.$$

10.2 Preliminaries

10.2.1 Statistical Setup and GEL

Let $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ be an iid sample of d_X-dimensional random vectors, drawn from an unknown distribution F. The symbol $P_F^{(N)}$ (and P_F) denotes the distribution of $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ (and $\mathbf{X} \sim F$), and the phrase "as the sample size N goes to the infinity" is dropped (it is implicit, unless otherwise stated).

Suppose that there is a p-dimensional parameter $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ associated with F, for which the interest is in a p_1-dimensional parameter $\boldsymbol{\theta}_{(1)} \in \boldsymbol{\Theta}_{(1)}$ under the presence of a p_2-dimensional nuisance parameter $\boldsymbol{\theta}_{(2)} \in \boldsymbol{\Theta}_{(2)}$, where $p = p_1 + p_2$, $\boldsymbol{\theta} = (\boldsymbol{\theta}_{(1)}^\top, \boldsymbol{\theta}_{(2)}^\top)^\top$, and $\boldsymbol{\Theta} = \boldsymbol{\Theta}_1 \times \boldsymbol{\Theta}_2 \subset \mathbb{R}^{p_1} \times \mathbb{R}^{p_2}$. More precisely, given a known function

$$\mathbf{g}(\mathbf{x}, \boldsymbol{\theta}) = [g_\beta(\mathbf{x}, \boldsymbol{\theta})]_{\beta \in \{1,\ldots,M\}} ,$$

the information about $\boldsymbol{\theta}$ is summarized through the moment restrictions, such that

$$E_F[\mathbf{g}(\mathbf{X}, \boldsymbol{\theta}_0)] = \mathbf{0}_M \text{ for a unique } \boldsymbol{\theta}_0 = (\boldsymbol{\theta}_{(1)0}^\top, \boldsymbol{\theta}_{(2)0}^\top)^\top \in \boldsymbol{\Theta} ,$$

where the symbol E_F stands for expectation taken with respect to $\mathbf{X} \sim F$. We assume $M \geq p$, i.e., the just-identified case $M = p$ or the over-identified case $M > p$.

To describe the GEL objective function (Newey and Smith [56]), let $\rho(\cdot)$ be a (smooth) function that is concave on its domain $\mathcal{V}_\rho$; an open interval containing zero. With $\rho_i(v) = (d^i/dv^i)\rho(v)$, we assume that $\rho(0) = 0$ and $\rho_1(0) = \rho_2(0) = -1$, since, as long as $\rho_1 \neq 0$ and $\rho_2 < 0$, where $\rho_1 = \rho_1(0)$ and $\rho_2 = \rho_2(0)$, we can replace $\rho(v)$ by $-(\rho_2/\rho_1^2)[\rho\{(\rho_1/\rho_2)v\} - \rho(0)]$. Here are a few practical examples:

[EL] $\rho_{\mathrm{EL}}(v) = \log(1 - v)$ and $\mathcal{V}_{\mathrm{EL}} = (-\infty, 1)$;
[CU] $\rho_{\mathrm{CU}}(v) = -(v + v^2/2)$ and $\mathcal{V}_{\mathrm{CU}} = \mathbb{R}$;
[ET] $\rho_{\mathrm{ET}}(v) = -\mathrm{e}^v + 1$ and $\mathcal{V}_{\mathrm{ET}} = \mathbb{R}$.

The GEL is a comprehensive framework, indexed by such a normalized ρ-function, hence, the resulting GEL-based test statistics are necessarily equivalent in the large sample theory, but possess different N^{-1}-asymptotics via $\rho_3 = \rho_3(0)$ and $\rho_4 = \rho_4(0)$ generally, as will be shown later.

Introducing an $M \times 1$ vector of auxiliary parameters $\boldsymbol{\lambda}$, let $\boldsymbol{\eta} = (\boldsymbol{\theta}^\top, \boldsymbol{\lambda}^\top)^\top$ be a $(p + M) \times 1$ vector of augmented parameters to be inferenced. A GEL saddlepoint problem $\inf_{\theta \in \Theta} \sup_{\lambda \in \Lambda_{\theta,N}^\rho} \ell^{(N)\rho}(\boldsymbol{\eta})$ was formulated by Newey and Smith [56], where

$$\ell^{(N)\rho}(\boldsymbol{\eta}) = \frac{1}{N} \sum_{i=1}^{N} \rho\{\boldsymbol{\lambda}^\top \mathbf{g}(\mathbf{X}_i, \boldsymbol{\theta})\} ,$$

with $\Lambda_{\theta,N}^\rho = \{\boldsymbol{\lambda} \in \mathbb{R}^M : \boldsymbol{\lambda}^\top \mathbf{g}(\mathbf{X}_i, \boldsymbol{\theta}) \in \mathcal{V}_\rho\ ,\ i \in \{1, \ldots, N\}\}$. Newey and Smith [56] and Ragusa [64] pointed out that a subclass of criteria based on the Cressie and Read [22] power divergence can be treated by using the following ρ-function;

$$\rho_{\mathrm{CR}}(v) = \begin{cases} -\dfrac{1}{\omega + 1}\{(1 + \omega v)^{(\omega+1)/\omega} - 1\} , & \omega \in \mathbb{R}\backslash\{-1, 0\} , \\ \log(1 - v) , & \omega = -1\, ;\ \text{EL}, \\ -\mathrm{e}^v + 1 , & \omega = 0\, ;\ \text{ET}, \end{cases}$$

whose domain $\mathcal{V}_{\mathrm{CR}}$ depends on the parameter ω (the CU is a special case of $\omega = 1$). Note that

$$\rho'''_{\mathrm{CR}}(0) = -(1-\omega) \quad \text{and} \quad \rho''''_{\mathrm{CR}}(0) = -(1-\omega)(1-2\omega).$$

For testing a null hypothesis $\boldsymbol{\theta}_{(1)} = \boldsymbol{\theta}_{(1)0}$, we solve the unrestricted GEL first-order condition

$$\boldsymbol{\ell}^{(N)\rho}\begin{pmatrix}\widehat{\boldsymbol{\theta}}^{(N)\rho}_{(1)}\\ \widehat{\boldsymbol{\theta}}^{(N)\rho}_{(2)}\\ \widehat{\boldsymbol{\lambda}}^{(N)\rho}\end{pmatrix} = \mathbf{0}_{p+M}$$

or the restricted GEL first-order condition

$$\boldsymbol{\ell}^{(N)\rho}_{(2)}\begin{pmatrix}\boldsymbol{\theta}_{(1)0}\\ \widetilde{\boldsymbol{\theta}}^{(N)\rho}_{(2)}\\ \widetilde{\boldsymbol{\lambda}}^{(N)\rho}\end{pmatrix} = \mathbf{0}_{p_2+M}$$

(by the hat, it is meant to be unrestricted; $\widehat{\boldsymbol{\eta}}^{(N)\rho} = ((\widehat{\boldsymbol{\theta}}^{(N)\rho}_{(1)})^\top, (\widehat{\boldsymbol{\theta}}^{(N)\rho}_{(2)})^\top, (\widehat{\boldsymbol{\lambda}}^{(N)\rho})^\top)^\top$, and, by the tilde, it is meant to be restricted; $\widetilde{\boldsymbol{\eta}}^{(N)\rho} = (\boldsymbol{\theta}^\top_{(1)0}, (\widetilde{\boldsymbol{\theta}}^{(N)\rho}_{(2)})^\top, (\widetilde{\boldsymbol{\lambda}}^{(N)\rho})^\top)^\top$), where $\boldsymbol{\ell}^{(N)\rho}(\boldsymbol{\eta}) = [\ell^{(N)\rho}_j(\boldsymbol{\eta})]_{j\in\{1,\ldots,p+M\}}$, $\boldsymbol{\ell}^{(N)\rho}_{(2)}(\boldsymbol{\eta}) = [\ell^{(N)\rho}_r(\boldsymbol{\eta})]_{r\in\{p_1+1,\ldots,p+M\}}$, with

$$\ell^{(N)\rho}_\alpha(\boldsymbol{\eta}) = \frac{1}{N}\sum_{i=1}^N \rho_1\{\boldsymbol{\lambda}^\top\mathbf{g}(\mathbf{X}_i,\boldsymbol{\theta})\}\boldsymbol{\lambda}^\top\frac{\partial}{\partial\theta_\alpha}\mathbf{g}(\mathbf{X}_i,\boldsymbol{\theta}), \quad \alpha\in\{1,\ldots,p\},$$

$$\ell^{(N)\rho}_{[\beta]}(\boldsymbol{\eta}) = \frac{1}{N}\sum_{i=1}^N \rho_1\{\boldsymbol{\lambda}^\top\mathbf{g}(\mathbf{X}_i,\boldsymbol{\theta})\}g_\beta(\mathbf{X}_i,\boldsymbol{\theta}), \quad \beta\in\{1,\ldots,M\}$$

(we set $[\beta] = p+\beta$). A crucial point of the N^{-1}-asymptotic theory is that, under certain conditions, both $\widehat{\boldsymbol{\eta}}^{(N)\rho}$ and $\widetilde{\boldsymbol{\eta}}^{(N)\rho}$ are well defined, with probability $1-o(N^{-1})$, lying in an $\epsilon^{(N)\rho}$-neighborhood of $\boldsymbol{\eta}_0 = (\boldsymbol{\theta}_0^\top, \mathbf{0}_M^\top)^\top$, where $\epsilon^{(N)\rho} \propto N^{-1/2}(\log N)^{1/2}$. See Lemma 10.4 of Sect. 10.6.1 for more details, including the stochastic expansions of $\widehat{\boldsymbol{\eta}}^{(N)\rho}$ and $\widetilde{\boldsymbol{\eta}}^{(N)\rho}$ based on Bhattacharya and Ghosh's [6] argument and Taniguchi [71, p. 76]. If there is no confusion, for any (nonrandom or random) scalar or vector or matrix function; $H(\cdot)$ or $H^{(N)}(\cdot)$, we write, e.g., $\widehat{H}$, $\widetilde{H}$, and H instead of $H(\widehat{\boldsymbol{\eta}}^{(N)\rho})$, $H(\widetilde{\boldsymbol{\eta}}^{(N)\rho})$, and $H(\boldsymbol{\eta}_0)$, respectively.

Not surprisingly, in parallel to the trinity of the three test statistics from the classical likelihood, the GELR test statistic $\mathrm{ELR}^{(N)\rho} = 2N(\widetilde{\ell}^{(N)\rho} - \widehat{\ell}^{(N)\rho})$, and the Wald-type and score-type test statistics $\mathrm{W}^{(N)\rho}$ and $\mathrm{S}^{(N)\rho}$, defined as quadratic forms of $\widehat{\boldsymbol{\theta}}^{(N)\rho}_{(1)} - \boldsymbol{\theta}_{(1)0}$ and $[\widetilde{\ell}^{(N)\rho}_a]_{a\in\{1,\ldots,p_1\}}$, respectively, are available even from the present GEL framework, and they have the same limiting χ^2 distribution under suitable regularity conditions. For more details, see Sect. 10.3.

Remark 10.1 Unless otherwise stated, we use the letter "j" or "k" as the indices of the full vector $\boldsymbol{\eta} = (\boldsymbol{\theta}^\top_{(1)}, \boldsymbol{\theta}^\top_{(2)}, \boldsymbol{\lambda}^\top)^\top$ that runs from 1 to $p+M$, the letter "a" or "b" as the indices of the subvector $\boldsymbol{\eta}_{(1)} = \boldsymbol{\theta}_{(1)}$ of $\boldsymbol{\eta}$ that runs from 1 to p_1, and the letter "r" as the indices of the subvector $\boldsymbol{\eta}_{(2)} = (\boldsymbol{\theta}_{(2)}^\top, \boldsymbol{\lambda}^\top)^\top$ of $\boldsymbol{\eta}$ that runs from $p_1 + 1$ to $p + M$. These distinctions are useful for two purposes in the present paper, first to denote a typical element of any (nonrandom or random) R-way array $[H_{j_1 \cdots j_R}]_{j_1,\ldots,j_R \in \{1,\ldots,p+M\}}$ or $[H^{(N)}_{j_1 \cdots j_R}]_{j_1,\ldots,j_R \in \{1,\ldots,p+M\}}$, and second to indicate the range of a sum in the Einstein summation convention.

10.2.2 Notation

The Rth partial derivatives of $\ell^\rho(\boldsymbol{\eta}) = \rho\{\boldsymbol{\lambda}^\top \mathbf{g}(\mathbf{X}, \boldsymbol{\theta})\}$ and $\ell^{(N)\rho}(\boldsymbol{\eta})$ with respect to $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_{p+M})^\top (= (\boldsymbol{\theta}^\top, \boldsymbol{\lambda}^\top)^\top)$ are denoted by

$$\ell^\rho_{j_1 \cdots j_R}(\boldsymbol{\eta}) = \frac{\partial^R \ell^\rho(\boldsymbol{\eta})}{\partial \eta_{j_1} \cdots \partial \eta_{j_R}} \quad \text{and} \quad \ell^{(N)\rho}_{j_1 \cdots j_R}(\boldsymbol{\eta}) = \frac{\partial^R \ell^{(N)\rho}(\boldsymbol{\eta})}{\partial \eta_{j_1} \cdots \partial \eta_{j_R}},$$

respectively, where $j_1, \ldots, j_R \in \{1, \ldots, p + M\}$. For the sake of brevity, we may write $I_R = j_1 \cdots j_R$, where $R = |I_R|$ is the length of I_R.

In order to distinguish $(\eta_1, \ldots, \eta_p)^\top (= \boldsymbol{\theta})$ and $(\eta_{p+1}, \ldots, \eta_{p+M})^\top (= \boldsymbol{\lambda})$, we use the letter "$\alpha$" as the indices running from 1 to p, compared with the definition $[\beta] = p + \beta$, in which "β" runs from 1 to M. After some algebra, we can see that the partial derivatives (up to four) of $\ell^\rho(\boldsymbol{\eta})$ evaluated at $\boldsymbol{\eta} = \boldsymbol{\eta}_0 (= (\boldsymbol{\theta}_0^\top, \mathbf{0}_M^\top)^\top)$ are explicitly given, as follows (we sometimes suppress the superscript "ρ" for the quantities that are independent of the ρ-function): For $\alpha_1 \in \{1, \ldots, p\}$ and $\beta_1, \beta_2, \beta_3, \beta_4 \in \{1, \ldots, M\}$, $\ell_{\alpha_1} = \ell_{\alpha_1\alpha_2} = \ell_{\alpha_1\alpha_2\alpha_3} = \ell_{\alpha_1\alpha_2\alpha_3\alpha_4} = 0$,

$$\begin{aligned}
\ell_{[\beta_1]} &= -g_{\beta_1}(\mathbf{X}, \boldsymbol{\theta}_0), \quad \ell_{[\beta_1][\beta_2]} = -\prod_{i=1}^{2} g_{\beta_i}(\mathbf{X}, \boldsymbol{\theta}_0), \\
\ell^\rho_{[\beta_1][\beta_2][\beta_3]} &= \rho_3 \prod_{i=1}^{3} g_{\beta_i}(\mathbf{X}, \boldsymbol{\theta}_0), \quad \ell^\rho_{[\beta_1][\beta_2][\beta_3][\beta_4]} = \rho_4 \prod_{i=1}^{4} g_{\beta_i}(\mathbf{X}, \boldsymbol{\theta}_0), \\
\ell_{\alpha_1[\beta_2]} &= -\frac{\partial g_{\beta_2}(\mathbf{X}, \boldsymbol{\theta}_0)}{\partial \theta_{\alpha_1}}, \quad \ell_{\alpha_1[\beta_2][\beta_3]} = -\frac{\partial \prod_{i=2}^{3} g_{\beta_i}(\mathbf{X}, \boldsymbol{\theta}_0)}{\partial \theta_{\alpha_1}}, \qquad (10.2) \\
\ell^\rho_{\alpha_1[\beta_2][\beta_3][\beta_4]} &= \rho_3 \frac{\partial \prod_{i=2}^{4} g_{\beta_i}(\mathbf{X}, \boldsymbol{\theta}_0)}{\partial \theta_{\alpha_1}}, \quad \ell_{\alpha_1\alpha_2[\beta_3]} = -\frac{\partial^2 g_{\beta_3}(\mathbf{X}, \boldsymbol{\theta}_0)}{\partial \theta_{\alpha_2} \partial \theta_{\alpha_2}}, \\
\ell_{\alpha_1\alpha_2[\beta_3][\beta_4]} &= -\frac{\partial^2 \prod_{i=3}^{4} g_{\beta_i}(\mathbf{X}, \boldsymbol{\theta}_0)}{\partial \theta_{\alpha_1} \partial \theta_{\alpha_2}}, \quad \ell_{\alpha_1\alpha_2\alpha_3[\beta_4]} = -\frac{\partial^3 g_{\beta_4}(\mathbf{X}, \boldsymbol{\theta}_0)}{\partial \theta_{\alpha_1} \partial \theta_{\alpha_2} \partial \theta_{\alpha_3}}.
\end{aligned}$$

For $\nu \in \{1, 2, 3, 4\}$ and $1 \le \sum_{i=1}^{\nu} R_i \le 4$, the joint cumulants of the $\ell^\rho_{I_{R_i}}$'s, defined by

$$\nu^\rho{}_{I_{R_1},\ldots,I_{R_\nu}} = Cum_F(\ell^\rho_{I_{R_1}}, \ldots, \ell^\rho_{I_{R_\nu}}),$$

Table 10.1 The νth cumulants $\nu^{\rho}{}_{I_{R_1},\ldots,I_{R_\nu}}$, for $\nu \in \{1,2,3,4\}$ and $1 \le \sum_{i=1}^{\nu} R_i \le 4$, where the arguments "$\mathbf{X}, \boldsymbol{\theta}_0$" from $g_{\#}$ and its higher-order derivatives, e.g., $g_{\beta/\alpha}$, are omitted here, for simplicity

Means
$\nu_{\alpha_1} = \nu_{\alpha_1\alpha_2} = \nu_{\alpha_1\alpha_2\alpha_3} = \nu_{\alpha_1\alpha_2\alpha_3\alpha_4} = \nu_{[\beta_1]} = 0,\ \nu_{[\beta_1][\beta_2]} = -E_F[g_{\beta_1} g_{\beta_2}],$
$\nu^{\rho}{}_{[\beta_1][\beta_2][\beta_3]} = \rho_3 E_F[g_{\beta_1} g_{\beta_2} g_{\beta_3}],\ \nu^{\rho}{}_{[\beta_1][\beta_2][\beta_3][\beta_4]} = \rho_4 E_F[g_{\beta_1} g_{\beta_2} g_{\beta_3} g_{\beta_4}],$
$\nu_{\alpha_1[\beta_2]} = -E_F[g_{\beta_2/\alpha_1}],\ \nu_{\alpha_1[\beta_2][\beta_3]} = -E_F\left[\frac{\partial(g_{\beta_2} g_{\beta_3})}{\partial\theta_{\alpha_1}}\right],$
$\nu^{\rho}{}_{\alpha_1[\beta_2][\beta_3][\beta_4]} = \rho_3 E_F\left[\frac{\partial(g_{\beta_2} g_{\beta_3} g_{\beta_4})}{\partial\theta_{\alpha_1}}\right],\ \nu_{\alpha_1\alpha_2[\beta_3]} = -E_F[g_{\beta_3/\alpha_1\alpha_2}],$
$\nu_{\alpha_1\alpha_2[\beta_3][\beta_4]} = -E_F\left[\frac{\partial^2(g_{\beta_3} g_{\beta_4})}{\partial\theta_{\alpha_1}\partial\theta_{\alpha_2}}\right],\ \nu_{\alpha_1\alpha_2\alpha_3[\beta_4]} = -E_F[g_{\beta_4/\alpha_1\alpha_2\alpha_3}];$
Covariances
$\nu_{\alpha_1,\alpha_2} = \nu_{\alpha_1,[\beta_2]} = 0,\ \nu_{[\beta_1],[\beta_2]} = E_F[g_{\beta_1} g_{\beta_2}],$
$\nu_{\alpha_1\alpha_2,\alpha_3} = \nu_{\alpha_1\alpha_2,[\beta_3]} = \nu_{\alpha_1[\beta_2],\alpha_3} = \nu_{[\beta_1][\beta_2],\alpha_3} = 0,$
$\nu_{\alpha_1[\beta_2],[\beta_3]} = E_F[g_{\beta_2/\alpha_1} g_{\beta_3}],\ \nu_{[\beta_1][\beta_2],[\beta_3]} = E_F[g_{\beta_1} g_{\beta_2} g_{\beta_3}],$
$\nu_{\alpha_1\alpha_2\alpha_3,\alpha_4} = \nu_{\alpha_1\alpha_2[\beta_3],\alpha_4} = \nu_{\alpha_1[\beta_2][\beta_3],\alpha_4} = \nu_{[\beta_1][\beta_2][\beta_3],\alpha_4} = \nu_{\alpha_1\alpha_2\alpha_3,[\beta_4]} = 0,$
$\nu_{\alpha_1\alpha_2[\beta_3],[\beta_4]} = E_F[g_{\beta_3/\alpha_1\alpha_2} g_{\beta_4}],\ \nu_{\alpha_1[\beta_2][\beta_3],[\beta_4]} = E_F\left[\frac{\partial(g_{\beta_2} g_{\beta_3})}{\partial\theta_{\alpha_1}} g_{\beta_4}\right],$
$\nu^{\rho}{}_{[\beta_1][\beta_2][\beta_3],[\beta_4]} = -\rho_3 E_F[g_{\beta_1} g_{\beta_2} g_{\beta_3} g_{\beta_4}],$
$\nu_{\alpha_1\alpha_2,\alpha_3\alpha_4} = \nu_{\alpha_1\alpha_2,\alpha_3[\beta_4]} = \nu_{\alpha_1\alpha_2,[\beta_3][\beta_4]} = 0,$
$\nu_{\alpha_1[\beta_2],\alpha_3[\beta_4]} = E_F[g_{\beta_2/\alpha_1} g_{\beta_4/\alpha_3}] - E_F[g_{\beta_2/\alpha_1}]E_F[g_{\beta_4/\alpha_3}],$
$\nu_{\alpha_1[\beta_2],[\beta_3][\beta_4]} = -E_F[g_{\beta_2/\alpha_1} g_{\beta_3} g_{\beta_4}] + E_F[g_{\beta_2/\alpha_1}]E_F[g_{\beta_3} g_{\beta_4}],$
$\nu_{[\beta_1][\beta_2],[\beta_3][\beta_4]} = E_F[g_{\beta_1} g_{\beta_2} g_{\beta_3} g_{\beta_4}] - E_F[g_{\beta_1} g_{\beta_2}]E_F[g_{\beta_3} g_{\beta_4}];$
3rd cumulants
$\nu_{\alpha_1,\alpha_2,\alpha_3} = \nu_{\alpha_1,\alpha_2,[\beta_3]} = \nu_{\alpha_1,[\beta_2],[\beta_3]} = 0,\ \nu_{[\beta_1],[\beta_2],[\beta_3]} = -E_F[g_{\beta_1} g_{\beta_2} g_{\beta_3}],$
$\nu_{\alpha_1\alpha_2,\alpha_3,\alpha_4} = \nu_{\alpha_1[\beta_2],\alpha_3,\alpha_4} = \nu_{[\beta_1][\beta_2],\alpha_3,\alpha_4} = 0,$
$\nu_{\alpha_1\alpha_2,\alpha_3,[\beta_4]} = \nu_{\alpha_1[\beta_2],\alpha_3,[\beta_4]} = \nu_{[\beta_1][\beta_2],\alpha_3,[\beta_4]} = \nu_{\alpha_1\alpha_2,[\beta_3],[\beta_4]} = 0,$
$\nu_{\alpha_1[\beta_2],[\beta_3],[\beta_4]} = -E_F[g_{\beta_2/\alpha_1} g_{\beta_3} g_{\beta_4}] + E_F[g_{\beta_2/\alpha_1}]E_F[g_{\beta_3} g_{\beta_4}],$
$\nu_{[\beta_1][\beta_2],[\beta_3],[\beta_4]} = -E_F[g_{\beta_1} g_{\beta_2} g_{\beta_3} g_{\beta_4}] + E_F[g_{\beta_1} g_{\beta_2}]E_F[g_{\beta_3} g_{\beta_4}];$
4th cumulants
$\nu_{\alpha_1,\alpha_2,\alpha_3,\alpha_4} = \nu_{\alpha_1,\alpha_2,\alpha_3,[\beta_4]} = \nu_{\alpha_1,\alpha_2,[\beta_3],[\beta_4]} = \nu_{\alpha_1,[\beta_2],[\beta_3],[\beta_4]} = 0,$
$\nu_{[\beta_1],[\beta_2],[\beta_3],[\beta_4]} = E_F[g_{\beta_1} g_{\beta_2} g_{\beta_3} g_{\beta_4}] - \langle 3\rangle E_F[g_{\beta_1} g_{\beta_2}]E_F[g_{\beta_3} g_{\beta_4}].$

are given in Table 10.1, where the Rth partial derivative of $g_\beta(\mathbf{x}, \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ is denoted by

$$g_{\beta/\alpha_1\cdots\alpha_R}(\mathbf{x}, \boldsymbol{\theta}) = \frac{\partial^R g_\beta(\mathbf{x}, \boldsymbol{\theta})}{\partial\theta_{\alpha_1}\cdots\partial\theta_{\alpha_R}}, \quad \alpha_1, \ldots, \alpha_R \in \{1, \ldots, p\},\ \beta \in \{1, \ldots, M\}.$$

We observe that the four patterns $\nu^{\rho}{}_{[\beta_1][\beta_2][\beta_3]}$, $\nu^{\rho}{}_{\alpha_1[\beta_2][\beta_3][\beta_4]}$, $\nu^{\rho}{}_{[\beta_1][\beta_2][\beta_3][\beta_4]}$, and $\nu^{\rho}{}_{[\beta_1][\beta_2][\beta_3],[\beta_4]}$ depend on the choice of $\rho(\cdot)$ through the third and fourth derivatives evaluated at $v = 0$; $\rho_3 = \rho_3(0)$ and $\rho_4 = \rho_4(0)$, whereas the others are independent of $\rho(\cdot)$.

Importantly, we always have the identities

$$\nu_{[\beta_1]} = 0\,, \quad \nu_{[\beta_1][\beta_2]} + \nu_{[\beta_1],[\beta_2]} = 0 \quad \text{for}\, \beta_1, \beta_2 \in \{1, \ldots, M\}. \tag{10.3}$$

However, the *partial* Bartlett identities (with respect to $\boldsymbol{\lambda}$) do not hold generally, unless $(\rho_3, \rho_4) = (-2, -6)$. That is, for $\beta_1, \beta_2, \beta_3, \beta_4 \in \{1, \ldots, M\}$,

$$\begin{aligned}
&\nu^{\rho}{}_{[\beta_1][\beta_2][\beta_3]} + \langle 3\rangle \nu_{[\beta_1][\beta_2],[\beta_3]} + \nu_{[\beta_1],[\beta_2],[\beta_3]} \\
&= (\rho_3 + 2) E_F\Big[\prod_{i=1}^{3} g_{\beta_i}(\mathbf{X}, \boldsymbol{\theta}_0)\Big], &(10.4)\\
&\nu^{\rho}{}_{[\beta_1][\beta_2][\beta_3][\beta_4]} + \langle 4\rangle \nu^{\rho}{}_{[\beta_1][\beta_2][\beta_3],[\beta_4]} \\
&\quad + \langle 3\rangle \nu_{[\beta_1][\beta_2],[\beta_3][\beta_4]} + \langle 6\rangle \nu_{[\beta_1][\beta_2],[\beta_3],[\beta_4]} + \nu_{[\beta_1],[\beta_2],[\beta_3],[\beta_4]} \\
&= (\rho_4 - 4\rho_3 - 2) E_F\Big[\prod_{i=1}^{4} g_{\beta_i}(\mathbf{X}, \boldsymbol{\theta}_0)\Big], &(10.5)
\end{aligned}$$

where $\langle n\rangle$ before a term with indices is a sum of n similar terms obtained by index permutation, e.g., $\langle 3\rangle \nu_{[\beta_1][\beta_2],[\beta_3]} = \nu_{[\beta_1][\beta_2],[\beta_3]} + \nu_{[\beta_1][\beta_3],[\beta_2]} + \nu_{[\beta_2][\beta_3],[\beta_1]}$. Within a class of the ECRs, we see that

$$(\rho_{\mathrm{CR}}'''(0), \rho_{\mathrm{CR}}''''(0)) = (-2, -6) \quad \text{iff}\, \omega = -1,$$

hence, the EL has a unique characteristic that, in addition to (10.3), both (10.4) and (10.5) are equal to zero, i.e., the partial Bartlett identities (with respect to $\boldsymbol{\lambda}$) hold up to the fourth degree.

Remark 10.2 (i) In the above-mentioned setup, recall $[\nu_{\alpha\alpha'}]_{\alpha,\alpha'\in\{1,\ldots,p\}} = \mathbf{O}_{p,p}$. We partition the $(p+M)\times(p+M)$ symmetric matrix $[\nu_{jj'}]_{j,j'\in\{1,\ldots,p+M\}} = \boldsymbol{\nu}$ (say) into

$$\boldsymbol{\nu} = \begin{pmatrix} \mathbf{O}_{p,p} & \boldsymbol{\nu}_{\theta\lambda} \\ \boldsymbol{\nu}_{\lambda\theta} & \boldsymbol{\nu}_{\lambda\lambda} \end{pmatrix} = \begin{pmatrix} \mathbf{O}_{p_1,p_1} & \mathbf{O}_{p_1,p_2} & \boldsymbol{\nu}_{\theta_{(1)}\lambda} \\ \mathbf{O}_{p_2,p_1} & \mathbf{O}_{p_2,p_2} & \boldsymbol{\nu}_{\theta_{(2)}\lambda} \\ \boldsymbol{\nu}_{\lambda\theta_{(1)}} & \boldsymbol{\nu}_{\lambda\theta_{(2)}} & \boldsymbol{\nu}_{\lambda\lambda} \end{pmatrix}$$

according to the partition $\boldsymbol{\eta} = (\boldsymbol{\theta}^\top, \boldsymbol{\lambda}^\top)^\top = (\boldsymbol{\theta}_{(1)}^\top, \boldsymbol{\theta}_{(2)}^\top, \boldsymbol{\lambda}^\top)^\top$, respectively, where $\boldsymbol{\nu}_{\theta\lambda}^\top = \boldsymbol{\nu}_{\lambda\theta} = (\boldsymbol{\nu}_{\lambda\theta_{(1)}}\ \boldsymbol{\nu}_{\lambda\theta_{(2)}})$.

(ii) Let

$$\boldsymbol{\nu}_{(21)}^\top = \boldsymbol{\nu}_{(12)} = (\mathbf{O}_{p_1,p_2}\ \boldsymbol{\nu}_{\theta_{(1)}\lambda})\,, \quad \boldsymbol{\nu}_{(22)} = \begin{pmatrix} \mathbf{O}_{p_2,p_2} & \boldsymbol{\nu}_{\theta_{(2)}\lambda} \\ \boldsymbol{\nu}_{\lambda\theta_{(2)}} & \boldsymbol{\nu}_{\lambda\lambda} \end{pmatrix}.$$

If $\boldsymbol{\nu}_{\lambda\lambda}$ is negative definite and $\boldsymbol{\nu}_{\lambda\theta}$ has full column rank, it is easily verified that the $(p+M)\times(p+M)$ matrix $\boldsymbol{\nu}$ (the $(p_2+M)\times(p_2+M)$ matrix $\boldsymbol{\nu}_{(22)}$) has just p (p_2) positive eigenvalues and M negative eigenvalues (both $\boldsymbol{\nu}$ and $\boldsymbol{\nu}_{(22)}$ are nonsingular), using Sylvester's law of inertia (see Horn and Johnson [38, p. 223]), since, e.g.,

$$\begin{pmatrix} \mathbf{I}_p & -\boldsymbol{\nu}_{\theta\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1} \\ \mathbf{O}_{M,p} & \mathbf{I}_M \end{pmatrix} \begin{pmatrix} \mathbf{O}_{p,p} & \boldsymbol{\nu}_{\theta\lambda} \\ \boldsymbol{\nu}_{\lambda\theta} & \boldsymbol{\nu}_{\lambda\lambda} \end{pmatrix} \begin{pmatrix} \mathbf{I}_p & \mathbf{O}_{p,M} \\ -\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta} & \mathbf{I}_M \end{pmatrix} = \begin{pmatrix} -\boldsymbol{\nu}_{\theta\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta} & \mathbf{O}_{p,M} \\ \mathbf{O}_{M,p} & \boldsymbol{\nu}_{\lambda\lambda} \end{pmatrix} .$$

It follows that

$$\begin{aligned} ||\boldsymbol{\nu}^{-1}|| &= \frac{1}{\min\{\rho_{M+1}^{\uparrow}(\boldsymbol{\nu}), -\rho_M^{\uparrow}(\boldsymbol{\nu})\}} , \\ ||\boldsymbol{\nu}_{(22)}^{-1}|| &= \frac{1}{\min\{\rho_{M+1}^{\uparrow}(\boldsymbol{\nu}_{(22)}), -\rho_M^{\uparrow}(\boldsymbol{\nu}_{(22)})\}} . \end{aligned} \tag{10.6}$$

Before formulating our results of Sect. 10.3, we prepare some additional notations. According to the partition $\boldsymbol{\eta} = (\boldsymbol{\theta}_{(1)}^{\top}, \boldsymbol{\eta}_{(2)}^{\top})^{\top}$, we partition the $(p+M) \times 1$ vector $\mathbf{Z}^{(N)\rho}(\boldsymbol{\eta}) = [Z_j^{(N)\rho}(\boldsymbol{\eta})]_{j\in\{1,\dots,p+M\}}$ and the $(p+M) \times (p+M)$ symmetric matrix $\mathbf{L}^{(N)\rho}(\boldsymbol{\eta}) = [\ell_{jj'}^{(N)\rho}(\boldsymbol{\eta})]_{j,j'\in\{1,\dots,p+M\}}$ into

$$\begin{aligned} \mathbf{Z}^{(N)\rho}(\boldsymbol{\eta}) &= \begin{pmatrix} \mathbf{Z}_{(1)}^{(N)\rho}(\boldsymbol{\eta}) \\ \mathbf{Z}_{(2)}^{(N)\rho}(\boldsymbol{\eta}) \end{pmatrix} , \\ \mathbf{L}^{(N)\rho}(\boldsymbol{\eta}) &= \begin{pmatrix} \mathbf{L}_{(11)}^{(N)\rho}(\boldsymbol{\eta}) & \mathbf{L}_{(12)}^{(N)\rho}(\boldsymbol{\eta}) \\ \mathbf{L}_{(21)}^{(N)\rho}(\boldsymbol{\eta}) & \mathbf{L}_{(22)}^{(N)\rho}(\boldsymbol{\eta}) \end{pmatrix} , \end{aligned}$$

respectively, where

$$Z_j^{(N)\rho}(\boldsymbol{\eta}) = N^{1/2}\ell_j^{(N)\rho}(\boldsymbol{\eta}) .$$

Let

$$\begin{aligned} \mathcal{G}^{(N)\rho}(\boldsymbol{\eta}) &= [\mathcal{G}_{ja}^{(N)\rho}(\boldsymbol{\eta})]_{j\in\{1,\dots,p+M\},a\in\{1,\dots,p_1\}} \\ &= \begin{pmatrix} \mathbf{I}_{p_1} \\ -\{\mathbf{L}_{(22)}^{(N)\rho}(\boldsymbol{\eta})\}^{-1}\mathbf{L}_{(21)}^{(N)\rho}(\boldsymbol{\eta}) \end{pmatrix} , \\ \mathbf{L}_{(11\cdot 2)}^{(N)\rho}(\boldsymbol{\eta}) &= [\ell_{(11\cdot 2)aa'}^{(N)\rho}(\boldsymbol{\eta})]_{a,a'\in\{1,\dots,p_1\}} \\ &= \{\mathcal{G}^{(N)\rho}(\boldsymbol{\eta})\}^{\top}\mathbf{L}^{(N)\rho}(\boldsymbol{\eta})\mathcal{G}^{(N)\rho}(\boldsymbol{\eta}) \\ &= \mathbf{L}_{(11)}^{(N)\rho}(\boldsymbol{\eta}) - \mathbf{L}_{(12)}^{(N)\rho}(\boldsymbol{\eta})\{\mathbf{L}_{(22)}^{(N)\rho}(\boldsymbol{\eta})\}^{-1}\mathbf{L}_{(21)}^{(N)\rho}(\boldsymbol{\eta}) , \end{aligned}$$

which are sample analogues of

$$\begin{aligned} \mathcal{G} = [\mathcal{G}_{ja}]_{j\in\{1,\dots,p+M\},a\in\{1,\dots,p_1\}} &= \begin{pmatrix} \mathbf{I}_{p_1} \\ -\boldsymbol{\nu}_{(22)}^{-1}\boldsymbol{\nu}_{(21)} \end{pmatrix} , \\ \boldsymbol{\nu}_{(11\cdot 2)} = -\boldsymbol{\nu}_{(12)}\boldsymbol{\nu}_{(22)}^{-1}\boldsymbol{\nu}_{(21)} &\ (= \mathcal{G}^{\top}\boldsymbol{\nu}\,\mathcal{G}) , \end{aligned}$$

respectively. We denote by $[\ell^{(N)\rho}(\boldsymbol{\eta})]^{jj'}$ the (j, j')th element of $\{\mathbf{L}^{(N)\rho}(\boldsymbol{\eta})\}^{-1}$ and by $[\ell^{(N)\rho}_{(11\cdot2)}(\boldsymbol{\eta})]^{aa'}$, $[\nu_{(11\cdot2)}]^{aa'}$, and $\nu_{(11\cdot2)aa'}$, the (a, a')th element of $\{\mathbf{L}^{(N)\rho}_{(11\cdot2)}(\boldsymbol{\eta})\}^{-1}$, $\boldsymbol{\nu}^{-1}_{(11\cdot2)}$, and $\boldsymbol{\nu}_{(11\cdot2)}$. Also, $[\mathbf{v}]_j$ or v_j stands for the jth element of any vector $\mathbf{v}$.

Lastly, for any scalar random variable $h_N(\mathbf{X}_1, \ldots, \mathbf{X}_N) = Y^{(N)}$ (say), we use the notation

$$Y^{(N)} = o_F^{(N)}(q_1, q_2)\,, \tag{10.7}$$

if $P_F^{(N)}[|Y^{(N)}| \geq \gamma(\log N)^{q_2}] = o(N^{-q_1})$ for some constants $\gamma > 0$ and $q_1, q_2 \geq 0$, independent of N; see, e.g., Magdalinos [54].[5] This notation can be also used for any finite-dimensional random vector or matrix, if it is defined via the sum of the absolute values of all elements.

10.2.3 Assumptions

The following assumptions are made for some $Q \geq 16$:

(C$_0$) $\rho(\cdot)$ is four times continuously differentiable and concave on its domain $\mathcal{V}_\rho$; an open interval containing zero, such that $\rho(0) = 0$, $\rho_1(0) = \rho_2(0) = -1$, and $|\rho_4(v) - \rho_4(0)| \leq \mathrm{Lip}^{(4)}_\rho |v|$ in a neighborhood $\mathcal{N}_\rho(\subset \mathcal{V}_\rho)$ of $v = 0$ for some constant $\mathrm{Lip}^{(4)}_\rho \geq 0$, where $\rho_i(v) = (d^i/dv^i)\rho(v)$.

(C$_1$) (i) The parameter space $\boldsymbol{\Theta}$ is an open convex subset of $\mathbb{R}^p$.
(ii) For each $R \in \{1, 2, 3, 4\}$ and $\beta \in \{1, \ldots, M\}$, $g_\beta(\mathbf{x}, \boldsymbol{\theta})$ has the Rth partial derivative with respect to $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^\top$.

(C$_2$)$_Q$ There exists $B(\cdot) = B_\Theta(\cdot)$ with $E_F[B^Q(\mathbf{X})] < \infty$, such that

$$\sup_{\theta\in\Theta} \sum_{\beta=1}^{M} \left\{ |g_\beta(\mathbf{x}, \boldsymbol{\theta})| + \sum_{R=1}^{4} \sum_{\alpha_1,\ldots,\alpha_R=1}^{p} |g_{\beta/\alpha_1\cdots\alpha_R}(\mathbf{x}, \boldsymbol{\theta})| \right\} \leq B(\mathbf{x})$$

and

[5] The following properties hold:

- If $Y^{(N)} = o_F^{(N)}(q_1, q_2)$, then,

  (i) $Y^{(N)} = o_F^{(N)}(q_1', q_2')$ for any $q_1 \geq q_1'(\geq 0)$ and $q_2' \geq q_2$;

  (ii) $N^{-\tau} Y^{(N)} = o_F^{(N)}(q_1, q_2')$ for any $\tau > 0$ and $q_2' \geq 0$.
- If $Y_{\mathrm{I}}^{(N)} = o_F^{(N)}(q, q_2)$ and $Y_{\mathrm{II}}^{(N)} = o_F^{(N)}(q, q_2')$, then,

$$Y_{\mathrm{I}}^{(N)} + Y_{\mathrm{II}}^{(N)} = o_F^{(N)}(q, \max(q_2, q_2')) \quad \text{and} \quad Y_{\mathrm{I}}^{(N)} Y_{\mathrm{II}}^{(N)} = o_F^{(N)}(q, q_2 + q_2')\,.$$

$$\sum_{\beta=1}^{M}\Big\{|g_\beta(\mathbf{x},\boldsymbol{\theta})-g_\beta(\mathbf{x},\boldsymbol{\theta}_0)| \\ +\sum_{R=1}^{3}\sum_{\alpha_1,\ldots,\alpha_R=1}^{p}|g_{\beta/\alpha_1\cdots\alpha_R}(\mathbf{x},\boldsymbol{\theta})-g_{\beta/\alpha_1\cdots\alpha_R}(\mathbf{x},\boldsymbol{\theta}_0)|\Big\}\le B(\mathbf{x})||\boldsymbol{\theta}-\boldsymbol{\theta}_0||$$

for all $\boldsymbol{\theta}$; $||\boldsymbol{\theta}-\boldsymbol{\theta}_0||\le\epsilon$, where $\boldsymbol{\theta}_0\in\boldsymbol{\Theta}$ is the unique solution to the moment restrictions $E_F[\mathbf{g}(\mathbf{X},\boldsymbol{\theta})]=\mathbf{0}_M$ (we can choose a small $\epsilon=\epsilon_{\theta_0}>0$ such that $B_p(\boldsymbol{\theta}_0:2\epsilon)$ is contained in the parameter space $\boldsymbol{\Theta}$, where $B_n(\mathbf{c}:\varrho)$ stands for the open ball of center $\mathbf{c}\in\mathbb{R}^n$ and radius $\varrho>0$).

(C_3) (i) $\boldsymbol{\nu}_{\lambda\lambda}=(-E_F[g_\beta(\mathbf{X},\boldsymbol{\theta}_0)g_{\beta'}(\mathbf{X},\boldsymbol{\theta}_0)])_{\beta,\beta'\in\{1,\ldots,M\}}$ is negative definite.
(ii) $\boldsymbol{\nu}_{\lambda\theta}=(-E_F[g_{\beta/\alpha}(\mathbf{X},\boldsymbol{\theta}_0)])_{\beta\in\{1,\ldots,M\},\alpha\in\{1,\ldots,p\}}$ has full column rank.

In order to state the last condition (C_4), let $\mathbf{G}_{1:3}(\mathbf{x},\boldsymbol{\theta}_0)$ be a column vector that consists of the following elements (there are six patterns):

$$g_{\beta_1}(\mathbf{x},\boldsymbol{\theta}_0)\,,\quad \prod_{i=1}^{2}g_{\beta_i}(\mathbf{x},\boldsymbol{\theta}_0)\,,\quad \prod_{i=1}^{3}g_{\beta_i}(\mathbf{x},\boldsymbol{\theta}_0)\,,\quad g_{\beta_1/\alpha_1}(\mathbf{x},\boldsymbol{\theta}_0)\,,$$
$$g_{\beta_1/\alpha_1}(\mathbf{x},\boldsymbol{\theta}_0)g_{\beta_2}(\mathbf{x},\boldsymbol{\theta}_0)+g_{\beta_2/\alpha_1}(\mathbf{x},\boldsymbol{\theta}_0)g_{\beta_1}(\mathbf{x},\boldsymbol{\theta}_0)\,,\quad\text{and } g_{\beta_1/\alpha_1\alpha_2}(\mathbf{x},\boldsymbol{\theta}_0)\,,$$

with the indices $p\ge\alpha_1\ge\alpha_2\ge1$ and $M\ge\beta_1\ge\beta_2\ge\beta_3\ge1$ in descending order. Note that each partial derivative (up to three) of $\ell^\rho(\boldsymbol{\eta})$ evaluated at $\boldsymbol{\eta}=\boldsymbol{\eta}_0$, except for $\ell_{\alpha_1}=\ell_{\alpha_1\alpha_2}=\ell_{\alpha_1\alpha_2\alpha_3}=0$, is equal to any of the elements of $\mathbf{G}_{1:3}(\mathbf{x},\boldsymbol{\theta}_0)$, multiplied the scaling -1 or ρ_3; see (10.2). Under the assumed nonsingularity of the $M\times M$ variance matrix $\mathrm{Var}_F[\mathbf{g}(\mathbf{X},\boldsymbol{\theta}_0)]$ (see (C_3)(i)), there exists a subvector $\dot{\mathbf{G}}_{1:3}(\mathbf{x},\boldsymbol{\theta}_0)$ of $\mathbf{G}_{1:3}(\mathbf{x},\boldsymbol{\theta}_0)$, containing at least the M elements $g_1(\mathbf{x},\boldsymbol{\theta}_0),\ldots,g_M(\mathbf{x},\boldsymbol{\theta}_0)$, such that $\mathrm{Var}_F[\dot{\mathbf{G}}_{1:3}(\mathbf{X},\boldsymbol{\theta}_0)]$ is nonsingular, and that $\mathbf{G}_{1:3}(\mathbf{x},\boldsymbol{\theta}_0)=\mathbf{A}(\boldsymbol{\theta}_0)\dot{\mathbf{G}}_{1:3}(\mathbf{x},\boldsymbol{\theta}_0)+\mathbf{b}(\boldsymbol{\theta}_0)$ for some block lower triangular matrix $\mathbf{A}(\boldsymbol{\theta}_0)$ and vector $\mathbf{b}(\boldsymbol{\theta}_0)$, not depending on $\mathbf{x}$. Then, Cramér's condition $\dot{\mathbf{G}}_{1:3}(\mathbf{X},\boldsymbol{\theta}_0)$ is additionally imposed:

(C_4) $\limsup_{||\mathbf{s}||\to\infty}|E_F[\exp\{i\mathbf{s}^\top\dot{\mathbf{G}}_{1:3}(\mathbf{X},\boldsymbol{\theta}_0)\}]|<1$.

10.3 Higher-Order Result of GEL Testing Inference

We want to test H: $\boldsymbol{\theta}_{(1)}=\boldsymbol{\theta}_{(1)0}$ against A: $\boldsymbol{\theta}_{(1)}\ne\boldsymbol{\theta}_{(1)0}$, for the GEL framework in the possibly over-identified moment restrictions. Lemma 10.1 motivates us to consider a class of test statistics admitting the stochastic expansion

$$
\begin{aligned}
T^{(N)\rho,\tau} &= (\widetilde{\mathbf{Z}}_{(1)}^{(N)\rho})^\top (\widetilde{\mathbf{L}}_{(11\cdot 2)}^{(N)\rho})^{-1}\widetilde{\mathbf{Z}}_{(1)}^{(N)\rho} + \frac{1}{N^{1/2}}\,\tau_1 \widetilde{\ell}_{j_1j_2j_3}^{(N)\rho} \prod_{i=1}^{3}[\widetilde{\mathcal{G}}^{(N)\rho}(\widetilde{\mathbf{L}}_{(11\cdot 2)}^{(N)\rho})^{-1}\widetilde{\mathbf{Z}}_{(1)}^{(N)\rho}]_{j_i} \\
&\quad + \frac{1}{N}\,[\widetilde{\ell}_{jj_1j_2}^{(N)\rho}\{\tau_2 \widetilde{\mathcal{G}}_{ja}^{(N)\rho}(\widetilde{\ell}_{(11\cdot 2)}^{(N)\rho})^{aa'}\widetilde{\mathcal{G}}_{j'a'}^{(N)\rho} + \tau_3(\widetilde{\ell}^{(N)\rho})^{jj'}\}\widetilde{\ell}_{j'j_3j_4}^{(N)\rho} + \tau_4 \widetilde{\ell}_{j_1j_2j_3j_4}^{(N)\rho}] \\
&\quad\quad \prod_{i=1}^{4}[\widetilde{\mathcal{G}}^{(N)\rho}(\widetilde{\mathbf{L}}_{(11\cdot 2)}^{(N)\rho})^{-1}\widetilde{\mathbf{Z}}_{(1)}^{(N)\rho}]_{j_i} \\
&\quad + \frac{1}{N^{3/2}}\, o_F^{(N)}(1, q)
\end{aligned}
\tag{10.8}
$$

for some constants $q \geq 0$ and $(\tau_1, \tau_2, \tau_3, \tau_4)^\top \in \mathbb{R}^4$, independent of N.

Lemma 10.1 *Suppose that assumptions* (C_0)–(C_3) *hold with* $Q \geq 16$*. In addition to the score-type test statistic*

$$
\mathrm{S}^{(N)\rho} = (\widetilde{\mathbf{Z}}_{(1)}^{(N)\rho})^\top (\widetilde{\mathbf{L}}_{(11\cdot 2)}^{(N)\rho})^{-1}\widetilde{\mathbf{Z}}_{(1)}^{(N)\rho}\ (\stackrel{d}{\longrightarrow} \chi^2_{p_1})\,,
$$

the following test statistics admit the stochastic expansion (10.8):

1. *the GELR test statistic*[6]

$$
\mathrm{ELR}^{(N)\rho} = 2N(\widetilde{\ell}^{(N)\rho} - \widehat{\ell}^{(N)\rho})\ (\stackrel{d}{\longrightarrow} \chi^2_{p_1})\,;
$$

2. *the Wald-type test statistic (and its modified version)*[7]

$$
\begin{aligned}
\mathrm{W}^{(N)\rho} &= N(\widehat{\boldsymbol{\theta}}_{(1)}^{(N)\rho} - \boldsymbol{\theta}_{(1)0})^\top \widetilde{\mathbf{L}}_{(11\cdot 2)}^{(N)\rho}(\widehat{\boldsymbol{\theta}}_{(1)}^{(N)\rho} - \boldsymbol{\theta}_{(1)0})\ (\stackrel{d}{\longrightarrow} \chi^2_{p_1})\,, \\
\mathrm{W}_{\dagger}^{(N)\rho} &= N(\widehat{\boldsymbol{\theta}}_{(1)}^{(N)\rho} - \boldsymbol{\theta}_{(1)0})^\top \widehat{\mathbf{L}}_{(11\cdot 2)}^{(N)\rho}(\widehat{\boldsymbol{\theta}}_{(1)}^{(N)\rho} - \boldsymbol{\theta}_{(1)0})\ (\stackrel{d}{\longrightarrow} \chi^2_{p_1})\,;
\end{aligned}
$$

3. *the modified score-type test statistic*[8]

$$
\mathrm{S}_{\dagger}^{(N)\rho} = (\widetilde{\mathbf{Z}}_{(1)}^{(N)\rho})^\top (\widehat{\mathbf{L}}_{(11\cdot 2)}^{(N)\rho})^{-1}\widetilde{\mathbf{Z}}_{(1)}^{(N)\rho}\ (\stackrel{d}{\longrightarrow} \chi^2_{p_1})\,;
$$

4. *the gradient-type test statistic*[9]

$$
\mathrm{grad}^{(N)\rho} = (-\widetilde{\mathbf{Z}}_{(1)}^{(N)\rho})^\top N^{1/2}(\widehat{\boldsymbol{\theta}}_{(1)}^{(N)\rho} - \boldsymbol{\theta}_{(1)0})\ (\stackrel{d}{\longrightarrow} \chi^2_{p_1})\,.
$$

---

[6] The GELR is a GEL counterpart of Wilks [76] from the parametric likelihood. As a special case of $\rho(\cdot) = \rho_{\mathrm{EL}}(\cdot)$, the ELR by Qin and Lawless [62] is most popular in the literature.

[7] The Wald-type is a GEL counterpart of Wald [75] from the parametric likelihood.

[8] The score-type (or Lagrange multiplier) is a GEL counterpart of [65] (see also [69]) from the parametric likelihood. For a history about Rao's score test, we refer the readers to [5]. See also [8] for the score-type test statistic from the estimating equation models.

[9] The gradient-type is a GEL counterpart of [73] from the parametric likelihood.

Table 10.2 Examples of the GEL-based test statistic $T^{(N)\rho,\tau}$

Test statistic	$(\tau_1, \tau_2, \tau_3, \tau_4)$
GELR; $\mathrm{ELR}^{(N)\rho}$	$(1/3, 0, 1/4, -1/12)$
Wald-type; $\mathrm{W}^{(N)\rho}$	$(1, 1/4, 1, -1/3)$
Modified Wald-type; $\mathrm{W}_{\dagger}^{(N)\rho}$	$(0, 5/4, 1/2, 1/6)$
Score-type; $\mathrm{S}^{(N)\rho}$	$(0, 0, 0, 0)$
Modified score-type; $\mathrm{S}_{\dagger}^{(N)\rho}$	$(1, 0, 3/2, -1/2)$
Gradient-type; $\mathrm{grad}^{(N)\rho}$	$(1/2, 0, 1/2, -1/6)$

See Table 10.2 for each parameter $(\tau_1, \tau_2, \tau_3, \tau_4)$.

The proof of this lemma is postponed to Sect. 10.6.2.

It is worth giving a brief description about the asymptotic distribution result. First, we know that $-\boldsymbol{\nu}_{\theta_{(2)}\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta_{(2)}}$ is positive definite (by (C_3)), hence, with

$$\mathbf{M}_{\lambda\lambda} = \boldsymbol{\nu}_{\lambda\lambda}^{-1} - \boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta_{(2)}}(\boldsymbol{\nu}_{\theta_{(2)}\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta_{(2)}})^{-1}\boldsymbol{\nu}_{\theta_{(2)}\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1},$$

we notice that the $p_1 \times p_1$ matrix $-\boldsymbol{\nu}_{\theta_{(1)}\lambda}\mathbf{M}_{\lambda\lambda}\boldsymbol{\nu}_{\lambda\theta_{(1)}}$ is positive definite, since it is the Schur complement of $-\boldsymbol{\nu}_{\theta_{(2)}\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta_{(2)}}$ in the $p \times p$ positive definite matrix

$$-\boldsymbol{\nu}_{\theta\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta} = -\begin{pmatrix} \boldsymbol{\nu}_{\theta_{(1)}\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta_{(1)}} & \boldsymbol{\nu}_{\theta_{(1)}\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta_{(2)}} \\ \boldsymbol{\nu}_{\theta_{(2)}\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta_{(1)}} & \boldsymbol{\nu}_{\theta_{(2)}\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta_{(2)}} \end{pmatrix} \quad \text{(by (C}_3\text{) again).}$$

Next, it is easy to see that $\widetilde{\mathbf{Z}}_{(1)}^{(N)\rho} = \mathcal{G}^\top\mathbf{Z}^{(N)} + o_p(1)$, $\widetilde{\mathbf{L}}_{(11\cdot2)}^{(N)\rho} = \boldsymbol{\nu}_{(11\cdot2)} + o_p(1)$, and $\widehat{\mathbf{L}}_{(11\cdot2)}^{(N)\rho} = \boldsymbol{\nu}_{(11\cdot2)} + o_p(1)$, where

$$\mathbf{Z}^{(N)} = \begin{pmatrix} \mathbf{0}_{p_1} \\ \mathbf{0}_{p_2} \\ -N^{1/2}\overline{\mathbf{g}} \end{pmatrix} \quad \text{with } \overline{\mathbf{g}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{g}(\mathbf{X}_i, \boldsymbol{\theta}_0).$$

Below, we can verify that

$$\mathcal{G}^\top\mathrm{Var}_F^{(N)}[\mathbf{Z}^{(N)}]\,\mathcal{G} = \boldsymbol{\nu}_{(11\cdot2)} \tag{10.9}$$

is positive definite. Finally, $(\mathcal{G}^\top\mathbf{Z}^{(N)})^\top\boldsymbol{\nu}_{(11\cdot2)}^{-1}(\mathcal{G}^\top\mathbf{Z}^{(N)}) \xrightarrow{d} \chi_{p_1}^2$ is a consequence of the central limit theorem $\mathcal{G}^\top\mathbf{Z}^{(N)} \xrightarrow{d} \mathrm{N}(\mathbf{0}_{p_1}, \boldsymbol{\nu}_{(11\cdot2)})$.

Now, recall

$$\boldsymbol{\nu}_{(22)}^{-1} = \begin{pmatrix} -(\boldsymbol{\nu}_{\theta_{(2)}\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta_{(2)}})^{-1} & (\boldsymbol{\nu}_{\theta_{(2)}\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta_{(2)}})^{-1}\boldsymbol{\nu}_{\theta_{(2)}\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1} \\ \boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta_{(2)}}(\boldsymbol{\nu}_{\theta_{(2)}\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta_{(2)}})^{-1} & \mathbf{M}_{\lambda\lambda} \end{pmatrix}.$$

By definition, $\boldsymbol{\nu}_{(11\cdot2)} = -\boldsymbol{\nu}_{(12)}\boldsymbol{\nu}_{(22)}^{-1}\boldsymbol{\nu}_{(21)} = -\boldsymbol{\nu}_{\theta_{(1)}\lambda}\mathbf{M}_{\lambda\lambda}\boldsymbol{\nu}_{\lambda\theta_{(1)}}$ (this matrix is positive definite, as mentioned above). On the other hand, using $\mathbf{M}_{\lambda\lambda} = \mathbf{M}_{\lambda\lambda}\boldsymbol{\nu}_{\lambda\lambda}\mathbf{M}_{\lambda\lambda}$ and $\boldsymbol{\nu}_{\theta_{(2)}\lambda}\mathbf{M}_{\lambda\lambda} = \mathbf{O}_{p_2,M}$, we have

$$\boldsymbol{\nu}_{(22)}^{-1}\begin{pmatrix}\mathbf{O}_{p_2,p_2} & \mathbf{O}_{p_2,M}\\ \mathbf{O}_{M,p_2} & \boldsymbol{\nu}_{\lambda\lambda}\end{pmatrix}\boldsymbol{\nu}_{(22)}^{-1} = \begin{pmatrix}(\boldsymbol{\nu}_{\theta_{(2)}\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta_{(2)}})^{-1} & \mathbf{O}_{p_2,M}\\ \mathbf{O}_{M,p_2} & \mathbf{M}_{\lambda\lambda}\end{pmatrix}.$$

Then,

$$\begin{aligned}\mathcal{G}^\top \mathrm{Var}_F^{(N)}[\mathbf{Z}^{(N)}]\mathcal{G} &= \mathcal{G}^\top\left(\begin{array}{c|cc}\mathbf{O}_{p_1,p_1} & \mathbf{O}_{p_1,p_2} & \mathbf{O}_{p_1,M}\\ \hline \mathbf{O}_{p_2,p_1} & \mathbf{O}_{p_2,p_2} & \mathbf{O}_{p_2,M}\\ \mathbf{O}_{M,p_1} & \mathbf{O}_{M,p_2} & -\boldsymbol{\nu}_{\lambda\lambda}\end{array}\right)\mathcal{G}\\ &= -\boldsymbol{\nu}_{(12)}\boldsymbol{\nu}_{(22)}^{-1}\begin{pmatrix}\mathbf{O}_{p_2,p_2} & \mathbf{O}_{p_2,M}\\ \mathbf{O}_{M,p_2} & \boldsymbol{\nu}_{\lambda\lambda}\end{pmatrix}\boldsymbol{\nu}_{(22)}^{-1}\boldsymbol{\nu}_{(21)}\\ &= -\boldsymbol{\nu}_{(12)}\begin{pmatrix}(\boldsymbol{\nu}_{\theta_{(2)}\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta_{(2)}})^{-1} & \mathbf{O}_{p_2,M}\\ \mathbf{O}_{M,p_2} & \mathbf{M}_{\lambda\lambda}\end{pmatrix}\boldsymbol{\nu}_{(21)}\\ &= -\boldsymbol{\nu}_{\theta_{(1)}\lambda}\mathbf{M}_{\lambda\lambda}\boldsymbol{\nu}_{\lambda\theta_{(1)}} = \boldsymbol{\nu}_{(11\cdot2)}\,.\end{aligned}$$

This verifies (10.9).[10]

We are ready to state the higher-order result on the GEL testing inference. The finding is a nontrivial extension of Baggerly [2] to the GEL framework in the possibly over-identified moment restrictions.

Theorem 10.1 *Suppose that assumptions (C_0)–(C_4) hold with $Q \geq 16$. Then, (10.8) is B-correctable if* $(\rho_3, \rho_4, \tau_1, \tau_2, \tau_3, \tau_4) = (-2, -6, 1/3, 0, 1/4, -1/12)$, *i.e., the*

[10] Newey and Smith [56] established that

$$N^{1/2}(\widehat{\boldsymbol{\eta}}^{(N)\rho} - \boldsymbol{\eta}_0) \xrightarrow{d} \mathrm{N}\left(\mathbf{0}_{p+M}\,,\ \begin{pmatrix}(\boldsymbol{\nu}_{\theta\lambda}\boldsymbol{\nu}_{\lambda,\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta})^{-1} & \mathbf{O}_{p,M}\\ \mathbf{O}_{M,p} & \mathbf{P}\end{pmatrix}\right),$$

where $\boldsymbol{\nu}_{\lambda,\lambda} = -\boldsymbol{\nu}_{\lambda\lambda}$ and

$$\mathbf{P} = \boldsymbol{\nu}_{\lambda,\lambda}^{-1} - \boldsymbol{\nu}_{\lambda,\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta}(\boldsymbol{\nu}_{\theta\lambda}\boldsymbol{\nu}_{\lambda,\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta})^{-1}\boldsymbol{\nu}_{\theta\lambda}\boldsymbol{\nu}_{\lambda,\lambda}^{-1}\,.$$

We can see that

$$(\boldsymbol{\nu}_{\theta\lambda}\boldsymbol{\nu}_{\lambda,\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta})^{-1} = -(\boldsymbol{\nu}_{\theta\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta})^{-1} = -\begin{pmatrix}\boldsymbol{\nu}_{\theta_{(1)}\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta_{(1)}} & \boldsymbol{\nu}_{\theta_{(1)}\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta_{(2)}}\\ \boldsymbol{\nu}_{\theta_{(2)}\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta_{(1)}} & \boldsymbol{\nu}_{\theta_{(2)}\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta_{(2)}}\end{pmatrix}^{-1},$$

which yields $N^{1/2}(\widehat{\boldsymbol{\theta}}_{(1)}^{(N)\rho} - \boldsymbol{\theta}_{(1)0}) \xrightarrow{d} \mathrm{N}(\mathbf{0}_{p_1}, \boldsymbol{\nu}_{(11\cdot2)}^{-1})$, since

$$(\mathbf{I}_{p_1}\ \mathbf{O}_{p_1,p_2})(\boldsymbol{\nu}_{\theta\lambda}\boldsymbol{\nu}_{\lambda\lambda}^{-1}\boldsymbol{\nu}_{\lambda\theta})^{-1}\begin{pmatrix}\mathbf{I}_{p_1}\\ \mathbf{O}_{p_2,p_1}\end{pmatrix} = -(\boldsymbol{\nu}_{\theta_{(1)}\lambda}\mathbf{M}_{\lambda\lambda}\boldsymbol{\nu}_{\lambda\theta_{(1)}})^{-1} = \boldsymbol{\nu}_{(11\cdot2)}^{-1}\,.$$

It follows that $\mathrm{W}^{(N)\rho} \xrightarrow{d} \chi^2_{p_1}$ and $\mathrm{W}_\dagger^{(N)\rho} \xrightarrow{d} \chi^2_{p_1}$.

ELR test statistic $\mathrm{ELR}^{(N)} = 2N\{\ell_{\mathrm{EL}}^{(N)}(\widehat{\boldsymbol{\theta}}_{\mathrm{EL}}^{(N)}) - \ell_{\mathrm{EL}}^{(N)}(\widetilde{\boldsymbol{\theta}}_{\mathrm{EL}}^{(N)})\}$*, based on* $\rho(\cdot) = \rho_{\mathrm{EL}}(\cdot)$*, is B-correctable.*

Proof Proposition 10.2 of Sect. 10.6.3 shows that $T^{(N)\rho,\tau}$ is B-correctable iff

$$0 = \Big(\frac{1}{4}\kappa_{b_1,b_2,b_3,b_4}^{\rho_3,\rho_4,\tau_1,\tau_2,\tau_3,\tau_4} + \kappa_{b_1}^{\rho_3,\tau_1}\kappa_{b_2,b_3,b_4}^{\rho_3,\tau_1}\Big)\nu_{(11\cdot 2)}^{b_1 b_2}\nu_{(11\cdot 2)}^{b_3 b_4},$$

$$0 = 6\kappa_{b_1,b_2,b_3}^{\rho_3,\tau_1}\Big(\prod_{i=1}^{3}\nu_{(11\cdot 2)}^{b_i b_i'}\Big)\kappa_{b_1',b_2',b_3'}^{\rho_3,\tau_1} + 9(\nu_{(11\cdot 2)}^{b_1 b_2}\kappa_{b_1,b_2,b}^{\rho_3,\tau_1})\nu_{(11\cdot 2)}^{bb'}(\nu_{(11\cdot 2)}^{b_1' b_2'}\kappa_{b_1',b_2',b'}^{\rho_3,\tau_1}),$$

equivalently,

$$\begin{aligned} 0 &= \kappa_{a_1,a_2,a_3}^{\rho_3,\tau_1} \\ &= (3\tau_1 - 1)\nu^{\rho\mathcal{G}\mathcal{G}\mathcal{G}}_{a_1 a_2 a_3} + (\rho_3 + 2)\sum_{\beta_1,\beta_2,\beta_3=1}^{M}\Big(\prod_{i=1}^{3}\mathcal{G}_{[\beta_i]a_i}\Big)E_F\Big[\prod_{i=1}^{3}g_{\beta_i}(\mathbf{X},\boldsymbol{\theta}_0)\Big] \end{aligned} \tag{10.10}$$

for $a_1, a_2, a_3 \in \{1,\ldots,p_1\}$ and $0 = \kappa_{b_1,b_2,b_3,b_4}^{\rho_3,\rho_4,\tau_1,\tau_2,\tau_3,\tau_4}\nu_{(11\cdot 2)}^{b_1 b_2}\nu_{(11\cdot 2)}^{b_3 b_4}$ (for simplicity, the notation $\#^{\mathcal{G}}_{\ldots a\ldots}$ is used instead of writing $\#_{\ldots j\ldots}\mathcal{G}_{ja}$). Obviously, $(\rho_3, \tau_1) = (-2, 1/3)$ is a sufficient condition for (10.10). The assertion follows by noting that

$$\begin{aligned} \kappa_{a_1,a_2,a_3,a_4}^{-2,\rho_4,1/3,\tau_2,\tau_3,\tau_4} &= 4\tau_2\langle 3\rangle\nu^{\rho\mathcal{G}\mathcal{G}\mathcal{G}}_{a\,a_1 a_2}\nu_{(11\cdot 2)}^{aa'}\nu^{\rho\mathcal{G}\mathcal{G}\mathcal{G}}_{a' a_3 a_4} \\ &\quad + (4\tau_3 - 1)\langle 3\rangle\nu^{\rho\mathcal{G}\mathcal{G}}_{a_1 a_2 j}\nu^{jj'}\nu^{\rho\mathcal{G}\mathcal{G}}_{a_3 a_4 j'} + (12\tau_4 + 1)\nu^{\rho\mathcal{G}\mathcal{G}\mathcal{G}\mathcal{G}}_{a_1 a_2 a_3 a_4} \\ &\quad + (\rho_4 + 6)\sum_{\beta_1,\beta_2,\beta_3,\beta_4=1}^{M}\Big(\prod_{i=1}^{4}\mathcal{G}_{[\beta_i]a_i}\Big)E_F\Big[\prod_{i=1}^{4}g_{\beta_i}(\mathbf{X},\boldsymbol{\theta}_0)\Big] \end{aligned}$$

for $a_1, a_2, a_3, a_4 \in \{1,\ldots,p_1\}$. □

To implement the Bartlett correction

$$\Big(1 + \frac{b}{N}\Big)\mathrm{ELR}^{(N)} \approx \frac{\mathrm{ELR}^{(N)}}{E_F^{(N)}[\mathrm{ELR}^{(N)}/p_1]},$$

the Bartlett factor $b = -(1/p_1)(\kappa_{b_1,b_2}^{-2,-6,1/3,0,1/4,-1/12} + \kappa_{b_1}^{-2,1/3}\kappa_{b_2}^{-2,1/3})\nu_{(11\cdot 2)}^{b_1 b_2}$ has to be given explicitly.[11] If an analytical expression of b were available, a sample analogue $b^{(N)}$ of b could be obtained according to routine that the population moments are replaced by the corresponding sample moments. However, due to the rather lengthy form of $\kappa_{b_1,b_2}^{-2,-6,1/3,0,1/4,-1/12}$, we need the following bootstrap procedure:

[11] By Proposition 10.2 of Sect. 10.6.3, we have, in principle,

$$E_F^{(N)}[\mathrm{ELR}^{(N)}] = p_1 + \frac{1}{N}(\kappa_{b_1,b_2}^{-2,-6,1/3,0,1/4,-1/12} + \kappa_{b_1}^{-2,1/3}\kappa_{b_2}^{-2,1/3})\nu_{(11\cdot 2)}^{b_1 b_2} + o(N^{-1}).$$

Step 1. Resample with replacement from a given observed sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ to obtain a bootstrap sample $\{\mathbf{X}_1^*, \ldots, \mathbf{X}_N^*\}$. Then, compute

$$\mathrm{ELR}^{*(N)} = 2N \left\{ \ell_{\mathrm{EL}}^{*(N)}(\widehat{\boldsymbol{\eta}}_{\mathrm{EL}}^{*(N)}) - \ell_{\mathrm{EL}}^{*(N)} \begin{pmatrix} \widehat{\boldsymbol{\theta}}_{\mathrm{EL}(1)}^{(N)} \\ \widetilde{\boldsymbol{\eta}}_{\mathrm{EL}(2)}^{*(N)} \end{pmatrix} \right\},$$

where $\ell_{\mathrm{EL}}^{*(N)}(\boldsymbol{\eta}) = N^{-1} \sum_{i=1}^{N} \log\{1 - \boldsymbol{\lambda}^\top \mathbf{g}(\mathbf{X}_i^*, \boldsymbol{\theta})\}$.

Step 2. Repeat Step 1 (B times) to get $\mathrm{ELR}_1^{*(N)}, \ldots, \mathrm{ELR}_B^{*(N)}$. Then, a bootstrap estimate of b is given by

$$b^* = \frac{1}{B} \sum_{i=1}^{B} N\Big(1 - \frac{\mathrm{ELR}_i^{*(N)}}{p_1}\Big).$$

Given a significance level α, the resulting bootstrap-based Bartlett corrected test is to reject the null hypothesis if

$$\Big(1 + \frac{b^*}{N}\Big)\mathrm{ELR}^{(N)} > \chi_{p_1}^2(\alpha),$$

where $\chi_\nu^2(\alpha)$ is the upper α-quantile of the χ_ν^2 random variable.

Remark 10.3 This kind of a hybrid procedure was first suggested by Rocke [67] for Zellner's seemingly unrelated regression models. Since then, making use of Efron's bootstrap method is a routine to estimate the Bartlett factor in the literature.[12] One may prefer a direct use of "bootstrap test", i.e., an alternative method is to estimate the critical value of $\mathrm{ELR}^{(N)}$ as the upper α-quantile of $\{\mathrm{ELR}_1^{*(N)}, \ldots, \mathrm{ELR}_B^{*(N)}\}$. Rocke [67] pointed out that estimating (tail) quantile requires, in general, a larger number (B) of the bootstrap iterations than estimating the constant b; this is an advantage of the bootstrap-based Bartlett correction.

We illustrate how the bootstrap-based Bartlett correction works. As a benchmark analysis, we consider here the Bartlett [4] test of homogeneity of variances among two or more groups. This test is available in most of the statistical software (e.g., the R software has a function "bartlett.test"), which enables us to test whether the k samples from the normal populations have equal variances. More precisely, the LR test statistic after the Bartlett correction is given by

$$T_{\mathrm{Bart}} = \frac{T}{1 + \Big(\sum_{i=1}^{k} \frac{1}{N_i - 1} + \frac{1}{N-k}\Big)\frac{1}{3(k-1)}},$$

[12] Some authors may use the terminology of the Bartlett correction in the sense that the expectation of the test statistic is closer to that of the original (uncorrected) test statistic.

Table 10.3 The type I errors (10,000 replications) of homogeneity tests of variances among three groups ($k = 3$), using the original or (analytical or bootstrap-based) Bartlett corrected LR test statistics, denoted by T, T_{Bart}, and T^*_{Bart} (with significance levels $\alpha \in \{0.1, 0.05, 0.01\}$), for the balanced case $(N_1, N_2, N_3) \in \{(7, 7, 7), (15, 15, 15), (20, 20, 20)\}$, i.e., $N \in \{21, 45, 60\}$

		LR	Bartlett corrected LR			
			Analytical	Bootstrap-based		
				$B = 100$	200	500
$N = 21$	$\alpha = 0.1$	0.1200	0.1018	0.1004	0.0999	0.1009
	$\alpha = 0.05$	0.0627	0.0502	0.0522	0.0505	0.0496
	$\alpha = 0.01$	0.0141	0.0099	0.0112	0.0112	0.0098
$N = 45$	$\alpha = 0.1$	0.1064	0.1008	0.0985	0.1001	0.1002
	$\alpha = 0.05$	0.0559	0.0510	0.0510	0.0509	0.0506
	$\alpha = 0.01$	0.0118	0.0109	0.0110	0.0100	0.0101
$N = 60$	$\alpha = 0.1$	0.1025	0.0975	0.0994	0.0966	0.0981
	$\alpha = 0.05$	0.0533	0.0505	0.0510	0.0500	0.0505
	$\alpha = 0.01$	0.0111	0.0098	0.0114	0.0109	0.0104

where

$$T = (N - k) \log s^2_{\text{pool}} - \sum_{i=1}^{k} (N_i - 1) \log s_i^2 \ (\xrightarrow{d} \chi^2_{k-1})$$

is the LR test statistic for testing the equality of variances in k normal populations $N(\mu_i, \sigma_i^2)$'s, $N = \sum_{i=1}^k N_i$ is the total sample size, s_i^2 is the unbiased variance of the sample of size N_i from the ith group, and $s^2_{\text{pool}} = \sum_{i=1}^k (N_i - 1)s_i^2/(N - k)$ is the pooled variance. Table 10.3 shows the simulation results (from 10,000 replications) of the type I errors of the large sample LR test (T), the (analytical) Bartlett corrected LR test (T_{Bart}), and the bootstrap-based Bartlett corrected LR tests (T^*_{Bart}), for the balanced case $(N_1, N_2, N_3) \in \{(7, 7, 7), (15, 15, 15), (20, 20, 20)\}$ when $k = 3$ (we set $B \in \{100, 200, 500\}$). We observe that

- the large sample test, with the rejection region $T > \chi^2_k(\alpha)$, is oversized;
- the type I error of the test using the analytical Bartlett correction T_{Bart} is close to the nominal size, and the bootstrap-based Bartlett correction T^*_{Bart} works even for the small bootstrap iterations.

As we argued in Theorem 10.1, several GEL-based test statistics, except for the ELR test statistic, are, in general, not B-correctable. In the next section, we will reduce the error of the $\chi^2_{p_1}$ approximation to $o(N^{-1})$, by means of the Bartlett-type correction with bootstrap.

10.4 Bartlett-Type Correction with Bootstrap

Suppose that we have the asymptotic expansion in the form of[13]

$$P^{(N)}[S^{(N)} \le x] = G_{p_1}(x) - \frac{2}{N}\sum_{\ell=1}^{3} \pi_{S,\ell}\, g_{p_1+2\ell}(x) + o(N^{-1}), \tag{10.11}$$

as in, e.g., Proposition 10.2 of Sect. 10.6.3, where $G_\nu(x) = \int_0^x g_\nu(y)\,dy$ for $x \ge 0$, and $g_\nu(\cdot)$ is the density function of the χ^2_ν random variable. Our interest here is to refine the approximation of (10.11).

Using the relation $g_{p_1+2\ell}(x) = \frac{\Gamma(p_1/2)}{2^\ell \Gamma(p_1/2+\ell)}\, x^\ell g_{p_1}(x)$, we rewrite (10.11) as

$$\begin{aligned} &P^{(N)}[S^{(N)} \le x] \\ &= G_{p_1}(x) - \frac{2}{N}\left\{\frac{\pi_{S,1}}{p_1} + \frac{\pi_{S,2}\, x}{p_1(p_1+2)} + \frac{\pi_{S,3}\, x^2}{p_1(p_1+2)(p_1+4)}\right\} x g_{p_1}(x) + o(N^{-1}) \\ &= G_{p_1}\left[x - \frac{2}{N}\left\{\frac{\pi_{S,1}}{p_1} + \frac{\pi_{S,2}\, x}{p_1(p_1+2)} + \frac{\pi_{S,3}\, x^2}{p_1(p_1+2)(p_1+4)}\right\} x\right] + o(N^{-1}). \end{aligned}$$

Then, one suggests to multiply $S^{(N)}$ by a polynomial of degree 2 (not the constant factor), i.e., construct a new test statistic, referred to as the Bartlett-type corrected test statistic (Cordeiro and Ferrari [20][14]), in the form of

$$S_{\rm CF}^{(N)} = \left[1 - \frac{2}{N}\left\{\frac{\pi_{S,1}}{p_1} + \frac{\pi_{S,2} S^{(N)}}{p_1(p_1+2)} + \frac{\pi_{S,3} (S^{(N)})^2}{p_1(p_1+2)(p_1+4)}\right\}\right] S^{(N)} = \mathcal{P}_S^{(N)}(S^{(N)}) \quad \text{(say)}.$$

In some contexts, the monotonicity of the transformation is a matter of concern. The polynomial $\mathcal{P}_S^{(N)}(x)$ may be not monotone in $x \ge 0$, although this is presumably asymptotically monotone increasing in $x \in [0,\ O((\log N)^a)]$ for sufficiently large N, where $a > 0$. A simple idea to preserve the monotonicity is the perturbation adding a term of order $O(N^{-2})$, i.e.,

---

[13] We stress that the classical likelihood-based parametric case is also allowed here. See Kakizawa [45–49].

[14] There were, at least for me, confusing expressions in Cordeiro and Ferrari [20, (1) and (2)]; indeed, using the relation $2g_{\nu+2}(x) = G_\nu(x) - G_{\nu+2}(x)$, one can rearrange Harris's [34, (3.2)] asymptotic expansion for the distribution of the Rao test statistic $S_{\rm R}$, as follows:

$$\begin{aligned} &\Pr(S_{\rm R} \le x) \\ &= G_m(x) + \frac{1}{24n}[A_3 G_{m+6}(x) + (A_2 - 3A_3)G_{m+4}(x) + (A_1 - 2A_2 + 3A_3)G_{m+2}(x) \\ &\qquad + (-A_1 + A_2 - A_3)G_m(x)] + o(n^{-1}) \\ &= G_m(x) - \frac{1}{12n}\left[\frac{A_1 - A_2 + A_3}{m} + \frac{(A_2 - 2A_3)x}{m(m+2)} + \frac{A_3 x^2}{m(m+2)(m+4)}\right] x g_m(x) + o(n^{-1}). \end{aligned}$$

$$
\begin{aligned}
&\mathcal{P}_S^{(N)}(x) + \frac{1}{4}\int_0^x \left\{\frac{d}{dt}\mathcal{P}_S^{(N)}(t) - 1\right\}^2 dt \\
&= \Big[1 - \frac{2}{N}\Big\{\frac{\pi_{S,1}}{p_1} + \frac{\pi_{S,2}x}{p_1(p_1+2)} + \frac{\pi_{S,3}x^2}{p_1(p_1+2)(p_1+4)}\Big\} \\
&\quad + \frac{1}{N^2}\Big\{\frac{\pi_{S,1}^2}{p_1^2} + \frac{\pi_{S,2}^2 4x^2}{3p_1^2(p_1+2)^2} + \frac{\pi_{S,3}^2 9x^4}{5p_1^2(p_1+2)^2(p_1+4)^2} \\
&\qquad + 2\frac{\pi_{S,1}\pi_{S,2}x}{p_1^2(p_1+2)} + 2\frac{\pi_{S,1}\pi_{S,3}x^2}{p_1^2(p_1+2)(p_1+4)} + \frac{\pi_{S,2}\pi_{S,3}3x^3}{p_1^2(p_1+2)^2(p_1+4)}\Big\}\Big]x \\
&= \mathcal{P}_S^{+(N)}(x) \quad \text{(say)}
\end{aligned}
$$

is monotone increasing in $x \geq 0$. The monotone Bartlett-type corrected test statistic $\mathcal{P}_S^{+(N)}(S^{(N)})$ was suggested by Kakizawa [42], although there were infinitely many forms; Fujikoshi [28], Fujisawa [29], Iwashita [39], and Kakizawa [43].

Remark 10.4 The Bartlett-type corrected test statistic $S_{\mathrm{CF}}^{(N)}$ is closely related to Cornish–Fisher's type expansion for the upper α-quantile of $S^{(N)}$, denoted by $S^{(N)}(\alpha)$, as follows:

$$
S^{(N)}(\alpha) = \Big[1 + \frac{2}{N}\Big\{\frac{\pi_{S,1}}{p_1} + \frac{\pi_{S,2}\chi_{p_1}^2(\alpha)}{p_1(p_1+2)} + \frac{\pi_{S,3}\{\chi_{p_1}^2(\alpha)\}^2}{p_1(p_1+2)(p_1+4)}\Big\}\Big]\chi_{p_1}^2(\alpha) + o(N^{-1}).
$$

Of course, for implementing the Bartlett-type correction $S_{\mathrm{CF}}^{(N)}$, three coefficients $\pi_{S,1}, \pi_{S,2}$, and $\pi_{S,3}$ have to be given explicitly. But, the closed-form expressions are, in general, difficult to derive. Kojima and Kubokawa [51] applied the parametric bootstrap for the hypothesis testing of regression coefficients in normal linear models with general error covariance matrices (note that $\pi_{S,3} = 0$ due to the normal models), including normal linear mixed models. Our proposal, which is applicable without assuming (i) $\pi_{S,3} = 0$ or (ii) any structure of the parametric models (hence, the GEL framework is allowed), is naturally introduced from a different perspective, i.e., it can be shown that, in principle, these coefficients $\pi_{S,1}, \pi_{S,2}$, and $\pi_{S,3}$ are associated with the first three moments of $S^{(N)}$;

$$
\begin{aligned}
&\begin{pmatrix} E^{(N)}[S^{(N)}] - p_1 \\ E^{(N)}[(S^{(N)})^2] - p_1(p_1+2) \\ E^{(N)}[(S^{(N)})^3] - p_1(p_1+2)(p_1+4) \end{pmatrix} \\
&= \frac{2}{N}\begin{pmatrix} 1 & 1 & 1 \\ 2(p_1+2) & 2(p_1+4) & 2(p_1+6) \\ 3(p_1+2)(p_1+4) & 3(p_1+4)(p_1+6) & 3(p_1+6)(p_1+8) \end{pmatrix}\begin{pmatrix} \pi_{S,1} \\ \pi_{S,2} \\ \pi_{S,3} \end{pmatrix} + o(N^{-1}).
\end{aligned}
$$

This fact is compartible with the bootstrap procedure, as follows: Let $\mu_1' = p_1$, $\mu_2' = p_1(p_1+2)$, and $\mu_3' = p_1(p_1+2)(p_1+4)$. Note that the ℓth raw moment of the $\chi_{p_1}^2$ random variable is given by $\mu_\ell' = 2^\ell \frac{\Gamma(p_1/2+\ell)}{\Gamma(p_1/2)}$.

Step 1. Resample with replacement from a given observed sample $\{\mathbf{X}_1, \ldots, \mathbf{X}_N\}$ to obtain a bootstrap sample $\{\mathbf{X}_1^*, \ldots, \mathbf{X}_N^*\}$. Then, compute a bootstrap analogue $S^{*(N)}$, corresponding to $S^{(N)}$.

Step 2. Repeat Step 1 (B times) to get $S_1^{*(N)}, \ldots, S_B^{*(N)}$, and compute the biases of the first three moments;

$$\delta_{S,\ell}^* = \frac{1}{B}\sum_{i=1}^{B}(S_i^{*(N)})^\ell - \mu'_\ell\,, \quad \ell \in \{1, 2, 3\}\,.$$

Then, the three coefficients $\pi_{S,1}$, $\pi_{S,2}$, and $\pi_{S,3}$ are estimated by

$$\begin{pmatrix} \pi_{S,1}^* \\ \pi_{S,2}^* \\ \pi_{S,3}^* \end{pmatrix} = \frac{N}{2}\begin{pmatrix} \frac{(p_1+4)(p_1+6)}{8} & -\frac{p_1+6}{8} & \frac{1}{24} \\ -\frac{(p_1+2)(p_1+6)}{4} & \frac{p_1+5}{4} & -\frac{1}{12} \\ \frac{(p_1+2)(p_1+4)}{8} & -\frac{p_1+4}{8} & \frac{1}{24} \end{pmatrix}\begin{pmatrix} \delta_{S,1}^* \\ \delta_{S,2}^* \\ \delta_{S,3}^* \end{pmatrix}.$$

Given a significance level α, the resulting bootstrap-based Bartlett-type corrected test is to reject the null hypothesis if

$$\left[1 - \frac{2}{N}\left\{\frac{\pi_{S,1}^*}{p_1} + \frac{\pi_{S,2}^* S^{(N)}}{p_1(p_1+2)} + \frac{\pi_{S,3}^*(S^{(N)})^2}{p_1(p_1+2)(p_1+4)}\right\}\right]S^{(N)} > \chi_{p_1}^2(\alpha)\,. \quad (10.12)$$

Remark 10.5 (i) When $\pi_{S,2} = \pi_{S,3} = 0$, one can use, instead of the test (10.12), the bootstrap-based Bartlett corrected test, described in the previous section, i.e., reject the null hypothesis if

$$\left(1 - \frac{\delta_{S,1}^*}{p_1}\right)S^{(N)} > \chi_{p_1}^2(\alpha)\,.$$

On the other hand, for the case $\pi_{S,3} = 0$ (e.g., in the GEL framework, consider a situation where $E_F[\prod_{j=1}^3 g_{\beta_j}(\mathbf{X}; \boldsymbol{\theta}_0)] = 0$ for every $\beta_1, \beta_2, \beta_3 \in \{1, \ldots, M\}$ when $\mathbf{g}(\mathbf{X}; \boldsymbol{\theta})$ is linear with respect to $\boldsymbol{\theta}$), the resulting test, instead of the test (10.12), is to reject the null hypothesis if

$$\left[1 - \frac{2}{N}\left\{\frac{\dot{\pi}_{S,1}^*}{p_1} + \frac{\dot{\pi}_{S,2}^* S^{(N)}}{p_1(p_1+2)}\right\}\right]S^{(N)} > \chi_{p_1}^2(\alpha)\,,$$

where

$$\begin{pmatrix} \dot{\pi}_{S,1}^* \\ \dot{\pi}_{S,2}^* \end{pmatrix} = \frac{N}{2}\begin{pmatrix} \frac{p_1+4}{2} & -\frac{1}{4} \\ -\frac{p_1+2}{2} & \frac{1}{4} \end{pmatrix}\begin{pmatrix} \delta_{S,1}^* \\ \delta_{S,2}^* \end{pmatrix}.$$

(ii) However, the situation $\pi_{S,2} = \pi_{S,3} = 0$ or $\pi_{S,3} = 0$ would be rare, generally; the test (10.12) is finally recommended if an N^{-1}-accurate testing inference is needed.

To illustrate the bootstrap-based Bartlett-type corrected test (10.12), we conduct some simulation experiments from the exponential regression model (independent but non-identical case), as follows:

$$Y_i \sim \text{Exp}(\vartheta_i)\,,$$
$$\vartheta_i = \exp(\beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \beta_{k+1})\,, \quad i \in \{1, \ldots, N\}\,.$$

Letting $k = 2$, we want to test the null hypothesis H: $\beta_1 = \beta_2 = 0$ (i.e., the data are generated according to iid exponential distribution). We set the $N \times 3$ design matrix as $\mathbf{X} = (x_{ij})_{i\in\{1,\ldots,N\},j\in\{1,2,3\}} = (\boldsymbol{x}_1\ \boldsymbol{x}_2\ \mathbf{1}_N)$ (in this case, we can test the equality of the three groups), where $\boldsymbol{x}_1 = (\mathbf{1}_n^\top, \mathbf{0}_n^\top, \mathbf{0}_n^\top)^\top$ and $\boldsymbol{x}_2 = (\mathbf{0}_n^\top, \mathbf{1}_n^\top, \mathbf{0}_n^\top)^\top \in \mathbb{R}^N$ ($N = 3n$), with $\mathbf{1}_\nu$ being the $\nu \times 1$ vector of ones. Table 10.4 shows the simulation results (from 10,000 replications) of the type I errors of the large sample Rao, LR, and Wald tests, the (analytical) Bartlett-type corrected Rao, LR, and Wald tests, and the bootstrap-based Bartlett-type corrected Rao, LR, and Wald tests, when $k = 2$ (we set $N \in \{21, 45, 60\}$ and $B \in \{100, 200, 500\}$). We observe that

- the large sample Rao test, with the rejection region Rao $> \chi_k^2(\alpha)$, tends to be undersized, whereas the large sample LR and Wald tests, with the rejection regions LR $> \chi_k^2(\alpha)$ and Wald $> \chi_k^2(\alpha)$, respectively, are oversized;
- the type I errors of the analytical Bartlett-type corrections, with the rejection regions $\text{Rao}_{\text{Bart-type}} > \chi_k^2(\alpha)$, $\text{LR}_{\text{Bart}} > \chi_k^2(\alpha)$, and $\text{Wald}_{\text{Bart-type}} > \chi_k^2(\alpha)$, respectively, are close to the nominal size, and the bootstrap-based Bartlett-type corrections work even for the small bootstrap iterations.

10.5 Concluding Remarks

To achieve an accurate testing inference for the nonparametric setup as well as the parametric setup, we have mainly addressed two issues that have probably never been addressed in the literature; (i) asymptotic expansions for the distributions of χ^2 type test statistics from the modern GEL framework in the possibly over-identified moment restrictions and (ii) a refinement of the distribution (10.11) by making use of the Bartlett-type adjustment. The null distribution of each adjusted test statistic can be approximated as a central chi-squared distribution (up to order N^{-1}). It is not difficult to see that the limiting nonnull distribution (under a sequence of local alternatives) is a noncentral chi-squared distribution. However, their higher-order local power properties are, in general, not identical. Higher-order comparison of local powers of several tests have been also an active area of research. We list below some important results for the parametric likelihood-based tests that the author worked on (the cited references are kept to a minimum).

Table 10.4 The type I errors (10,000 replications) of Rao, $\text{Rao}_{\text{Bart-type}}$, $\text{Rao}^*_{\text{Bart-type}}$, LR, LR_{Bart}, $\text{LR}^*_{\text{Bart}}$, Wald, $\text{Wald}_{\text{Bart-type}}$, and $\text{Wald}^*_{\text{Bart-type}}$ tests (with significance levels $\alpha \in \{0.1, 0.05, 0.01\}$) for testing H: $\beta_1 = \beta_2 = 0$ when $N \in \{21, 45, 60\}$

		Rao	Bartlett-type corrected Rao			
			Analytical	Bootstrap-based		
				$B = 100$	200	500
$N = 21$	$\alpha = 0.1$	0.0807	0.0971	0.0963	0.0941	0.0955
	$\alpha = 0.05$	0.0371	0.0465	0.0470	0.0455	0.0459
	$\alpha = 0.01$	0.0065	0.0100	0.0102	0.0100	0.0096
$N = 45$	$\alpha = 0.1$	0.0906	0.0981	0.0977	0.0956	0.0958
	$\alpha = 0.05$	0.0438	0.0499	0.0480	0.0489	0.0498
	$\alpha = 0.01$	0.0068	0.0094	0.0109	0.0100	0.0094
$N = 60$	$\alpha = 0.1$	0.0943	0.0991	0.0998	0.0990	0.1003
	$\alpha = 0.05$	0.0471	0.0510	0.0522	0.0513	0.0517
	$\alpha = 0.01$	0.0086	0.0105	0.0111	0.0099	0.0099
		LR	Bartlett corrected LR			
			Analytical	Bootstrap-based		
				$B = 100$	200	500
$N = 21$	$\alpha = 0.1$	0.1056	0.0987	0.1015	0.0977	0.0969
	$\alpha = 0.05$	0.0557	0.0498	0.0508	0.0498	0.0504
	$\alpha = 0.01$	0.0114	0.0097	0.0107	0.0108	0.0101
$N = 45$	$\alpha = 0.1$	0.0994	0.0961	0.1003	0.0953	0.0973
	$\alpha = 0.05$	0.0508	0.0483	0.0486	0.0477	0.0471
	$\alpha = 0.01$	0.0107	0.0103	0.0100	0.0104	0.0103
$N = 60$	$\alpha = 0.1$	0.1034	0.1015	0.1000	0.0999	0.0993
	$\alpha = 0.05$	0.0543	0.0534	0.0535	0.0524	0.0537
	$\alpha = 0.01$	0.0099	0.0092	0.0100	0.0098	0.0095
		Wald	Bartlett-type corrected Wald			
			Analytical	Bootstrap-based		
				$B = 100$	200	500
$N = 21$	$\alpha = 0.1$	0.1179	0.0975	0.0937	0.0965	0.0973
	$\alpha = 0.05$	0.0648	0.0480	0.0476	0.0453	0.0458
	$\alpha = 0.01$	0.0178	0.0090	0.0122	0.0100	0.0088
$N = 45$	$\alpha = 0.1$	0.1033	0.0950	0.0951	0.0920	0.0957
	$\alpha = 0.05$	0.0560	0.0481	0.0495	0.0467	0.0476
	$\alpha = 0.01$	0.0129	0.0103	0.0133	0.0109	0.0094
$N = 60$	$\alpha = 0.1$	0.1068	0.1004	0.1003	0.0997	0.1003
	$\alpha = 0.05$	0.0574	0.0529	0.0524	0.0520	0.0531
	$\alpha = 0.01$	0.0118	0.0096	0.0117	0.0100	0.0082

1. For the multivariate analysis, in addition to the LR test, Lawley–Hotelling's (LH) trace test and Bartlett–Nanda–Pillai's (BNP) trace test are standard in multivariate regression or (G)MANOVA model. The local powers of three tests for a general linear hypothesis under a sequence of contiguous alternatives were discussed in

Anderson [1, p. 336].[15] See also Kakizawa [44] for nonnormal case, including the use of monotone Bartlett-type adjustment [42].

2. For a general parametric model, the local power analysis (up to order $N^{-1/2}$) among the LR, Rao, and Wald tests under a sequence of contiguous alternatives dates back to Peers [61], Hayakawa [36], Harris and Peers [35] in the absence/presence of the nuisance parameters, which indicates that there is, in general, no uniform ordering (with regard to the point-by-point local power) among the LR, Rao, and Wald tests without adjustment.[16]
3. One thus needs to adjust the tests in a meaningful way or use some alternative criterion to the point-by-point local power. We refer the readers to Mukerjee [55] for a comprehensive literature review on this subject till the early 1990s. As a unified treatment about the likelihood-based testing theory for a general statistical model, including iid/non-identical models or Gaussian stationary time series models, Taniguchi [72] explicitly derived N^{-1}-asymptotic expansion for the distribution (under a sequence of contiguous alternatives) of a broad class of test statistics about a scalar parameter of interest. Professor Taniguchi's [72] main contribution was the introduction of the Bartlett-type adjustment (T approach) in order to make N^{-1}-local power comparison of tests. According to Taniguchi's N^{-1}-asymptotic expansion, Rao and Mukerjee [66] further posed the issue of a comparative power study of various Bartlett-type adjustments.
4. The works in these directions have been extended to a subvector hypothesis even in the presence of the nuisance parameters, including $N^{-1/2}$-sensitivity analysis for the change of the nuisance parameters, $N^{-1/2}$-local power identity in a class of (adjusted) tests, and N^{-1}-average power comparison. See Kakizawa [45–49] and the references cited therein.

10.6 Technical Details

In this section, we prove Lemma 10.1 and derive an asymptotic expansion for the distribution of a class of the test statistics, given in (10.8). For this purpose, let

$$Z_{j_1}^{(N)} = N^{1/2}\ell_{j_1}^{(N)}(\boldsymbol{\eta}_0)\,, \quad Z_{j_1 j_2}^{(N)} = N^{1/2}\{\ell_{j_1 j_2}^{(N)}(\boldsymbol{\eta}_0) - \nu_{j_1 j_2}\}\,,$$
$$Z_{j_1\cdots j_R}^{(N)\rho} = N^{1/2}\{\ell_{j_1\cdots j_R}^{(N)\rho}(\boldsymbol{\eta}_0) - \nu^{\rho}{}_{j_1\cdots j_R}\}\,,\ R \in \{3,4\}\,.$$

[15] A seminal work in this area is an unpublished working paper of Rothenberg [68], cited by Anderson's book (see also Fujikoshi [27]); it would, however, not available almost anywhere from the world.

[16] Similar conclusion holds for the modern GEL framework; indeed, Bravo [10] explicitly derived $N^{-1/2}$-asymptotic expansion for the distribution (under a sequence of contiguous alternatives) of a class of ECR test statistics for testing a simple full vector parameter hypothesis in the just-identified moment restrictions. To the best of our knowledge, the literature is, however, little.

Also, we denote by $[\nu^{jj'}]_{j,j'\in\{1,\dots,p+M\}}$ and $[\nu^{rr'}_{(22)}]_{r,r'\in\{p_1+1,\dots,p+M\}}$ the inverse of the matrix $\boldsymbol{\nu}$ and $\boldsymbol{\nu}_{(22)}$, respectively. The dot notation is introduced for simplicity, e.g., $|\#^{(N)}_{\bullet\cdots\bullet}| = \sum_{j_1,\dots,j_R=1}^{p+M} |\#^{(N)}_{j_1\cdots j_R}|$.

10.6.1 Some Auxiliary Lemmas

Before proving Lemma 10.1 (Sect. 10.6.2) and then deriving an asymptotic expansion (Sect. 10.6.3) for the distribution of a class of the test statistics, given in (10.8), this subsection collects several lemmas. To the best of our knowledge, the related mathematical issues in the (G)EL framework are found in Bravo [11] and Ragusa [64].

The following fundamental results (Lemmas 10.2–10.6) are indispensable, as in the case of M-estimators for the parametric models (see, e.g., Bhattacharya and Ghosh [6], Taniguchi [71], and Kakizawa [45–49]).

Lemma 10.2 *Suppose that* $E_F[\sup_{\theta\in\Theta} ||\mathbf{g}(\mathbf{X},\boldsymbol{\theta})||^Q] < \infty$ *for some* $Q > 4$. *Then,*

$$P_F^{(N)}\Big[\max_{i\in\{1,\dots,N\}} \sup_{\theta\in\Theta} ||\mathbf{g}(\mathbf{X}_i,\boldsymbol{\theta})|| > \frac{1}{2} N^{1/2-\xi_0}\Big] = o(N^{-1})$$

for any constant $\xi_0 \in (0, 1/2 - 2/Q)$.

Proof Using

$$E_F^{(N)}\Big[\max_{i\in\{1,\dots,N\}} \sup_{\theta\in\Theta} ||\mathbf{g}(\mathbf{X}_i,\boldsymbol{\theta})||^Q\Big] \le N E_F\Big[\sup_{\theta\in\Theta} ||\mathbf{g}(\mathbf{X},\boldsymbol{\theta})||^Q\Big] = N K_\Theta \text{ (say)},$$

Markov's inequality yields

$$P_F^{(N)}\Big[\max_{i\in\{1,\dots,N\}} \sup_{\theta\in\Theta} ||\mathbf{g}(\mathbf{X}_i,\boldsymbol{\theta})|| > 2N^{1/2-\xi_0}\Big] \le 2^{-Q} N^{Q(\xi_0-1/2)+1} K_\Theta\,.$$

□

Remark 10.6 Suppose that assumptions (C_0)–(C_2) hold. If the event

$$\mathcal{X}_0^{(N)} = \Big\{\max_{i\in\{1,\dots,N\}} \sup_{\lambda\in B_M(0_M:2N^{-1/2}\log N),\theta\in\Theta} |\boldsymbol{\lambda}^\top \mathbf{g}(\mathbf{X}_i,\boldsymbol{\theta})| \le N^{-\xi_0}\log N\Big\}$$

occurs,[17] then,

[17] Note that

$$1 - P_F^{(N)}(\mathcal{X}_0^{(N)}) \le P_F^{(N)}\Big[\max_{i\in\{1,\dots,N\}} \sup_{\theta\in\Theta} ||\mathbf{g}(\mathbf{X}_i,\boldsymbol{\theta})|| > \frac{1}{2} N^{1/2-\xi_0}\Big] = o(N^{-1})$$

- the GEL objective function $\ell^{(N)\rho}(\boldsymbol{\eta})$ is four times continuously differentiable on $\boldsymbol{\Theta} \times B_M(\mathbf{0}_M : 2N^{-1/2} \log N)$ (say),[18] and
- there exists a constant $c_\rho = c_{\rho,\Theta} > 0$ (independent of N), such that, for all $\boldsymbol{\eta}_{\mathrm{I}} = (\boldsymbol{\theta}_{\mathrm{I}}^\top, \boldsymbol{\lambda}_{\mathrm{I}}^\top)^\top$ and $\boldsymbol{\eta}_{\mathrm{II}} = (\boldsymbol{\theta}_{\mathrm{II}}^\top, \boldsymbol{\lambda}_{\mathrm{II}}^\top)^\top$; $||\boldsymbol{\theta}_\# - \boldsymbol{\theta}_0|| \le \epsilon$ (see the comment in (C_2)) and $||\boldsymbol{\lambda}_\#|| \le N^{-1/2} \log N$ $(N \ge N_{0,\rho})$,

$$|\ell^{(N)\rho}_{\bullet\bullet\bullet\bullet}(\boldsymbol{\eta}_{\mathrm{I}}) - \ell^{(N)\rho}_{\bullet\bullet\bullet\bullet}(\boldsymbol{\eta}_{\mathrm{II}})| \le c_\rho ||\boldsymbol{\eta}_{\mathrm{I}} - \boldsymbol{\eta}_{\mathrm{II}}|| \sum_{\nu=1}^{5} (|Z^{(N)}_{B^\nu}| + \mu_{B^\nu}),$$

where $\mu_{B^\nu} = E_F[B^\nu(\mathbf{X})]$ and $Z^{(N)}_{B^\nu} = N^{-1} \sum_{i=1}^N \{B^\nu(\mathbf{X}_i) - \mu_{B^\nu}\}$.

Lemma 10.3 *Suppose that assumptions* (C_0)–(C_2) *hold with* $Q \ge 16$. *Then,*

(i) $P_F^{(N)}\Big[\sum_{\nu=1}^{5} |Z^{(N)}_{B^\nu}| > 1\Big] = o(N^{-1})$;

(ii) $P_F^{(N)}\big[|Z^{(N)}_{\bullet}| + |Z^{(N)}_{\bullet\bullet}| + |Z^{(N)\rho}_{\bullet\bullet\bullet}| + |Z^{(N)\rho}_{\bullet\bullet\bullet\bullet}| > d_{1,\rho}(\log N)^{1/2}\big] = o(N^{-1}(\log N)^{-2})$,

where $d_{1,\rho} = d_{1,\rho,\theta_0} > 0$ *is a constant (independent of* N*);*
(iii) $P_F^{(N)}[\mathcal{X}^{(N)\rho}] = 1 - o(N^{-1})$, *where* $\mathcal{X}^{(N)\rho} = \mathcal{X}_0^{(N)} \bigcap \mathcal{X}_0^{(N)\rho}$, *with*

$$\mathcal{X}_0^{(N)\rho} = \Bigg\{ \sum_{\nu=1}^{5} |Z^{(N)}_{B^\nu}| \le 1 \text{ and } |Z^{(N)}_{\bullet}| + |Z^{(N)}_{\bullet\bullet}| + |Z^{(N)\rho}_{\bullet\bullet\bullet}| + |Z^{(N)\rho}_{\bullet\bullet\bullet\bullet}| \le d_{1,\rho}(\log N)^{1/2} \Bigg\}.$$

Proof Since $B^\nu(\mathbf{X})$, $\nu \in \{1, \ldots, 5\}$, have the sth absolute moment for $s \in (2, Q/5]$, Markov's and Rosenthal's inequalities yield

$$P_F^{(N)}\big[|Z^{(N)}_{B^\nu}| > 1\big] \le O(N^{-s/2}) = o(N^{-1}).$$

On the other hand, for the proof of (ii), we have only to apply the moderate deviation estimate (we set $t = 4$); see, e.g., Bhattacharya and Rao [7, Corollary 17.12]: If $\psi(\mathbf{X}, \boldsymbol{\theta}_0)$, taking values in $\mathbb{R}$, satisfies

$$E_F[\psi(\mathbf{X}, \boldsymbol{\theta}_0)] = 0, \quad Var_F[\psi(\mathbf{X}, \boldsymbol{\theta}_0)] > 0, \quad \text{and} \quad E_F[|\psi(\mathbf{X}, \boldsymbol{\theta}_0)|^t] < \infty$$

for some integer $t \ge 3$, then,

(see Lemma 10.2).

18 There exists an integer $N_{0,\rho}$ such that $\pm N^{-\xi_0} \log N \in \mathcal{N}_\rho (\subset \mathcal{V}_\rho)$ for all $N \ge N_{0,\rho}$.

$$P_F^{(N)}\Big[\frac{1}{N^{1/2}}\Big|\sum_{i=1}^{N}\psi(\mathbf{X}_i,\boldsymbol{\theta}_0)\Big| > Var_F^{1/2}[\psi(\mathbf{X},\boldsymbol{\theta}_0)]\{(t-1)\log N\}^{1/2}\Big]$$
$$= o(N^{-(t-2)/2}(\log N)^{-t/2})$$

(this estimate remains valid even if $E_F[\psi(\mathbf{X},\boldsymbol{\theta}_0)] = Var_F[\psi(\mathbf{X},\boldsymbol{\theta}_0)] = 0$, since $\psi(\mathbf{X},\boldsymbol{\theta}_0)$ is then degenerate, i.e., $P_F[\psi(\mathbf{X},\boldsymbol{\theta}_0) = 0] = 1$). □

Recall that we write $Y^{(N)} = o_F^{(N)}(q_1, q_2)$, if $P_F^{(N)}[|Y^{(N)}| \geq \gamma(\log N)^{q_2}] = o(N^{-q_1})$ for some constants $\gamma > 0$ and $q_1, q_2 \geq 0$, independent of N (see (10.7)).

Lemma 10.4 *Suppose that assumptions* (C$_0$)–(C$_3$) *hold with* $Q \geq 16$. *Then, there exist sequences of statistics* $\{\widehat{\boldsymbol{\eta}}^{(N)\rho}\}_{N\geq 1}$ *and* $\{\widetilde{\boldsymbol{\eta}}_{(2)}^{(N)\rho}\}_{N\geq 1}$, *such that*

$$P_F^{(N)}\Big[\bigcap_{i=1}^{4}\mathcal{X}_i^{(N)\rho}\Big] = 1 - o(N^{-1})\,, \tag{10.13}$$

with

$$\begin{aligned}
\mathcal{X}_1^{(N)\rho} &= \Big\{||\widehat{\boldsymbol{\eta}}^{(N)\rho} - \boldsymbol{\eta}_0|| < d_\rho\frac{(\log N)^{1/2}}{N^{1/2}}\,,\ \widehat{\boldsymbol{\eta}}^{(N)\rho} \textit{ solves } \widehat{\mathbf{Z}}^{(N)\rho} = \mathbf{0}_{p+M}\Big\}\,,\\
\mathcal{X}_2^{(N)\rho} &= \Big\{||\widetilde{\boldsymbol{\eta}}^{(N)\rho} - \boldsymbol{\eta}_0|| < d_\rho\frac{(\log N)^{1/2}}{N^{1/2}}\,,\ \widetilde{\boldsymbol{\eta}}^{(N)\rho} = \begin{pmatrix}\boldsymbol{\theta}_{(1)0}\\ \widetilde{\boldsymbol{\eta}}_{(2)}^{(N)\rho}\end{pmatrix} \textit{ solves } \widetilde{\mathbf{Z}}_{(2)}^{(N)\rho} = \mathbf{0}_{p_2+M}\Big\}\,,\\
\mathcal{X}_3^{(N)\rho} &= \Big\{\widehat{\mathbf{L}}^{(N)\rho} \textit{ is nonsingular, } ||(\widehat{\mathbf{L}}^{(N)\rho})^{-1}|| < 2||\boldsymbol{\nu}^{-1}||\Big\}\,,\\
\mathcal{X}_4^{(N)\rho} &= \Big\{\widetilde{\mathbf{L}}^{(N)\rho} \textit{ is nonsingular, } ||(\widetilde{\mathbf{L}}^{(N)\rho})^{-1}|| < 2||\boldsymbol{\nu}^{-1}||\Big\}\,,
\end{aligned}$$

where $d_\rho = 2d_{1,\rho}\max(||\boldsymbol{\nu}^{-1}||, ||\boldsymbol{\nu}_{(22)}^{-1}||)$*; besides, both* $\widehat{\boldsymbol{\eta}}^{(N)\rho}$ *and* $\widetilde{\boldsymbol{\eta}}_{(2)}^{(N)\rho}$ *admit the stochastic expansions, as follows:*

$$\boldsymbol{\eta}^{(N)\rho} = \boldsymbol{\eta}^{0(N)} + \frac{\boldsymbol{\eta}^{1(N)\rho}}{N^{1/2}} + \frac{\boldsymbol{\eta}^{2(N)\rho}}{N} + \frac{1}{N^{3/2}}\,o_F^{(N)}(1,2) \tag{10.14}$$

and

$$\boldsymbol{\zeta}_{(2)}^{(N)\rho} = \boldsymbol{\zeta}_{(2)}^{0(N)} + \frac{\boldsymbol{\zeta}_{(2)}^{1(N)\rho}}{N^{1/2}} + \frac{\boldsymbol{\zeta}_{(2)}^{2(N)\rho}}{N} + \frac{1}{N^{3/2}}\,o_F^{(N)}(1,2) \tag{10.15}$$

(we write $\boldsymbol{\eta}^{(N)\rho} = N^{1/2}(\widehat{\boldsymbol{\eta}}^{(N)\rho} - \boldsymbol{\eta}_0)$ *and* $\boldsymbol{\zeta}_{(2)}^{(N)\rho} = N^{1/2}(\widetilde{\boldsymbol{\eta}}_{(2)}^{(N)\rho} - \boldsymbol{\eta}_{(2)0})$*, respectively), where the* j*th elements of* $\boldsymbol{\eta}^{0(N)}$ *and* $\boldsymbol{\eta}^{i(N)\rho}$*,* $i \in \{1, 2\}$*, are given by*

$$\eta_j^{0(N)} = -\nu^{jj'} Z_{j'}^{(N)}, \quad \eta_j^{1(N)\rho} = -\nu^{jj'} \Big(Z_{j'j_2}^{(N)} \eta_{j_2}^{0(N)} + \frac{1}{2}\, \nu^{\rho}{}_{j'j_2j_3} \prod_{i=2}^{3} \eta_{j_i}^{0(N)} \Big),$$

$$\eta_j^{2(N)\rho} = -\nu^{jj'} \Big(Z_{j'j_2}^{(N)} \eta_{j_2}^{1(N)\rho} + \nu^{\rho}{}_{j'j_2j_3} \eta_{j_2}^{1(N)\rho} \eta_{j_3}^{0(N)} + \frac{1}{2}\, Z_{j'j_2j_3}^{(N)\rho} \prod_{i=2}^{3} \eta_{j_i}^{0(N)} + \frac{1}{6}\, \nu^{\rho}{}_{j'j_2j_3j_4} \prod_{i=2}^{4} \eta_{j_i}^{0(N)} \Big),$$

whereas the rth elements of $\boldsymbol{\zeta}_{(2)}^{0(N)}$ *and* $\boldsymbol{\zeta}_{(2)}^{i(N)\rho}$, $i \in \{1, 2\}$, *are given by*

$$\zeta_r^{0(N)} = -\nu_{(22)}^{rr'} Z_{r'}^{(N)}, \quad \zeta_r^{1(N)\rho} = -\nu_{(22)}^{rr'} \Big(Z_{r'r_2}^{(N)} \zeta_{r_2}^{0(N)} + \frac{1}{2}\, \nu^{\rho}{}_{r'r_2r_3} \prod_{i=2}^{3} \zeta_{r_i}^{0(N)} \Big),$$

$$\zeta_r^{2(N)\rho} = -\nu_{(22)}^{rr'} \Big(Z_{r'r_2}^{(N)} \zeta_{r_2}^{1(N)\rho} + \nu^{\rho}{}_{r'r_2r_3} \zeta_{r_2}^{1(N)\rho} \zeta_{r_3}^{0(N)} + \frac{1}{2}\, Z_{r'r_2r_3}^{(N)\rho} \prod_{i=2}^{3} \zeta_{r_i}^{0(N)} + \frac{1}{6}\, \nu^{\rho}{}_{r'r_2r_3r_4} \prod_{i=2}^{4} \zeta_{r_i}^{0(N)} \Big).$$

Proof We first prove (10.13) by using the argument of Bhattacharya and Ghosh [6], except for the use of Lemma 10.3, and then derive the stochastic expansion (10.15) (since we can obtain (10.14) analogously, we omit its detail, to save space) by the successive substitution as in Taniguchi [71, p. 76].

Proof of (10.13): Consider the event $\mathcal{X}^{(N)\rho}$ (see Remark 10.6 and Lemma 10.3). It suffices to prove that $\mathcal{X}^{(N)\rho} \subset \bigcap_{i=1}^{4} \mathcal{X}_i^{(N)\rho}$; (10.13) is concluded immediately from Lemma 10.3(iii).

With regard to the $p + M$ equations $\ell_j^{(N)\rho}(\boldsymbol{\eta}) = 0$, $j \in \{1, \ldots, p + M\}$, we rewrite $\boldsymbol{\eta} = \boldsymbol{\eta}_0 - \boldsymbol{\nu}^{-1} \boldsymbol{\psi}^{(N)\rho}(\boldsymbol{\eta})$, where the jth element of $\boldsymbol{\psi}^{(N)\rho}(\boldsymbol{\eta})$ is given by

$$\psi_j^{(N)\rho}(\boldsymbol{\eta}) = \frac{1}{N^{1/2}}\, Z_j^{(N)} + \frac{1}{N^{1/2}}\, Z_{jj_2}^{(N)} [\boldsymbol{\eta} - \boldsymbol{\eta}_0]_{j_2} + \frac{1}{2}\, \nu^{\rho}{}_{jj_2j_3} \prod_{i=2}^{3} [\boldsymbol{\eta} - \boldsymbol{\eta}_0]_{j_i} + \frac{1}{2N^{1/2}}\, Z_{jj_2j_3}^{(N)\rho} \prod_{i=2}^{3} [\boldsymbol{\eta} - \boldsymbol{\eta}_0]_{j_i} + \frac{1}{6}\, \nu^{\rho}{}_{jj_2j_3j_4} \prod_{i=2}^{4} [\boldsymbol{\eta} - \boldsymbol{\eta}_0]_{j_i} + R_j^{(N)\rho}(\boldsymbol{\eta}),$$

and

$$R_j^{(N)\rho}(\boldsymbol{\eta}) = \frac{1}{6N^{1/2}}\, Z_{jj_2j_3j_4}^{(N)\rho} \prod_{i=2}^{4} [\boldsymbol{\eta} - \boldsymbol{\eta}_0]_{j_i} + \frac{1}{2} \prod_{i=2}^{4} [\boldsymbol{\eta} - \boldsymbol{\eta}_0]_{j_i} \int_0^1 (1 - u)^2 [\ell_{jj_2j_3j_4}^{(N)\rho} \{\boldsymbol{\eta}_0 + u(\boldsymbol{\eta} - \boldsymbol{\eta}_0)\} - \ell_{jj_2j_3j_4}^{(N)\rho}]\, du,$$

with

$$|R_{\bullet}^{(N)\rho}(\boldsymbol{\eta})| \le \frac{1}{6}\,||\boldsymbol{\eta}-\boldsymbol{\eta}_0||^3\Big[\frac{1}{N^{1/2}}\,|Z_{\bullet\bullet\bullet\bullet}^{(N)\rho}|+c_\rho||\boldsymbol{\eta}-\boldsymbol{\eta}_0||\sum_{v=1}^{5}(|Z_{B^v}^{(N)}|+\mu_{B^v})\Big].$$

Let $\epsilon^{(N)\rho}=d_\rho N^{-1/2}(\log N)^{1/2}$, where $d_\rho=2d_{1,\rho}\max(||\boldsymbol{\nu}^{-1}||,||\boldsymbol{\nu}_{(22)}^{-1}||)$. There exists an integer $N_{1,\rho}=N_{1,\rho,\theta_0}$, such that $N^{-1/2}(\log N)^{1/2}<\min(\epsilon/d_\rho,d_{1,\rho}/D_\rho)$ for all $N\ge N_{1,\rho}$, where

$$D_\rho=d_\rho d_{1,\rho}+d_\rho^2(|\nu^\rho{}_{\bullet\bullet\bullet}|+d_{1,\rho})+d_\rho^3(|\nu^\rho{}_{\bullet\bullet\bullet\bullet}|+d_{1,\rho})+c_\rho d_\rho^4\Big(1+\sum_{v=1}^{5}\mu_{B^v}\Big).$$

On the event $\mathcal{X}^{(N)\rho}$ $(N\ge\max(N_{0,\rho},N_{1,\rho},\mathrm{e}^{d_\rho^2}))$, it is easy to see that

$$\begin{aligned}\sup_{||\eta-\eta_0||\le\epsilon^{(N)\rho}}||\boldsymbol{\nu}^{-1}\boldsymbol{\psi}^{(N)\rho}(\boldsymbol{\eta})|| &\le ||\boldsymbol{\nu}^{-1}||\sup_{||\eta-\eta_0||\le\epsilon^{(N)\rho}}|\psi_{\bullet}^{(N)\rho}(\boldsymbol{\eta})|\\ &\le ||\boldsymbol{\nu}^{-1}||\Big(\frac{\log N}{N}\Big)^{1/2}\Big[d_{1,\rho}+D_\rho\Big(\frac{\log N}{N}\Big)^{1/2}\Big]\\ &<\epsilon^{(N)\rho},\end{aligned}$$

hence, the equation $\boldsymbol{\eta}=\boldsymbol{\eta}_0-\boldsymbol{\nu}^{-1}\boldsymbol{\psi}^{(N)\rho}(\boldsymbol{\eta})$, equivalently, the GEL first-order condition $\mathbf{Z}^{(N)\rho}(\boldsymbol{\eta})=\mathbf{0}_{p+M}$, has a unique root $\boldsymbol{\eta}=\dot{\boldsymbol{\eta}}^{(N)\rho}$ lying in $B_{p+M}(\boldsymbol{\eta}_0:\epsilon^{(N)\rho})$ (apply the Brower fixed point theorem). In the exact same way (except for $\boldsymbol{\eta}=(\boldsymbol{\theta}_{(1)0}^\top,\boldsymbol{\eta}_{(2)}^\top)^\top$), with regard to the p_2+M equations $\ell_r^{(N)\rho}(\boldsymbol{\eta})=0$, $r\in\{p_1+1,\dots,p+M\}$, we rewrite $\boldsymbol{\eta}_{(2)}=(\boldsymbol{\theta}_{(2)0}^\top,\mathbf{0}_M^\top)^\top-\boldsymbol{\nu}_{(22)}^{-1}\boldsymbol{\psi}_{(2)}^{(N)\rho}(\boldsymbol{\eta})$, where $\boldsymbol{\psi}_{(2)}^{(N)\rho}(\boldsymbol{\eta})=[\psi_r^{(N)\rho}(\boldsymbol{\eta})]_{r\in\{p_1+1,\dots,p+M\}}$; then, on the event $\mathcal{X}^{(N)\rho}$ $(N\ge\max(N_{0,\rho},N_{1,\rho},\mathrm{e}^{d_\rho^2}))$,

$$\begin{aligned}&\sup_{\theta_{(1)}=\theta_{(1)0}\,,\,||\eta_{(2)}-\eta_{(2)0}||\le\epsilon^{(N)\rho}}||\boldsymbol{\nu}_{(22)}^{-1}\boldsymbol{\psi}_{(2)}^{(N)\rho}(\boldsymbol{\eta})||\\ &\le||\boldsymbol{\nu}_{(22)}^{-1}||\sup_{\theta_{(1)}=\theta_{(1)0}\,,\,||\eta_{(2)}-\eta_{(2)0}||\le\epsilon^{(N)\rho}}\sum_{r=p_1+1}^{p+M}|\psi_r^{(N)\rho}(\boldsymbol{\eta})|<\epsilon^{(N)\rho},\end{aligned}$$

which implies that the (restricted) GEL first-order condition $\mathbf{Z}_{(2)}^{(N)\rho}(\boldsymbol{\eta})=\mathbf{0}_{p_2+M}$ under the constraint $\boldsymbol{\eta}=(\boldsymbol{\theta}_{(1)0}^\top,\boldsymbol{\eta}_{(2)}^\top)^\top$ has a unique root $\boldsymbol{\eta}=(\boldsymbol{\theta}_{(1)0}^\top,(\ddot{\boldsymbol{\eta}}_{(2)}^{(N)\rho})^\top)^\top$ lying in $B_{p+M}(\boldsymbol{\eta}_0:\epsilon^{(N)\rho})$. Thus, we have $\mathcal{X}^{(N)\rho}\subset\mathcal{X}_1^{(N)\rho}\bigcap\mathcal{X}_2^{(N)\rho}$.

Furthermore, we rewrite $\mathbf{L}^{(N)\rho}(\boldsymbol{\eta})=\boldsymbol{\nu}+\boldsymbol{\Psi}^{(N)\rho}(\boldsymbol{\eta})$, where the (j,j')th element of $\boldsymbol{\Psi}^{(N)\rho}(\boldsymbol{\eta})$ is given by

$$\begin{aligned}\psi_{jj'}^{(N)\rho}(\boldsymbol{\eta})=\frac{1}{N^{1/2}}\,Z_{jj'}^{(N)}+\nu^\rho{}_{jj'j_3}[\boldsymbol{\eta}-\boldsymbol{\eta}_0]_{j_3}+\frac{1}{N^{1/2}}\,Z_{jj'j_3}^{(N)\rho}[\boldsymbol{\eta}-\boldsymbol{\eta}_0]_{j_3}\\ +\frac{1}{2}\,\nu^\rho{}_{jj'j_3j_4}\prod_{i=3}^{4}[\boldsymbol{\eta}-\boldsymbol{\eta}_0]_{j_i}+R_{jj'}^{(N)\rho}(\boldsymbol{\eta})\,,\end{aligned}$$

with

$$|R_{\bullet\bullet}^{(N)\rho}(\boldsymbol{\eta})| \leq \frac{1}{2}\,||\boldsymbol{\eta}-\boldsymbol{\eta}_0||^2\Big[\frac{1}{N^{1/2}}\,|Z_{\bullet\bullet\bullet\bullet}^{(N)\rho}| + c_\rho||\boldsymbol{\eta}-\boldsymbol{\eta}_0||\sum_{v=1}^{5}(|Z_{B^v}^{(N)}|+\mu_{B^v})\Big].$$

On the event $\mathcal{X}^{(N)\rho}$ $(N \geq \max(N_{0,\rho}, N_{1,\rho}, \mathrm{e}^{d_\rho^2}))$, it is easy to see that

$$\begin{aligned}\sup_{||\eta-\eta_0||\leq\epsilon^{(N)\rho}} ||\boldsymbol{\Psi}^{(N)\rho}(\boldsymbol{\eta})|| &\leq \sup_{||\eta-\eta_0||\leq\epsilon^{(N)\rho}} \sum_{jj'=1}^{p+M} |{}_\rho\psi_{jj'}^{(N)}(\boldsymbol{\eta})| \\ &\leq \frac{D_\rho}{d_\rho}\Big(\frac{\log N}{N}\Big)^{1/2} < \frac{1}{2||\boldsymbol{\nu}^{-1}||}\,,\end{aligned}$$

hence,[19]

$$\begin{aligned}\inf_{||\eta-\eta_0||\leq\epsilon^{(N)\rho}} \rho_{M+1}^{\uparrow}\{\mathbf{L}^{(N)\rho}(\boldsymbol{\eta})\} &> \frac{1}{2||\boldsymbol{\nu}^{-1}||}\,,\\ \sup_{||\eta-\eta_0||\leq\epsilon^{(N)\rho}} \rho_{M}^{\uparrow}\{\mathbf{L}^{(N)\rho}(\boldsymbol{\eta})\} &< -\frac{1}{2||\boldsymbol{\nu}^{-1}||}\end{aligned}$$

(it turns out that, uniformly in $\boldsymbol{\eta}$; $||\boldsymbol{\eta}-\boldsymbol{\eta}_0|| \leq \epsilon^{(N)\rho}$, $\mathbf{L}^{(N)\rho}(\boldsymbol{\eta})$ is nonsingular, with $||\{\mathbf{L}^{(N)\rho}(\boldsymbol{\eta})\}^{-1}|| < 2||\boldsymbol{\nu}^{-1}||$). Thus, we also have $\mathcal{X}^{(N)\rho} \subset \mathcal{X}_3^{(N)\rho} \bigcap \mathcal{X}_4^{(N)\rho}$.

Proof of (10.15): We start with

$$\begin{aligned}\zeta_r^{(N)\rho} = \zeta_r^{0(N)} - \nu_{(22)}^{rr'}\Big[&\frac{1}{N^{1/2}}\Big(Z_{r'r_2}^{(N)}\zeta_{r_2}^{(N)\rho} + \frac{1}{2}\,\nu^{\rho}{}_{r'r_2r_3}\prod_{i=2}^{3}\zeta_{r_i}^{(N)\rho}\Big)\\ &+\frac{1}{N}\Big(\frac{1}{2}\,Z_{r'r_2r_3}^{(N)\rho}\prod_{i=2}^{3}\zeta_{r_i}^{(N)\rho} + \frac{1}{6}\,\nu^{\rho}{}_{r'r_2r_3r_4}\prod_{i=2}^{4}\zeta_{r_i}^{(N)\rho}\Big) + \frac{1}{N^{3/2}}\,o_F^{(N)}(1,2)\Big]\end{aligned} \tag{10.16}$$

for $r \in \{p_1+1, \ldots, M\}$.

By (10.16), $\boldsymbol{\zeta}_{(2)}^{(N)\rho} = \boldsymbol{\zeta}_{(2)}^{0(N)} + N^{-1/2}\mathbf{e}_{(2)}^{0(N)\rho}$, where $\mathbf{e}_{(2)}^{0(N)\rho} = o_F^{(N)}(1,1)$.

[19] Apply the fact (10.1) to get

$$\begin{aligned}\rho_{M+1}^{\uparrow}(\boldsymbol{\nu}+\boldsymbol{\Psi}^{(N)\rho}(\boldsymbol{\eta})) &\geq \rho_{M+1}^{\uparrow}(\boldsymbol{\nu}) - ||\boldsymbol{\Psi}^{(N)\rho}(\boldsymbol{\eta})|| \geq \frac{1}{||\boldsymbol{\nu}^{-1}||} - ||\boldsymbol{\Psi}^{(N)\rho}(\boldsymbol{\eta})|| > \frac{1}{2||\boldsymbol{\nu}^{-1}||}\,,\\ \rho_{M}^{\uparrow}(\boldsymbol{\nu}+\boldsymbol{\Psi}^{(N)\rho}(\boldsymbol{\eta})) &\leq \rho_{M}^{\uparrow}(\boldsymbol{\nu}) + ||\boldsymbol{\Psi}^{(N)\rho}(\boldsymbol{\eta})|| \leq -\frac{1}{||\boldsymbol{\nu}^{-1}||} + ||\boldsymbol{\Psi}^{(N)\rho}(\boldsymbol{\eta})|| < -\frac{1}{2||\boldsymbol{\nu}^{-1}||}\,,\end{aligned}$$

since $\min\{\rho_{M+1}^{\uparrow}(\boldsymbol{\nu}), -\rho_{M}^{\uparrow}(\boldsymbol{\nu})\} = 1/||\boldsymbol{\nu}^{-1}||$ (see (10.6)).

Next, by substituting $\widetilde{\boldsymbol{\eta}}_{(2)}^{(N)\rho} = \boldsymbol{\eta}_{(2)0} + N^{-1/2}\boldsymbol{\zeta}_{(2)}^{0(N)} + N^{-1}\mathbf{e}_{(2)}^{0(N)\rho}$ for the $N^{-1/2}$-terms in the right-hand side of (10.16), we have $\boldsymbol{\zeta}_{(2)}^{(N)\rho} = \boldsymbol{\zeta}_{(2)}^{0(N)} + N^{-1/2}\boldsymbol{\zeta}_{(2)}^{1(N)\rho} + N^{-1}\mathbf{e}_{(2)}^{1(N)\rho}$, where $\mathbf{e}_{(2)}^{1(N)\rho} = o_F^{(N)}(1, 3/2)$.

Finally, by substituting $\widetilde{\boldsymbol{\eta}}_{(2)}^{(N)\rho} = \boldsymbol{\eta}_{(2)0} + N^{-1/2}\boldsymbol{\zeta}_{(2)}^{0(N)} + N^{-1}\boldsymbol{\zeta}_{(2)}^{1(N)\rho} + N^{-3/2}\mathbf{e}_{(2)}^{1(N)\rho}$ ($\widetilde{\boldsymbol{\eta}}_{(2)}^{(N)\rho} = \boldsymbol{\eta}_{(2)0} + N^{-1/2}\boldsymbol{\zeta}_{(2)}^{0(N)} + N^{-1}\mathbf{e}_{(2)}^{0(N)\rho}$) for the $N^{-1/2}$-terms (N^{-1}-terms) in the right-hand side of (10.16), we have $\boldsymbol{\zeta}_{(2)}^{(N)\rho} = \boldsymbol{\zeta}_{(2)}^{0(N)} + \sum_{i=1}^{2} N^{-i/2}\boldsymbol{\zeta}_{(2)}^{i(N)\rho} + N^{-3/2}\mathbf{e}_{(2)}^{2(N)\rho}$, where $\mathbf{e}_{(2)}^{2(N)\rho} = o_F^{(N)}(1, 2)$. □

We can connect $\widehat{\boldsymbol{\eta}}^{(N)\rho}$ to $\widetilde{\boldsymbol{\eta}}^{(N)\rho}$, as follows: Write $\boldsymbol{\delta}^{(N)\rho} = N^{1/2}(\widehat{\boldsymbol{\eta}}^{(N)\rho} - \widetilde{\boldsymbol{\eta}}^{(N)\rho})$, where the jth element of $\boldsymbol{\delta}^{(N)\rho}$ is denoted by $\delta_j^{(N)\rho}$.

Lemma 10.5 *Suppose that assumptions* (C$_0$)–(C$_3$) *hold with* $Q \geq 16$. *Then,*

$$\boldsymbol{\delta}^{(N)\rho} = \boldsymbol{\delta}^{0(N)\rho} + \frac{1}{N^{1/2}}\,\boldsymbol{\delta}^{1(N)\rho} + \frac{1}{N}\,\boldsymbol{\delta}^{2(N)\rho} + \frac{1}{N^{3/2}}\,o_F^{(N)}(1, 2)\,, \tag{10.17}$$

where the jth elements of $\boldsymbol{\delta}^{i(N)\rho}$, $i \in \{0, 1, 2\}$, are given by

$$\begin{aligned}
\delta_j^{0(N)\rho} &= -\left[(\widetilde{\mathbf{L}}^{(N)\rho})^{-1}\begin{pmatrix}\widetilde{\mathbf{Z}}_{(1)}^{(N)\rho}\\ \mathbf{0}_{p_2+M}\end{pmatrix}\right]_j = -[\widetilde{\mathcal{G}}^{(N)\rho}(\widetilde{\mathbf{L}}_{(11\cdot 2)}^{(N)\rho})^{-1}\widetilde{\mathbf{Z}}_{(1)}^{(N)\rho}]_j\,,\\
\delta_j^{1(N)\rho} &= -\frac{1}{2}\,(\widetilde{\ell}^{(N)\rho})^{jj'}\widetilde{\ell}_{j'j_2j_3}^{(N)\rho}\prod_{i=2}^{3}\delta_{j_i}^{0(N)\rho}\,,\\
\delta_j^{2(N)\rho} &= -(\widetilde{\ell}^{(N)\rho})^{jj'}\left(\widetilde{\ell}_{j'j_2j_3}^{(N)\rho}\delta_{j_2}^{1(N)\rho}\delta_{j_3}^{0(N)\rho} + \frac{1}{6}\,\widetilde{\ell}_{j'j_2j_3j_4}^{(N)\rho}\prod_{i=2}^{4}\delta_{j_i}^{0(N)\rho}\right).
\end{aligned}$$

Proof Expanding $\widehat{\ell}_j^{(N)\rho}$ around $\boldsymbol{\eta} = \widetilde{\boldsymbol{\eta}}^{(N)\rho}$, we have, for $j \in \{1, \ldots, p + M\}$,

$$\begin{aligned}
0 &= N^{1/2}\widehat{\ell}_j^{(N)\rho}\\
&= N^{1/2}\widetilde{\ell}_j^{(N)\rho} + \widetilde{\ell}_{jj_2}^{(N)\rho}\delta_{j_2}^{(N)\rho} + \frac{1}{2N^{1/2}}\,\widetilde{\ell}_{jj_2j_3}^{(N)\rho}\prod_{i=2}^{3}\delta_{j_i}^{(N)\rho} + \frac{1}{6N}\,\widetilde{\ell}_{jj_2j_3j_4}^{(N)\rho}\prod_{i=2}^{4}\delta_{j_i}^{(N)\rho}\\
&\quad + \frac{1}{N^{3/2}}\,o_F^{(N)}(1, 2)\,,
\end{aligned}$$

equivalently,

$$\begin{aligned}
-\widetilde{\ell}_{jj_2}^{(N)\rho}\delta_{j_2}^{(N)\rho} &= \chi_{\{j=a\}}\widetilde{Z}_a^{(N)\rho} + \frac{1}{2N^{1/2}}\,\widetilde{\ell}_{jj_2j_3}^{(N)\rho}\prod_{i=2}^{3}\delta_{j_i}^{(N)\rho} + \frac{1}{6N}\,\widetilde{\ell}_{jj_2j_3j_4}^{(N)\rho}\prod_{i=2}^{4}\delta_{j_i}^{(N)\rho}\\
&\quad + \frac{1}{N^{3/2}}\,o_F^{(N)}(1, 2)\,,
\end{aligned}$$

where the indicator function $\chi_{\{\cdot\}}$ takes on a value of 1 if the expression inside braces is true, and 0 otherwise. Stating with

$$\delta_j^{(N)\rho} = \delta_j^{0(N)\rho} - (\widetilde{\ell}^{(N)\rho})^{jj'}\Big[\frac{1}{2N^{1/2}}\,\widetilde{\ell}^{(N)\rho}_{j'j_2j_3}\prod_{i=2}^{3}\delta_{j_i}^{(N)\rho} + \frac{1}{6N}\,\widetilde{\ell}^{(N)\rho}_{j'j_2j_3j_4}\prod_{i=2}^{4}\delta_{j_i}^{(N)\rho} + \frac{1}{N^{3/2}}\,o_F^{(N)}(1,2)\Big]$$

for $j \in \{1, \dots, p + M\}$, (10.17) is concluded in the same way as the proof of (10.15), except for the repeatedly use of

$$\sum_{a=1}^{p_1} |\widetilde{Z}_a^{(N)\rho}| = o_F^{(N)}(1, 1/2)\,, \quad ||(\widetilde{\mathbf{L}}^{(N)\rho})^{-1}|| + |\widetilde{\ell}^{(N)\rho}_{\bullet\bullet\bullet}| + |\widetilde{\ell}^{(N)\rho}_{\bullet\bullet\bullet\bullet}| = o_F^{(N)}(1, 0) \tag{10.18}$$

(see Lemmas 10.3, 10.4, and 10.6). □

Lemma 10.6 *Suppose that assumptions* (C_0)–(C_3) *hold with* $Q \geq 16$. *Then,*

$$(i)\quad \widetilde{Z}_a^{(N)\rho} = Z_a^{0(N)} + \frac{1}{N^{1/2}}\,Z_a^{1(N)\rho} + \frac{1}{N}\,Z_a^{2(N)\rho} + \frac{1}{N^{3/2}}\,o_F^{(N)}(1, 2)\,;$$

$$(ii)\quad \widetilde{\ell}^{(N)\rho}_{jj'} = \nu_{jj'} + \frac{1}{N^{1/2}}\,Z_{jj'}^{0(N)\rho} + \frac{1}{N}\,Z_{jj'}^{1(N)\rho} + \frac{1}{N^{3/2}}\,o_F^{(N)}(1, 3/2)\,;$$

$$(iii)\quad \widetilde{\ell}^{(N)\rho}_{jj'j''} = \nu^{\rho}{}_{jj'j''} + \frac{1}{N^{1/2}}\,(Z^{(N)\rho}_{jj'j''} + \nu^{\rho}{}_{jj'j''r_4}\zeta_{r_4}^{0(N)}) + \frac{1}{N}\,o_F^{(N)}(1, 1)\,;$$

$$(iv)\quad \widetilde{\ell}^{(N)\rho}_{jj'j''j'''} = \nu^{\rho}{}_{jj'j''j'''} + \frac{1}{N^{1/2}}\,o_F^{(N)}(1, 1/2)\,,$$

where

$$Z_a^{0(N)} = \mathcal{G}_{ja}Z_j^{(N)} = [\mathcal{G}^{\top}\mathbf{Z}^{(N)}]_a\,,$$

$$Z_a^{1(N)\rho} = \mathcal{G}_{ja}\Big(Z^{(N)}_{jr_2}\zeta_{r_2}^{0(N)} + \frac{1}{2}\,\nu^{\rho}{}_{jr_2r_3}\prod_{i=2}^{3}\zeta_{r_i}^{0(N)}\Big)\,,$$

$$Z_a^{2(N)\rho} = \mathcal{G}_{ja}\Big(Z^{(N)}_{jr_2}\zeta_{r_2}^{1(N)\rho} + \nu^{\rho}{}_{jr_2r_3}\zeta_{r_2}^{1(N)\rho}\zeta_{r_3}^{0(N)} + \frac{1}{2}\,Z^{(N)\rho}_{jr_2r_3}\prod_{i=2}^{3}\zeta_{r_i}^{0(N)} + \frac{1}{6}\,\nu^{\rho}{}_{jr_2r_3r_4}\prod_{i=2}^{4}\zeta_{r_i}^{0(N)}\Big)\,,$$

$$Z_{jj'}^{0(N)\rho} = Z^{(N)}_{jj'} + \nu^{\rho}{}_{jj'r_3}\zeta_{r_3}^{0(N)}\,,$$

$$Z_{jj'}^{1(N)\rho} = \nu^{\rho}{}_{jj'r_3}\zeta_{r_3}^{1(N)\rho} + Z^{(N)\rho}_{jj'r_3}\zeta_{r_3}^{0(N)} + \frac{1}{2}\,\nu^{\rho}{}_{jj'r_3r_4}\prod_{i=3}^{4}\zeta_{r_i}^{0(N)}\,.$$

Furthermore,

(v) $(\widetilde{\ell}^{(N)\rho})^{jj'}$

$$= \nu^{jj'} + \nu^{jk}\Big[-\frac{1}{N^{1/2}}\, Z^{0(N)\rho}_{kk'} + \frac{1}{N}\,(-Z^{1(N)\rho}_{kk_1} + Z^{0(N)\rho}_{kk_1}\nu^{k_1k_2}Z^{0(N)\rho}_{k'k_2})\Big]\nu^{j'k'} + \frac{1}{N^{3/2}}\, o^{(N)}_F(1, 3/2)\,.$$

Proof Expanding $\widetilde{\ell}^{(N)\rho}_a$ and $\widetilde{\ell}^{(N)\rho}_{j_1\cdots j_R}$, $R \in \{2, 3, 4\}$, around $\boldsymbol{\eta} = \boldsymbol{\eta}_0$, we can see that

$$\begin{aligned}
N^{1/2}\widetilde{\ell}^{(N)\rho}_a &= Z^{(N)}_a + \Big(\nu_{ar_2} + \frac{1}{N^{1/2}}\, Z^{(N)}_{ar_2}\Big)\eta^{(N)\rho}_{r_2} \\
&\quad + \frac{1}{2N^{1/2}}\Big({\nu^{\rho}}_{ar_2r_3} + \frac{1}{N^{1/2}}\, Z^{(N)\rho}_{ar_2r_3}\Big)\prod_{i=2}^{3}\eta^{(N)\rho}_{r_i} \\
&\quad + \frac{1}{6N}\,{\nu^{\rho}}_{ar_2r_3r_4}\prod_{i=2}^{4}\eta^{(N)\rho}_{r_i} + \frac{1}{N^{3/2}}\, o^{(N)}_F(1, 2)\,, \\
\widetilde{\ell}^{(N)\rho}_{jj'} &= \nu_{jj'} + \frac{1}{N^{1/2}}\, Z^{(N)}_{jj'} + \frac{1}{N^{1/2}}\Big({\nu^{\rho}}_{jj'r_3}\eta^{(N)\rho}_{r_3} + \frac{1}{N^{1/2}}\, Z^{(N)\rho}_{jj'r_3}\eta^{(N)\rho}_{r_3}\Big) \\
&\quad + \frac{1}{2N}\,{\nu^{\rho}}_{jj'r_3r_4}\prod_{i=3}^{4}\eta^{(N)\rho}_{r_i} + \frac{1}{N^{3/2}}\, o^{(N)}_F(1, 3/2)\,, \\
\widetilde{\ell}^{(N)\rho}_{jj'j''} &= {\nu^{\rho}}_{jj'j''} + \frac{1}{N^{1/2}}\, Z^{(N)\rho}_{jj'j''} + \frac{1}{N^{1/2}}\,{\nu^{\rho}}_{jj'j''r_4}\eta^{(N)\rho}_{r_4} + \frac{1}{N}\, o^{(N)}_F(1, 1)\,, \\
\widetilde{\ell}^{(N)\rho}_{jj'j''j'''} &= {\nu^{\rho}}_{jj'j''j'''} + \frac{1}{N^{1/2}}\, o^{(N)}_F(1, 1/2)\,.
\end{aligned}$$

Using (10.15), (i)–(iv) follow from Lemma 10.3(ii).

It remains to prove (v). For this, with $\mathbf{Z}^{i(N)\rho} = [Z^{i(N)\rho}_{k_1k_2}]_{k_1,k_2\in\{1,\ldots,p+M\}}$, $i \in \{0, 1\}$, we rewrite (ii) in the matrix from of $\widetilde{\mathbf{L}}^{(N)\rho} = \boldsymbol{\nu}(\mathbf{I}_{p+M} + \boldsymbol{\Delta}^{(N)\rho})$, where

$$\boldsymbol{\Delta}^{(N)\rho} = \frac{1}{N^{1/2}}\,\boldsymbol{\nu}^{-1}\mathbf{Z}^{0(N)\rho} + \frac{1}{N}\,\boldsymbol{\nu}^{-1}\mathbf{Z}^{1(N)\rho} + \frac{1}{N^{3/2}}\, o_F(1, 3/2)\,.$$

Using $\{\mathbf{I}_{p+M} - \boldsymbol{\Delta}^{(N)\rho} + (\boldsymbol{\Delta}^{(N)\rho})^2\}(\mathbf{I}_{p+M} + \boldsymbol{\Delta}^{(N)\rho}) = \mathbf{I}_{p+M} + (\boldsymbol{\Delta}^{(N)\rho})^3$, together with Lemma 10.3(ii), we have

$$\begin{aligned}
(\widetilde{\mathbf{L}}^{(N)\rho})^{-1} &= \{\mathbf{I}_{p+M} - \boldsymbol{\Delta}^{(N)\rho} + (\boldsymbol{\Delta}^{(N)\rho})^2\}\boldsymbol{\nu}^{-1} - (\boldsymbol{\Delta}^{(N)\rho})^3(\widetilde{\mathbf{L}}^{(N)\rho})^{-1} \\
&= \boldsymbol{\nu}^{-1} + \boldsymbol{\nu}^{-1}\Big[-\frac{1}{N^{1/2}}\,\mathbf{Z}^{0(N)\rho} + \frac{1}{N}\,(-\mathbf{Z}^{1(N)\rho} + \mathbf{Z}^{0(N)\rho}\boldsymbol{\nu}^{-1}\mathbf{Z}^{0(N)\rho})\Big]\boldsymbol{\nu}^{-1} \\
&\quad + \frac{1}{N^{3/2}}\, o^{(N)}_F(1, 3/2)\,.
\end{aligned}$$

□

10.6.2 Proof of Lemma 10.1

The stochastic expansions of $\mathrm{W}^{(N)\rho}$ and $\mathrm{grad}^{(N)\rho}$ follow from (10.17) and (10.18). Also, expanding $-2N\widehat{\ell}^{(N)\rho}$ around $\boldsymbol{\eta} = \widetilde{\boldsymbol{\eta}}^{(N)\rho}$, we have

$$\begin{aligned}
&\mathrm{ELR}^{(N)\rho}\\
&= -2N^{1/2}\widetilde{\ell}_a^{(N)\rho}\delta_a^{(N)\rho} - \widetilde{\ell}_{j_1j_2}^{(N)\rho}\prod_{i=1}^{2}\delta_{j_i}^{(N)\rho} - \frac{1}{3N^{1/2}}\widetilde{\ell}_{j_1j_2j_3}^{(N)\rho}\prod_{i=1}^{3}\delta_{j_i}^{(N)\rho}\\
&\quad -\frac{1}{12N}\widetilde{\ell}_{j_1j_2j_3j_4}^{(N)\rho}\prod_{i=1}^{4}\delta_{j_i}^{(N)\rho} + \frac{1}{N^{3/2}}o_F^{(N)}(1,5/2)\\
&= \widetilde{\ell}_{j_1j_2}^{(N)\rho}\delta_{j_1}^{0(N)\rho}\delta_{j_2}^{0(N)\rho} - \frac{1}{3N^{1/2}}\,{}_{\rho}\widetilde{\ell}_{j_1j_2j_3}^{(N)\rho}\prod_{i=1}^{3}\delta_{j_i}^{0(N)\rho}\\
&\quad +\frac{1}{N}\left\{\frac{1}{4}\widetilde{\ell}_{j_1j_2j}^{(N)\rho}(\widetilde{\ell}^{(N)\rho})^{jj'}\widetilde{\ell}_{j_3j_4j'}^{(N)\rho} - \frac{1}{12}\widetilde{\ell}_{j_1j_2j_3j_4}^{(N)\rho}\right\}\prod_{i=1}^{4}\delta_{j_i}^{0(N)\rho} + \frac{1}{N^{3/2}}o_F^{(N)}(1,5/2)\,,
\end{aligned}$$

using (10.17), (10.18), and $\widetilde{\mathbf{L}}_{(11\cdot 2)}^{(N)\rho} = (\widetilde{\mathcal{G}}^{(N)\rho})^{\top}\widetilde{\mathbf{L}}^{(N)\rho}\widetilde{\mathcal{G}}^{(N)\rho}$.

It remains to prove the stochastic expansions of $\mathrm{S}_{\dagger}^{(N)\rho}$ and $\mathrm{W}_{\dagger}^{(N)\rho}$. Expanding $\widehat{\ell}_{jj'}^{(N)\rho}$ around $\boldsymbol{\eta} = \widetilde{\boldsymbol{\eta}}^{(N)\rho}$, we have

$$\begin{aligned}
&\widehat{\ell}_{jj'}^{(N)\rho}\\
&= \widetilde{\ell}_{jj'}^{(N)\rho} + \frac{1}{N^{1/2}}\widetilde{\ell}_{jj'j_3}^{(N)\rho}\delta_{j_3}^{(N)\rho} + \frac{1}{2N}\widetilde{\ell}_{jj'j_3j_4}^{(N)\rho}\prod_{i=3}^{4}\delta_{j_i}^{(N)\rho} + \frac{1}{N^{3/2}}o_F^{(N)}(1,3/2)\\
&= \widetilde{\ell}_{jj'}^{(N)\rho} + \frac{1}{N^{1/2}}\widetilde{\ell}_{jj'j_3}^{(N)\rho}\delta_{j_3}^{0(N)\rho} + \frac{1}{2N}\{-\widetilde{\ell}_{kjj'}^{(N)\rho}(\widetilde{\ell}^{(N)\rho})^{kk'}\widetilde{\ell}_{k'j_3j_4}^{(N)\rho} + \widetilde{\ell}_{jj'j_3j_4}^{(N)\rho}\}\prod_{i=3}^{4}\delta_{j_i}^{0(N)\rho}\\
&\quad +\frac{1}{N^{3/2}}o_F^{(N)}(1,3/2)\,,
\end{aligned}$$

using (10.17) and (10.18). Then, we have $\widehat{\mathbf{L}}^{(N)\rho} = \widetilde{\mathbf{L}}^{(N)\rho}(\mathbf{I}_{p+M} + \mathbf{D}^{(N)\rho})$, where the (j, j')th element of $\mathbf{D}^{(N)\rho}$ is given by

$$\begin{aligned}
D_{jj'}^{(N)\rho} &= \frac{1}{N^{1/2}}(\widetilde{\ell}^{(N)\rho})^{jk''}\widetilde{\ell}_{k''j'j_3}^{(N)\rho}\delta_{j_3}^{0(N)\rho}\\
&\quad +\frac{1}{2N}(\widetilde{\ell}^{(N)\rho})^{jk''}\{-\widetilde{\ell}_{kk''j'}^{(N)\rho}(\widetilde{\ell}^{(N)\rho})^{kk'}\widetilde{\ell}_{k'j_3j_4}^{(N)\rho} + \widetilde{\ell}_{k''j'j_3j_4}^{(N)\rho}\}\prod_{i=3}^{4}\delta_{j_i}^{0(N)\rho}\\
&\quad +\frac{1}{N^{3/2}}o_F^{(N)}(1,3/2)\,.
\end{aligned}$$

Using $\{\mathbf{I}_{p+M} - \mathbf{D}^{(N)\rho} + (\mathbf{D}^{(N)\rho})^2\}(\mathbf{I}_{p+M} + \mathbf{D}^{(N)\rho}) = \mathbf{I}_{p+M} + (\mathbf{D}^{(N)\rho})^3$, we have

$$(\widehat{\mathbf{L}}^{(N)\rho})^{-1} = \{\mathbf{I}_{p+M} - \mathbf{D}^{(N)\rho} + (\mathbf{D}^{(N)\rho})^2\}(\widetilde{\mathbf{L}}^{(N)\rho})^{-1} - (\mathbf{D}^{(N)\rho})^3(\widehat{\mathbf{L}}^{(N)\rho})^{-1} \,,$$

hence,

$$\begin{aligned}
(\widehat{\ell}^{(N)\rho}_{(11\cdot 2)})^{ab} &= \text{the } (a,b)\text{th element of } \left[(\mathbf{I}_{p_1} \;\; \mathbf{O}_{p_1,p_2+M})(\widehat{\mathbf{L}}^{(N)\rho})^{-1}\begin{pmatrix}\mathbf{I}_{p_1}\\ \mathbf{O}_{p_2+M,p_1}\end{pmatrix}\right]^{-1} \\
&= (\widetilde{\ell}^{(N)\rho}_{(11\cdot 2)})^{ab} - \frac{1}{N^{1/2}}\,\widetilde{\ell}^{(N)\rho}_{j_1j_2j_3}\widetilde{\mathcal{G}}^{(N)\rho}_{j_1a'}\widetilde{\mathcal{G}}^{(N)\rho}_{j_2b'}(\widetilde{\ell}^{(N)\rho}_{(11\cdot 2)})^{a'a}(\widetilde{\ell}^{(N)\rho}_{(11\cdot 2)})^{b'b}\delta^{0(N)\rho}_{j_3} \\
&\quad + \frac{1}{N}\Big[\frac{1}{2}\,\{\widetilde{\ell}^{(N)\rho}_{kj_1j_2}(\widetilde{\ell}^{(N)\rho})^{kk'}\widetilde{\ell}^{(N)\rho}_{k'j_3j_4} - \widetilde{\ell}^{(N)\rho}_{j_1j_2j_3j_4}\} + \widetilde{\ell}^{(N)\rho}_{kj_1j_3}(\widetilde{\ell}^{(N)\rho})^{kk'}\widetilde{\ell}^{(N)\rho}_{k'j_2j_4}\Big] \\
&\quad\quad \widetilde{\mathcal{G}}^{(N)\rho}_{j_1a'}\widetilde{\mathcal{G}}^{(N)\rho}_{j_2b'}(\widetilde{\ell}^{(N)\rho}_{(11\cdot 2)})^{a'a}(\widetilde{\ell}^{(N)\rho}_{(11\cdot 2)})^{b'b}\prod_{i=3}^{4}\delta^{0(N)\rho}_{j_i} \\
&\quad + \frac{1}{N^{3/2}}\,o^{(N)}_F(1,3/2)\,.
\end{aligned}$$

Furthermore, we obtain

$$\begin{aligned}
&\widehat{\ell}^{(N)\rho}_{(11\cdot 2)ab} \\
&= \widetilde{\ell}^{(N)\rho}_{(11\cdot 2)ab} + \frac{1}{N^{1/2}}\,\widetilde{\ell}^{(N)\rho}_{j_1j_2j_3}\widetilde{\mathcal{G}}^{(N)\rho}_{j_1a}\widetilde{\mathcal{G}}^{(N)\rho}_{j_2b}\delta^{0(N)\rho}_{j_3} \\
&\quad + \frac{1}{N}\Big[\frac{1}{2}\,\{-\widetilde{\ell}^{(N)\rho}_{kj_1j_2}(\widetilde{\ell}^{(N)\rho})^{kk'}\widetilde{\ell}^{(N)\rho}_{k'j_3j_4} + \widetilde{\ell}^{(N)\rho}_{j_1j_2j_3j_4}\} - \widetilde{\ell}^{(N)\rho}_{kj_1j_3}(\widetilde{\ell}^{(N)\rho})^{kk'}\widetilde{\ell}^{(N)\rho}_{k'j_2j_4}\Big] \\
&\quad\quad \widetilde{\mathcal{G}}^{(N)\rho}_{j_1a}\widetilde{\mathcal{G}}^{(N)\rho}_{j_2b}\prod_{i=3}^{4}\delta^{0(N)\rho}_{j_i} \\
&\quad + \frac{1}{N}\,(\widetilde{\ell}^{(N)\rho}_{j_1j_2j_3}\widetilde{\mathcal{G}}^{(N)\rho}_{j_2a}\delta^{0(N)\rho}_{j_3})\widetilde{\mathcal{G}}^{(N)\rho}_{j_1b'}(\widetilde{\ell}^{(N)\rho}_{(11\cdot 2)})^{b'b''}\widetilde{\mathcal{G}}^{(N)\rho}_{k_1b''}(\widetilde{\ell}^{(N)\rho}_{k_1k_2k_3}\widetilde{\mathcal{G}}^{(N)\rho}_{k_2b}\delta^{0(N)\rho}_{k_3}) \\
&\quad + \frac{1}{N^{3/2}}\,o^{(N)}_F(1,3/2)\,.
\end{aligned}$$

It follows that

$$\begin{aligned}
\mathrm{S}^{(N)\rho}_{\dagger} &= \widetilde{Z}^{(N)\rho}_a(\widetilde{\ell}^{(N)\rho}_{(11\cdot 2)})^{ab}\widetilde{Z}^{(N)\rho}_b - \frac{1}{N^{1/2}}\,\widetilde{\ell}^{(N)\rho}_{j_1j_2j_3}\prod_{i=1}^{3}\delta^{0(N)\rho}_{j_i} \\
&\quad + \frac{1}{N}\left\{\frac{3}{2}\,\widetilde{\ell}^{(N)\rho}_{jj_1j_2}(\widetilde{\ell}^{(N)\rho})^{jj'}\widetilde{\ell}^{(N)\rho}_{j'j_3j_4} - \frac{1}{2}\,\widetilde{\ell}^{(N)\rho}_{j_1j_2j_3j_4}\right\}\prod_{i=1}^{4}\delta^{0(N)\rho}_{j_i} \\
&\quad + \frac{1}{N^{3/2}}\,o^{(N)}_F(1,5/2)\,, \\
\mathrm{W}^{(N)\rho}_{\dagger} &= \mathrm{W}^{(N)\rho}
\end{aligned}$$

$$+\frac{1}{N^{1/2}}\,\widetilde{\ell}^{(N)\rho}_{j_1j_2j_3}\Big(\prod_{i=1}^{2}\delta^{(N)\rho}_{j_i}\Big)\delta^{0(N)\rho}_{j_3}+\frac{1}{N}\Big[\widetilde{\ell}^{(N)\rho}_{jj_1j_2}\Big\{\widetilde{\mathcal{G}}^{(N)\rho}_{ja}(\widetilde{\ell}^{(N)\rho}_{(11\cdot2)})^{aa'}\widetilde{\mathcal{G}}^{(N)\rho}_{j'a'}$$
$$-\frac{3}{2}\,(\widetilde{\ell}^{(N)\rho})^{jj'}\Big\}\widetilde{\ell}^{(N)\rho}_{j'j_3j_4}+\frac{1}{2}\,\widetilde{\ell}^{(N)\rho}_{j_1j_2j_3j_4}\Big]\prod_{i=1}^{4}\delta^{0(N)\rho}_{j_i}+\frac{1}{N^{3/2}}\,o^{(N)}_F(1,5/2)\,,$$

using (10.17) and (10.18). □

10.6.3 Asymptotic Expansion for the Distribution of $T^{(N)\rho,\tau}$

A rigorous derivation of a χ^2 type asymptotic expansion for the distribution of $T^{(N)\rho,\tau}$ basically consists of the following two steps:

- the square-root decomposition of $T^{(N)\rho,\tau}$ (Proposition 10.1);
- for $x > 0$, the integration of the N^{-1}-Edgeworth expansion over the convex region $\{\mathbf{u}=(u_1,\ldots,u_{p_1})^\top : \mathbf{u}^\top \boldsymbol{\nu}^{-1}_{(11\cdot2)}\mathbf{u}\le x\}$ (Proposition 10.2).

Recall the notation $G_\nu(x)=\int_0^x g_\nu(y)\,dy$ for $x\ge 0$, where $g_\nu(\cdot)$ is the density function of the χ^2_ν random variable. Note that $2g_{\nu+2}(x)=G_\nu(x)-G_{\nu+2}(x)$.

Proposition 10.1 *Suppose that assumptions* (C$_0$)–(C$_3$) *hold with* $Q\ge 16$. *For any constant* C*, independent of* N*, we have*

$$\Big(1+\frac{C}{N}\Big)T^{(N)\rho,\tau}=U^{C(N)\rho,\tau}_a\,\nu^{ab}_{(11\cdot2)}U^{C(N)\rho,\tau}_b+\frac{1}{N^{3/2}}\,o^{(N)}_F(1,\max(5/2,q))\,,$$

with

$$U^{C(N)\rho,\tau}_{b_0}=\Big(1+\frac{C}{2N}\Big)[\boldsymbol{\mathcal{G}}^\top\mathbf{Z}^{(N)}]_{b_0}+\frac{1}{N^{1/2}}\,U^{1(N)\rho,\tau}_{b_0}+\frac{1}{N}\,U^{2(N)\rho,\tau}_{b_0}\,,$$

where

$$U^{1(N)\rho,\tau}_{b_0}=-\mathcal{G}_{jb_0}Z^{(N)}_{jr_2}\nu^{r_2r_2'}_{(22)}Z^{(N)}_{r_2'}+\frac{1}{2}\,\nu^{\rho\mathcal{G}}_{b_0r_2r_3}\prod_{i=2}^{3}\nu^{r_ir_i'}_{(22)}Z^{(N)}_{r_i'}$$
$$+\frac{1}{2}\,\tau_1\nu^{\rho\mathcal{G}\,\mathcal{G}\,\mathcal{G}}_{a_1a_2b_0}\prod_{i=1}^{2}[\boldsymbol{\nu}^{-1}_{(11\cdot2)}\boldsymbol{\mathcal{G}}^\top\mathbf{Z}^{(N)}]_{a_i}$$
$$-\frac{1}{2}\,(\mathcal{G}_{ja_1}\mathcal{G}_{j'b_0}Z^{(N)}_{jj'}-\nu^{\rho\mathcal{G}\,\mathcal{G}}_{a_1b_0r_3}\nu^{r_3r_3'}_{(22)}Z^{(N)}_{r_3'})[\boldsymbol{\nu}^{-1}_{(11\cdot2)}\boldsymbol{\mathcal{G}}^\top\mathbf{Z}^{(N)}]_{a_1}\,,$$

and $U_{b_0}^{2(N)\rho,\tau}$ *is a monomial of degree* 3 *in the* $Z_{\#}^{(N)\rho}$*'s (the lengthy expression is omitted, to save space).*

Proof Use Lemma 10.6 repeatedly, together with Lemma 10.3(ii). □

Proposition 10.2 *Suppose that assumptions* (C_0)–(C_4) *hold with* $Q \geq 16$. *For any constant* C*, independent of* N*, we have*

$$P_F^{(N)}\Big[\Big(1+\frac{C}{N}\Big)T^{(N)\rho,\tau} \leq x\Big]$$
$$= G_{p_1}(x) - \frac{2}{N}\Big\{\pi_1^{C,\rho,\tau} g_{p_1+2}(x) + \sum_{\ell=2}^{3} \pi_\ell^{\rho,\tau} g_{p_1+2\ell}(x)\Big\} + o(N^{-1})\,, \quad (10.19)$$

where $\pi_1^{C,\rho,\tau} = \pi_1^{\rho,\tau} + \frac{1}{2}\,Cp_1$,

$$\begin{aligned}
\pi_1^{\rho,\tau} &= \frac{1}{2}\,(\kappa_{b_1,b_2}^{\rho_3,\rho_4,\tau_1,\tau_2,\tau_3,\tau_4} + \kappa_{b_1}^{\rho_3,\tau_1}\kappa_{b_2}^{\rho_3,\tau_1})\nu_{(11\cdot 2)}^{b_1b_2} \\
&\quad -\frac{1}{2}\Big(\frac{1}{4}\,\kappa_{b_1,b_2,b_3,b_4}^{\rho_3,\rho_4,\tau_1,\tau_2,\tau_3,\tau_4} + \kappa_{b_1}^{\rho_3,\tau_1}\kappa_{b_2,b_3,b_4}^{\rho_3,\tau_1}\Big)\nu_{(11\cdot 2)}^{b_1b_2}\nu_{(11\cdot 2)}^{b_3b_4} \\
&\quad +\frac{1}{72}\,\kappa_{b_1,b_2,b_3}^{\rho_3,\tau_1}\kappa_{b_4,b_5,b_6}^{\rho_3,\tau_1}\langle 15\rangle\nu_{(11\cdot 2)}^{b_1b_2}\nu_{(11\cdot 2)}^{b_3b_4}\nu_{(11\cdot 2)}^{b_5b_6}\,, \\
\pi_2^{\rho,\tau} &= \frac{1}{2}\Big(\frac{1}{4}\,\kappa_{b_1,b_2,b_3,b_4}^{\rho_3,\rho_4,\tau_1,\tau_2,\tau_3,\tau_4} + \kappa_{b_1}^{\rho_3,\tau_1}\kappa_{b_2,b_3,b_4}^{\rho_3,\tau_1}\Big)\nu_{(11\cdot 2)}^{b_1b_2}\nu_{(11\cdot 2)}^{b_3b_4} \\
&\quad -\frac{1}{36}\,\kappa_{b_1,b_2,b_3}^{\rho_3,\tau_1}\kappa_{b_4,b_5,b_6}^{\rho_3,\tau_1}\langle 15\rangle\nu_{(11\cdot 2)}^{b_1b_2}\nu_{(11\cdot 2)}^{b_3b_4}\nu_{(11\cdot 2)}^{b_5b_6}\,, \\
\pi_3^{\rho,\tau} &= \frac{1}{72}\,\kappa_{b_1,b_2,b_3}^{\rho_3,\tau_1}\kappa_{b_4,b_5,b_6}^{\rho_3,\tau_1}\langle 15\rangle\nu_{(11\cdot 2)}^{b_1b_2}\nu_{(11\cdot 2)}^{b_3b_4}\nu_{(11\cdot 2)}^{b_5b_6}\,.
\end{aligned}$$

Here, the $\kappa_{a_1,\ldots,a_R}^{\#}$*'s are some constants*[20] *(independent of* N*), associated with the Rth cumulants* $cum_{a_1,\ldots,a_R}^{C(N)\rho,\tau} = Cum_F^{(N)}(U_{a_1}^{C(N)\rho,\tau},\ldots,U_{a_R}^{C(N)\rho,\tau})$, $R \in \{1,2,3,4\}$*, i.e.,*

[20] Some straightforward but tedious calculations show that

$$\begin{aligned}
\kappa_{a_1}^{\rho_3,\tau_1} &= -\nu_{a_1 r,r'}^{\mathcal{G}}\nu_{(22)}^{rr'} + \frac{1}{2}\,\nu^{\rho\mathcal{G}}{}_{a_1 r_1 r_2}\nu_{(22)}^{r_1r_1'}\nu_{r_1',r_2'}\nu_{(22)}^{r_2r_2'} \\
&\quad +\frac{1}{2}\Big(\tau_1\nu^{\rho\mathcal{G}}{}_{a_1 b}^{\mathcal{G}}{}_{b'}^{\mathcal{G}} - \nu_{a_1 b,b'}^{\mathcal{G}\,\mathcal{G}\,\mathcal{G}} - \sum_{\beta'=1}^{M}\mathcal{G}_{[\beta']b'}\nu^{\rho\mathcal{G}}{}_{a_1 b[\beta']}^{\mathcal{G}}\Big)\nu_{(11\cdot 2)}^{bb'}\,,
\end{aligned}$$

and that $\kappa_{a_1,a_2,a_3}^{\rho_3,\tau_1}$ and $\kappa_{a_1,a_2,a_3,a_4}^{-2,\rho_4,1/3,\tau_2,\tau_3,\tau_4}$ are explicitly given in the proof of Theorem 10.1; the lengthy general expression for $\kappa_{a_1,a_2,a_3,a_4}^{\rho_3,\rho_4,\tau_1,\tau_2,\tau_3,\tau_4}$ is, however, omitted here. Although, in principle, we can write down $\kappa_{a_1,a_2}^{\rho_3,\rho_4,\tau_1,\tau_2,\tau_3,\tau_4}$, we have not rearranged it, due to the rather lengthy algebra (the task would be no practical importance, in many cases).

$$cum_{a_1}^{C(N)\rho,\tau} = \frac{1}{N^{1/2}}\,\kappa_{a_1}^{\rho_3,\tau_1} + o(N^{-1})\,,$$
$$cum_{a_1,a_2}^{C(N)\rho,\tau} = \nu_{(11\cdot2)a_1a_2} + \frac{1}{N}\,(\kappa_{a_1,a_2}^{\rho_3,\rho_4,\tau_1,\tau_2,\tau_3,\tau_4} + C\nu_{(11\cdot2)a_1a_2}) + o(N^{-1})\,,$$
$$cum_{a_1,a_2,a_3}^{C(N)\rho,\tau} = \frac{1}{N^{1/2}}\,\kappa_{a_1,a_2,a_3}^{\rho_3,\tau_1} + o(N^{-1})\,,$$
$$cum_{a_1,a_2,a_3,a_4}^{C(N)\rho,\tau} = \frac{1}{N}\,\kappa_{a_1,a_2,a_3,a_4}^{\rho_3,\rho_4,\tau_1,\tau_2,\tau_3,\tau_4} + o(N^{-1})\,.$$

Proof We use the notation $\phi_{\nu_{(11\cdot2)}}^{b_1,\ldots,b_R}(\mathbf{u}) = (-1)^R(\partial/\partial u_{b_1})\cdots(\partial/\partial u_{b_R})\phi_{\nu_{(11\cdot2)}}(\mathbf{u})$ for the density $\phi_{\nu_{(11\cdot2)}}(\mathbf{u})$ of p_1-variate normal distribution $\mathrm{N}(\mathbf{0}_{p_1}, \boldsymbol{\nu}_{(11\cdot2)})$. As we mentioned in Sect. 10.2.3, the $Z_{\#}^{(N)}$'s are expressed as a linear combination of the elements of $\boldsymbol{\psi}_{1:3}^{(N)} = N^{-1/2}\sum_{i=1}^{N}\{\dot{\mathbf{G}}_{1:3}(\mathbf{X}_i,\boldsymbol{\theta}_0) - E_F[\dot{\mathbf{G}}_{1:3}(\mathbf{X},\boldsymbol{\theta}_0)]\}$, so that $\mathbf{U}^{C(N)\rho,\tau} = [U_a^{C(N)\rho,\tau}]_{a\in\{1,\ldots,p_1\}}$ (see Proposition 10.1) admits the N^{-1}-Edgeworth expansion (e.g., Bhattacharya and Ghosh [6]), i.e.,

$$\begin{aligned}
&\sup_{B\in C_{p_1}}\Bigg|P_F^{(N)}[\mathbf{U}^{C(N)\rho,\tau}\in B]\\
&\quad-\int_B\Big[\phi_{\nu_{(11\cdot2)}}(\mathbf{u})\\
&\qquad+\frac{1}{N^{1/2}}\Big\{\kappa_{b_1}^{\rho_3,\tau_1}\phi_{\nu_{(11\cdot2)}}^{b_1}(\mathbf{u}) + \frac{1}{6}\,\kappa_{b_1,b_2,b_3}^{\rho_3,\tau_1}\phi_{\nu_{(11\cdot2)}}^{b_1,b_2,b_3}(\mathbf{u})\Big\}\\
&\qquad+\frac{1}{N}\Big\{\frac{1}{2}\,(C\nu_{(11\cdot2)b_1b_2} + \kappa_{b_1,b_2}^{\rho_3,\rho_4,\tau_1,\tau_2,\tau_3,\tau_4} + \kappa_{b_1}^{\rho_3,\tau_1}\kappa_{b_2}^{\rho_3,\tau_1})\phi_{\nu_{(11\cdot2)}}^{b_1,b_2}(\mathbf{u})\\
&\qquad\quad+\Big(\frac{1}{24}\,\kappa_{b_1,b_2,b_3,b_4}^{\rho_3,\rho_4,\tau_1,\tau_2,\tau_3,\tau_4} + \frac{1}{6}\,\kappa_{b_1}^{\rho_3,\tau_1}\kappa_{b_2,b_3,b_4}^{\rho_3,\tau_1}\Big)\phi_{\nu_{(11\cdot2)}}^{b_1,b_2,b_3,b_4}(\mathbf{u})\\
&\qquad\quad+\frac{1}{72}\,\kappa_{b_1,b_2,b_3}^{\rho_3,\tau_1}\kappa_{b_4,b_5,b_6}^{\rho_3,\tau_1}\phi_{\nu_{(11\cdot2)}}^{b_1,b_2,b_3,b_4,b_5,b_6}(\mathbf{u})\Big\}\Big]d\mathbf{u}\Bigg|\\
&= o(N^{-1})\,,
\end{aligned}$$

where C_{p_1} is the set of all Borel measurable convex subsets of $\mathbb{R}^{p_1}$. Thus, we have

$$\begin{aligned}
&P_F^{(N)}[(\mathbf{U}^{C(N)\rho,\tau})^\top\boldsymbol{\nu}_{(11\cdot2)}^{-1}\mathbf{U}^{C(N)\rho,\tau}\le x]\\
&= G_{p_1}(x)\\
&\quad+\int_{\mathbf{u}^\top\nu_{(11\cdot2)}^{-1}\mathbf{u}\le x}\Big[\frac{1}{N^{1/2}}\Big\{\kappa_{b_1}^{\rho_3,\tau_1}\phi_{\nu_{(11\cdot2)}}^{b_1}(\mathbf{u}) + \frac{1}{6}\,\kappa_{b_1,b_2,b_3}^{\rho_3,\tau_1}\phi_{\nu_{(11\cdot2)}}^{b_1,b_2,b_3}(\mathbf{u})\Big\}\\
&\qquad+\frac{1}{N}\Big\{\frac{1}{2}\,(C\nu_{(11\cdot2)b_1b_2} + \kappa_{b_1,b_2}^{\rho_3,\rho_4,\tau_1,\tau_2,\tau_3,\tau_4} + \kappa_{b_1}^{\rho_3,\tau_1}\kappa_{b_2}^{\rho_3,\tau_1})\phi_{\nu_{(11\cdot2)}}^{b_1,b_2}(\mathbf{u})\\
&\qquad\quad+\Big(\frac{1}{24}\,\kappa_{b_1,b_2,b_3,b_4}^{\rho_3,\rho_4,\tau_1,\tau_2,\tau_3,\tau_4} + \frac{1}{6}\,\kappa_{b_1}^{\rho_3,\tau_1}\kappa_{b_2,b_3,b_4}^{\rho_3,\tau_1}\Big)\phi_{\nu_{(11\cdot2)}}^{b_1,b_2,b_3,b_4}(\mathbf{u})\\
&\qquad\quad+\frac{1}{72}\,\kappa_{b_1,b_2,b_3}^{\rho_3,\tau_1}\kappa_{b_4,b_5,b_6}^{\rho_3,\tau_1}\phi_{\nu_{(11\cdot2)}}^{b_1,b_2,b_3,b_4,b_5,b_6}(\mathbf{u})\Big\}\Big]d\mathbf{u}
\end{aligned}$$

$$
\begin{aligned}
&+o(N^{-1}) \\
&= G_{p_1}(x) - \frac{2}{N}\Big[\frac{1}{2}\{Cp_1 + (\kappa_{b_1,b_2}^{\rho_3,\rho_4,\tau_1,\tau_2,\tau_3,\tau_4} + \kappa_{b_1}^{\rho_3,\tau_1}\kappa_{b_2}^{\rho_3,\tau_1})\nu_{(11\cdot 2)}^{b_1b_2}\}g_{p_1+2}(x) \\
&+\frac{1}{2}\Big(\frac{1}{4}\kappa_{b_1,b_2,b_3,b_4}^{\rho_3,\rho_4,\tau_1,\tau_2,\tau_3,\tau_4} + \kappa_{b_1}^{\rho_3,\tau_1}\kappa_{b_2,b_3,b_4}^{\rho_3,\tau_1}\Big)\nu_{(11\cdot 2)}^{b_1b_2}\nu_{(11\cdot 2)}^{b_3b_4}\{-g_{p_1+2}(x) + g_{p_1+4}(x)\} \\
&+\frac{1}{72}\kappa_{b_1,b_2,b_3}^{\rho_3,\tau_1}\kappa_{b_4,b_5,b_6}^{\rho_3,\tau_1}\langle 15\rangle \nu_{(11\cdot 2)}^{b_1b_2}\nu_{(11\cdot 2)}^{b_3b_4}\nu_{(11\cdot 2)}^{b_5b_6}\{g_{p_1+2}(x) - 2g_{p_1+4}(x) + g_{p_1+6}(x)\}\Big] \\
&+o(N^{-1}).
\end{aligned}
$$

Then, (10.19) is concluded from Magdalinos (1992; Theorem 2 and Lemma 3) and Proposition 10.1. □

The following is a slight modification of Proposition 10.2 (the proof is omitted; see, e.g., Kakizawa [45] for the related asymptotic expansions).

Proposition 10.3 *Suppose that assumptions* (C_0)–(C_4) *hold with* $Q \geq 16$. *For any symmetric* R*-way arrays* $\mathbf{\Gamma}_R = [\Gamma_{a_1\cdots a_R}]_{a_1,\ldots,a_R\in\{1,\ldots,p_1\}}$, $R \in \{2, 4, 6\}$, *we consider*

$$
T^{\Gamma_{2,4,6}(N)\rho,\tau} = T^{(N)\rho,\tau} + \frac{1}{N}\sum_{R\in\{2,4,6\}} \Gamma_{a_1\cdots a_R}\prod_{i=1}^{R}[(\widetilde{\mathbf{L}}_{(11\cdot 2)}^{(N)\rho})^{-1}\widetilde{\mathbf{Z}}_{(1)}^{(N)\rho}]_{a_i}.
$$

Then,

$$
P_F^{(N)}[T^{\Gamma_{2,4,6}(N)\rho,\tau} \leq x] = G_{p_1}(x) - \frac{2}{N}\sum_{\ell=1}^{3}\pi_\ell^{\Gamma_{2\ell},\rho,\tau} g_{p_1+2\ell}(x) + o(N^{-1}),
$$

where

$$
\begin{aligned}
\pi_1^{\Gamma_2,\rho,\tau} &= \pi_1^{\rho,\tau} + \frac{1}{2}\Gamma_{b_1b_2}\nu_{(11\cdot 2)}^{b_1b_2}, \\
\pi_2^{\Gamma_4,\rho,\tau} &= \pi_2^{\rho,\tau} + \frac{3}{2}\Gamma_{b_1b_2b_3b_4}\nu_{(11\cdot 2)}^{b_1b_2}\nu_{(11\cdot 2)}^{b_3b_4}, \\
\pi_3^{\Gamma_6,\rho,\tau} &= \pi_3^{\rho,\tau} + \frac{15}{2}\Gamma_{b_1b_2b_3b_4b_5b_6}\nu_{(11\cdot 2)}^{b_1b_2}\nu_{(11\cdot 2)}^{b_3b_4}\nu_{(11\cdot 2)}^{b_5b_6}.
\end{aligned}
$$

Remark 10.7 (i) If $p_1 > 1$, there are infinitely many choices $\mathbf{\Gamma}_R$, $R \in \{2, 4, 6\}$, such that $\pi_1^{\Gamma_2,\rho,\tau} = \pi_2^{\Gamma_4,\rho,\tau} = \pi_3^{\Gamma_6,\rho,\tau} = 0$, i.e.,

$$
P_F^{(N)}[T^{\Gamma_{2,4,6}(N)\rho,\tau} \leq x] = G_{p_1}(x) + o(N^{-1}).
$$

This idea, referred to as generalized CF (GCF) method, is found in Kakizawa [45].

(ii) Unlike the GCF method, the ideas behind the CM/T methods are to adjust the test statistic $T^{(N)\rho,\tau}$ as $T^{(N)\rho,\tau} + \sum_{i=1}^{2} N^{-i/2}Q_i^{(N)}$, where $Q_1^{(N)}$ and $Q_2^{(N)}$ depend on $(\widetilde{\mathbf{L}}_{(11\cdot 2)}^{(N)\rho})^{-1}\widetilde{\mathbf{Z}}_{(1)}^{(N)\rho}$, in such a way that the $N^{-1/2}$-perturbating term contributes to the skewness correction for the square-root decomposition of $T^{(N)\rho,\tau} + \sum_{i=1}^{2} N^{-i/2} Q_i^{(N)}$ (see the argument given at the top of this subsection); in that case, the resulting asymptotic expansion necessarily corresponds to $\pi_3^{\rho,\tau} = 0$. In this sense, the CM/T methods are regarded as "double correction methods" of the skewness correction and the (G)CF Bartlett-type correction. The details are omitted.

References

1. ANDERSON, T. W. (2003). *An Introduction to Multivariate Statistical Analysis*, 3rd ed. New York: Wiley.
2. BAGGERLY, K.A. (1998). Empirical likelihood as a goodness-of-fit measure. *Biometrika* **85** 535–547.
3. BARNDORFF- NIELSEN, O.E. AND COX, D.R. (1994). *Inference and Asymptotics*. London: Chapman and Hall.
4. BARTLETT, M.S. (1937). Properties of sufficiency and statistical tests. *Proc.Roy.Soc.Lond.* A **160** 268–282.
5. BERA, A.K. AND BILIAS, Y. (2001). Rao's score, Neyman's $C(\alpha)$ and Silvey's LM tests: an essay on historical developments and some new results. *J.Statist.Plann.Inference* **97** 9–44.
6. BHATTACHARYA, R.N. AND GHOSH, J.K. (1978). On the validity of the formal Edgeworth expansion. *Ann.Statist.* **6** 434–451. Correction: (1980). **8** 1399.
7. BHATTACHARYA, R.N. AND RAO, R.R. (1976). *Normal Approximation and Asymptotic Expansions*. New York: Wiley.
8. BOOS, D.D. (1992). On generalized score tests. *Amer.Statist.* **46** 327–333.
9. BRAVO, F. (2002). Testing linear restrictions in linear models with empirical likelihood. *Econom.J.* **5** 104–130.
10. BRAVO, F. (2003). Second-order power comparisons for a class of nonparametric likelihood-based tests. *Biometrika* **90** 881–890.
11. BRAVO, F. (2004). Empirical likelihood based inference with applications to some econometric models. *Econometric Theory* **20** 231–264.
12. BRAVO, F. (2006). Bartlett-type adjustments for empirical discrepancy test statistics. *J.Statist.Plann.Inference* **136** 537–554.
13. CAMPONOVO, L. AND OTSU, T. (2014). On Bartlett correctability of empirical likelihood in generalized power divergence family. *Statist.Probab.Lett.* **86** 38–43.
14. CHANDRA, T.K. AND MUKERJEE, R. (1991). Bartlett-type modification for Rao's efficient score statistic. *J.Multivariate Anal.* **36** 103–112.
15. CHEN, S.X. (1993). On the accuracy of empirical likelihood confidence regions for linear regression model. *Ann.Inst.Statist.Math.* **45** 621–637.
16. CHEN, S.X. (1994). Empirical likelihood confidence intervals for linear regression coefficients. *J.Multivariate Anal.* **49** 24–40.
17. CHEN, S.X. AND CUI, H. (2006). On Bartlett correction of empirical likelihood in the presence of nuisance parameters. *Biometrika* **93** 215–220.
18. CHEN, S.X. AND CUI, H. (2007). On the second-order properties of empirical likelihood with moment restrictions. *J.Econometrics* **141** 492–516.
19. CORCORAN, S.A. (1998). Bartlett adjustment of empirical discrepancy statistics. *Biometrika* **85** 967–972.
20. CORDEIRO, G.M. AND FERRARI, S.L.P. (1991). A modified score test statistic having chi-squared distribution to order n^{-1}. *Biometrika* **78** 573–582.
21. COX, D.R. (1988). Some aspects of conditional and asymptotic inference: a review. *Sankhyā* A **50** 314–337.

22. Cressie, N. and Read, T. (1984). Multinomial goodness-of-fit tests. *J.R.Stat.Soc.Ser.* B **46** 440–464.
23. Cribari- Neto, F. and Cordeiro, G.M. (1996). On Bartlett and Bartlett-type corrections. *Econometric Rev.* **15** 339–367.
24. DiCiccio, T., Hall, P. and Romano, J. (1991). Empirical likelihood is Bartlett-correctable. *Ann.Statist.* **19** 1053–1061.
25. Efron, B. (1979). Bootstrap methods: another look at the jackknife. *Ann.Statist.* **7** 1–26.
26. Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. *Philos.Trans.R.Soc.Lond.Ser.A Math.Phys.Eng.Sci.* **222** 309–368.
27. Fujikoshi, Y. (1988). Comparison of powers of a class of tests for multivariate linear hypothesis and independence. *J.Multivariate Anal.* **26** 48–58.
28. Fujikoshi, Y. (1997). A method for improving the large-sample chi-squared approximations to some multivariate test statistics. *Amer.J.Math.Management Sci.* **17** 15–29.
29. Fujisawa, H. (1997). Improvement on chi-squared approximation by monotone transformation. *J.Multivariate Anal.* **60** 84–89.
30. Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*. New York: Springer.
31. Hall, P. and La Scala, B. (1990). Methodology and algorithms of empirical likelihood. *Internat.Statist.Rev.* **58** 109–127.
32. Hansen, L.P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica* **50** 1029–1054.
33. Hansen, L.P., Heaton, J. and Yaron, A. (1996). Finite-sample properties of some alternative GMM estimators. *J.Bus.Econom.Statist.* **14** 262–280.
34. Harris, P. (1985). An asymptotic expansion for the null distribution of the efficient score statistic. *Biometrika* **72** 653–659.
35. Harris, P. and Peers, H.W. (1980). The local power of the efficient scores test statistic. *Biometrika* **67** 525–529.
36. Hayakawa, T. (1975). The likelihood ratio criterion for a composite hypothesis under a local alternative. *Biometrika* **62** 451–460.
37. Hayakawa, T. (1977). The likelihood ratio criterion and the asymptotic expansion of its distribution. *Ann.Inst.Statist.Math.* **29** 359–378. Correction: (1987). **39** 681.
38. Horn, R.A. and Johnson, C.R. (1985). *Matrix Analysis*. Cambridge: Cambridge University Press.
39. Iwashita, T. (1997). Adjustment of Bartlett type to Hotelling's T^2-statistic under elliptical distributions. *Amer.J.Math.Management Sci.* **17** 91–96.
40. Jensen, J.L. (1993). A historical sketch and some new results on the improved log likelihood ratio statistic. *Scand.J.Statist.* **20** 1–15.
41. Jing, B.Y. and Wood, A.T.A. (1996). Exponential empirical likelihood is not Bartlett correctable. *Ann.Statist.* **24** 365–369.
42. Kakizawa, Y. (1996). Higher order monotone Bartlett-type adjustment for some multivariate test statistics. *Biometrika* **83** 923–927.
43. Kakizawa, Y. (1997). Higher-order Bartlett-type adjustment. *J.Stat.Plan.Inference* **65** 269–280.
44. Kakizawa, Y. (2009). Third-order power comparisons for a class of tests for multivariate linear hypothesis under general distributions. *J.Multivariate Anal.* **100** 473–496.
45. Kakizawa, Y. (2012a). Generalized Cordeiro–Ferrari Bartlett-type adjustment. *Statist.Probab.Lett.* **82** 2008–2016.
46. Kakizawa, Y. (2012b). Improved chi-squared tests for a composite hypothesis. *J.Multivariate Anal.* **107** 141–161.
47. Kakizawa, Y. (2013). Third-order local power properties of tests for a composite hypothesis. *J.Multivariate Anal.* **114** 303–317.
48. Kakizawa, Y. (2015). Third-order local power properties of tests for a composite hypothesis, II. *J.Multivariate Anal.* **140** 99–112.
49. Kakizawa, Y. (2017). Third-order average local powers of Bartlett-type adjusted tests: Ordinary versus adjusted profile likelihood. *J.Multivariate Anal.* **153** 98–120.

50. KITAMURA, Y. AND STUTZER, M. (1997). An information-theoretic alternative to generalized method of moments estimation. *Econometrica* **65** 861–874.
51. KOJIMA, M. AND KUBOKAWA, T. (2013). Bartlett-type adjustments for hypothesis testing in linear models with general error covariance matrices. *J.Multivariate Anal.* **122** 162–174.
52. KOLACZYK, E.D. (1994). Empirical likelihood for generalized linear models. *Statist.Sinica* **4** 199–218.
53. LAWLEY, D.N. (1956). A general method for approximating to the distribution of likelihood ratio criteria. *Biometrika* **43** 295–303.
54. MAGDALINOS, M.A. (1992). Stochastic expansions and asymptotic approximations. *Econometric Theory* **8** 343–367.
55. MUKERJEE, R. (1993). Rao's score test: Recent asymptotic results. In: *Handbook of Statistics, Vol.11*. Maddala, G.S., Rao, C.R. and Vinod, H.D. Eds. Amsterdam: North-Holland, pp.363–379.
56. NEWEY, W.K. AND SMITH, R.J. (2004). Higher order properties of GMM and generalized empirical likelihood estimators. *Econometrica* **72** 219–255.
57. OWEN, A.B. (1988). Empirical likelihood ratio confidence intervals for a single functional. *Biometrika* **75** 237–249.
58. OWEN, A.B. (1990). Empirical likelihood ratio confidence regions. *Ann.Statist.* **18** 90–120.
59. OWEN, A.B. (1991). Empirical likelihood for linear models. *Ann.Statist.* **19** 1725–1747.
60. OWEN, A.B. (2001). *Empirical Likelihood*. London: Chapman and Hall.
61. PEERS, H.W. (1971). Likelihood ratio and associated test criteria. *Biometrika* **58** 577–587.
62. QIN, J. AND LAWLESS, J. (1994). Empirical likelihood and general estimating equations. *Ann.Statist.* **22** 300–325.
63. QIN, J. AND LAWLESS, J. (1995). Estimating equations, empirical likelihood and constraints on parameters. *Canad.J.Statist.* **23** 145–159.
64. RAGUSA, G. (2011). Minimum divergence, generalized empirical likelihoods, and higher order expansions. *Econometric Rev.* **30** 406–456.
65. RAO, C.R. (1948). Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation. *Math.Proc.Cambridge Philos.Soc.* **44** 50–57.
66. RAO, C.R. AND MUKERJEE, R. (1997). Comparison of LR, score, and Wald tests in a non-iid setting. *J.Multivariate Anal.* **60** 99–110.
67. ROCKE, D.M. (1989). Bootstrap Bartlett adjustment in seemingly unrelated regression. *J.Amer.Statist.Assoc.* **84** 598–601.
68. ROTHENBERG, T.J. (1977). Edgeworth expansions for multivariate test statistics. IP–255, Center for Research in Management Science, University of California, Berkeley.
69. SILVEY, S.D. (1959). The Lagrange multiplier test. *Ann.Math.Statist.* **30** 389–407.
70. SIOTANI, M., HAYAKAWA, T. AND FUJIKOSHI, Y. (1985). *Modern Multivariate Analysis: A Graduate Course and Handbook*. Ohio: American Sciences Press.
71. TANIGUCHI, M. (1991a). *Higher Order Asymptotic Theory for Time Series Analysis*. Berlin: Springer-Verlag.
72. TANIGUCHI, M. (1991b). Third-order asymptotic properties of a class of test statistics under a local alternative. *J.Multivariate Anal.* **37** 223–238.
73. TERRELL, G.R. (2002). The gradient statistic. *Comput.Sci.Statist.* **34** 206–215.

74. THOMAS, D.R. AND GRUNKEMEIER, G.L. (1975). Confidence interval estimation of survival probabilities for censored data. *J.Amer.Statist.Assoc.* **70** 865–871.
75. WALD, A. (1943). Tests of statistical hypothesis concerning several parameters when the number of observations is large. *Trans.Amer.Math.Soc.* **54** 426–482.
76. WILKS, S.S. (1938). The large-sample distribution of the likelihood ratio for testing composite hypotheses. *Ann.Math.Statist.* **9** 60–62.
77. ZHANG, B. (1996). On the accuracy of empirical likelihood confidence intervals for M-functionals. *J.Nonparametr.Stat.* **6** 311–321.

Chapter 11
An Analog of the Bickel–Rosenblatt Test for Error Density in the Linear Regression Model

Fuxia Cheng, Hira L. Koul, Nao Mimoto, and Narayana Balakrishna

Abstract This paper addresses the problem of testing the goodness-of-fit hypothesis pertaining to error density in multiple linear regression models with non-random and random predictors. The proposed tests are based on the integrated square difference between a nonparametric density estimator based on the residuals and its expected value under the null hypothesis when all regression parameters are known. We derive the asymptotic distributions of this sequence of test statistics under the null hypothesis and under certain global alternatives. The asymptotic null distribution of a suitably standardized test statistic based on the residuals is the same as in the case of known underlying regression parameters. Under the global L_2 alternatives of [2], the asymptotic distribution of this sequence of statistics is affected by not knowing the parameters and, in general, is different from the one obtained in [2] for the zero intercept linear autoregressive time series context.

11.1 Introduction and Summary

The problem of testing the goodness-of-fit hypothesis pertaining to a density is a classical problem. In the case of the one sample setup where data consists of a random sample from the given density, one of the tests for this problem proposed in [3] is based on the integrated squared difference between the error density estimator and

F. Cheng
Illinois State University, Campus Box 4520, Normal, IL 61790-4520, USA
e-mail: fcheng@ilsu.edu

H. L. Koul (✉)
Michigan State University, East Lansing, MI 48824-1027, USA
e-mail: koul@msu.edu

N. Mimoto
The University of Akron, Akron, OH 44325-1913, USA
e-mail: nmimoto@uakron.edu

N. Balakrishna
Cochin University of Science and Technology, Kochi, KL, India
e-mail: nb@cusat.ac.in

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_11

its expected value under the null hypothesis. Under some conditions its asymptotic null distribution is a known normal distribution.

The papers [2, 4, 8, 9] observed that this fact continues to hold for the analog of this test statistic for fitting an error density based on the residuals in linear and nonlinear autoregressive and generalized autoregressive conditionally heteroscedastic time series models. In other words, the asymptotic null distribution of the analog of this statistic in all these cases is the same as in the case of known underlying parameters, i.e., not knowing the underlying nuisance parameters has no effect on the asymptotic null distribution of this test statistic in these models. In comparison, in general, the asymptotic null distributions of the goodness-of-fit tests based on suitably standardized residual empirical process depend on the estimators of the underlying nuisance parameters in these models in a complicated fashion, which renders them to be unknown. Consequently, in general, the goodness-of-fit tests based on residual empirical process are not implementable even for the large samples. Exception to this is provided by the tests based on Khmaladze's transform of the residual empirical process as described in [6, 7]. However, the implementation of such tests has proved to be difficult in general. For this reason, the above mentioned analog of the Bickel–Rosenblatt test is more appealing.

Relatively little is known in the literature about the asymptotic distributions of the analog of the above mentioned test statistic in regression models. In this paper, we derive the asymptotic distributions, under the null hypothesis and under the global L_2 alternatives in [2], $H_1 : f \neq f_0, \int (f - f_0)^2(y)dy > 0$, of the analogs of this statistic in the multiple linear regression models with non-random and random covariates.

To begin with, we shall focus on the case of non-random covariates. Accordingly, let n and p be known positive integers, $\mathbb{R}^p$ denote the p-dimensional Euclidean space. Especially, $\mathbb{R}$=$\mathbb{R}^1$. For any $x \in \mathbb{R}^p$, let x' denote its transpose. Let $x_i \in \mathbb{R}^p$, $1 \leq i \leq n$ be p-dimensional vectors of known real numbers and X denote the $n \times p$ matrix whose ith row is x_i', $1 \leq i \leq n$. In the linear regression model of interest with non-random predictors, one observes x_i, Y_i obeying the model

$$Y_i = \alpha + \beta' x_i + \varepsilon_i, \qquad \exists\, \alpha \in \mathbb{R},\ \beta \in \mathbb{R}^p, \forall\, 1 \leq i \leq n, \tag{11.1}$$

where ε, ε_i, $i \geq 1$ are independent and identically distributed (i.i.d.) random variables (r.v.'s) according to a Lebesgue density f on $\mathbb{R}$. We shall assume the above model is well identified, which is true if either $E(\varepsilon) = 0$ or ε is symmetrically distributed around zero. For our result neither of these two assumptions are needed. However, we need the availability of certain consistent estimators of α, β, see (11.9). All the results below in the case of non-random covariates are valid even when x_i's depend on n, i.e., when $x_i \equiv x_{ni}$. We do not exhibit this dependence on n for the sake of the brevity of the exposure.

Let f_0 be a known density. The problem of interest is to test

$$H_0 : \ f = f_0, \quad \text{versus} \quad H_1 : \ H_0 \text{ is not true},$$

based on the data (x_1, Y_1), ..., (x_n, Y_n) obeying the above model.

To describe the test statistic for this problem, let K be a density kernel and $h \equiv h_n$ be bandwidth sequence. Let $\widehat{\alpha}$, $\widehat{\beta}$ be estimators of α, β, respectively, and $\widehat{\varepsilon}_i := Y_i - \widehat{\alpha} - \widehat{\beta}' x_i$, $1 \le i \le n$. Define

$$\widehat{f}_n(y) := \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{y - \widehat{\varepsilon}_i}{h}\Big), \qquad f_n(y) := \frac{1}{nh} \sum_{i=1}^{n} K\Big(\frac{y - \varepsilon_i}{h}\Big),$$
$$\mu_n(y) := \frac{1}{h} \int K\Big(\frac{y - z}{h}\Big) f_0(z) dz, \qquad y \in \mathbb{R}, \tag{11.2}$$
$$\widehat{T}_n := \int \big(\widehat{f}_n(y) - \mu_n(y)\big)^2 dy, \qquad T_n := \int (f_n(y) - \mu_n(y))^2 dy.$$

Throughout this paper, $\int$ denotes a definite integral where the domain of the integral is omitted when there is no confusion.

The statistic T_n is the test statistic used in [3] for testing H_0 when ε_i's are observable, i.e., α, β are known in (11.1), while $\widehat{T}_n$ is its analog used for testing H_0 in the regression model (11.1) when α, β are unknown. Its asymptotic null distribution along with the required assumptions is described in the next section. Section 11.3 contains the asymptotic distribution of $\widehat{T}_n$ under the global L_2 alternatives $H_1 : f \ne f_0$, $\int (f - f_0)^2(y) dy > 0$. See [2] for the usefulness of these alternatives. Section 11.4 contains the asymptotic distributions, under H_0 and H_1, of $\widetilde{T}_n$, the analog of $\widehat{T}_n$ in the case of random covariates. The asymptotic null distributions of $\widehat{T}_n$ and $\widetilde{T}_n$ continue to be the same as that of T_n, i.e., as in the case when the parameters α, β are known.

The asymptotic distributions of $\widehat{T}_n$, $\widetilde{T}_n$ under the alternatives H_1 are seen to be affected by not knowing the parameters in general and are different from that obtained in [2] for the zero intercept linear autoregressive case. But if both f_0 and f in the alternatives are symmetric around zero then these asymptotic distributions are the same as in [2], for both non-random and random predictors cases. See Theorems 11.2 and 11.4. The findings of a finite sample simulation and an application to real data are given in Sect. 11.5.

In the sequel, all limits are taken as $n \to \infty$ and $\to_p$ ($\to_D$) denotes the convergence in probability (distribution).

11.2 Asymptotic Null Distribution of $\widehat{T}_n$

Before describing the asymptotic null distribution of $\widehat{T}_n$ in the current setup, we recall the following asymptotic normality results of T_n. Consider the following assumptions.

$$f,\ f_0 \text{ are square integrable and twice continuously differentiable densities with the second derivatives bounded and both having zero mean.} \tag{11.3}$$

$$K \text{ is continuous bounded symmetric kernel density with compact support and } \int v^2 K(v)dv < \infty. \tag{11.4}$$

$$h \to 0 \text{ and } nh^2 \to \infty. \tag{11.5}$$

For any two integrable functions $g_1,\ g_2$, let $g_1 * g_2(x) := \int g_1(x-t)g_2(t)dt$ and $K_h(\cdot) := (1/h)K(\cdot/h)$. Let $\tau^2 := 2\int f_0^2(y)dy \int (K * K(y))^2 dy$. Under Assumptions (11.3)–(11.5), the following normality results for T_n were proved in [2]:

$$\text{under } H_0 : n\sqrt{h}\Big(T_n - \frac{1}{nh}\int K^2(v)dv\Big) \to_D N(0,\tau^2); \tag{11.6}$$

$$\text{under } H_1 : f \neq f_0,\ \int (f(y) - f_0(y))^2 dy > 0,$$

$$\sqrt{n}\Big(T_n - \int \big\{K_h * (f - f_0)\big\}^2(y)dy\Big) \to_D N(0, 4\omega^2), \tag{11.7}$$

$$\omega^2 = \mathrm{Var}[f(\varepsilon) - f_0(\varepsilon)].$$

The result (11.6) was first proved in [3] under some different conditions. The proof in [2] uses a CLT for degenerate U statistics of [5] and is relatively simpler under somewhat weaker conditions.

To describe the asymptotic null distribution of $\widehat{T}_n$, we shall now state the required assumptions below, where for any smooth and square integrable functions φ, $\dot{\varphi}$, and $\ddot{\varphi}$ denote its first and second derivatives, respectively, $\|\varphi\|_2^2 := \int \varphi^2(y)dy$ and f is the error density.

$$X'X \text{ is nonsingular and with } A = (X'X)^{1/2},\ n^{1/2} \max_{1\le i\le n} \big\|A^{-1}x_i\big\| = O(1). \tag{11.8}$$

$$\big\|A\big(\widehat{\beta} - \beta\big)\big\| = O_p(1) \text{ and } n^{1/2}\big|\widehat{\alpha} - \alpha\big| = O_p(1). \tag{11.9}$$

$$f \text{ is twice continuously differentiable, and } \|f\|_2 + \|\dot{f}\|_2 + \|\ddot{f}\|_2 < \infty. \tag{11.10}$$

$$K \text{ is a twice differentiable density on } \mathbb{R},\ K(-v) \equiv K(v),\ \int v^2 K(v)dv < \infty,$$

$$\int \big|\dot{K}(v)\big|dv < \infty,\ \|\dot{K}\|_2 + \|\ddot{K}\|_2 < \infty \text{ and } \int \big|v^j \dot{K}(v)\big|dv < \infty,\ j = 1, 2. \tag{11.11}$$

$$h \to 0 \text{ and } nh^5 \to \infty. \tag{11.12}$$

The following theorem describes the asymptotic null distribution of $\widehat{T}_n$.

Theorem 11.1 *Suppose Assumptions* (11.8)–(11.10) *with* $f = f_0$, (11.11) *and* (11.12) *hold. Then, under* H_0,

$$nh^{1/2}\big|\widehat{T}_n - T_n\big| \to_p 0. \tag{11.13}$$

Hence, if in addition $\ddot{f}_0$ *is bounded and* K *is compactly supported, then, under* H_0*,*

$$nh^{1/2}\Big(\widehat{T}_n - \frac{1}{nh}\int K^2(v)dv\Big) \to_D N\big(0, \tau^2\big). \tag{11.14}$$

Proof The claim (11.14) follows from Slutsky's theorem, (11.13) and (11.6). We begin with the following decomposition:

$$\begin{aligned}\widehat{T}_n - T_n &= \int \big(\widehat{f}_n(y) - f_n(y)\big)^2 dy \\ &\quad +2\int \big(\widehat{f}_n(y) - f_n(y)\big)\big(f_n(y) - \mu_n(y)\big)dy.\end{aligned} \tag{11.15}$$

The claim (11.13) will thus be implied by the following lemma. □

Lemma 11.1 (a) *Assume the error density to be* f *and Assumptions* (11.8)–(11.12) *hold. Then,*

$$nh^{1/2}\int \big(\widehat{f}_n(y) - f_n(y)\big)^2 dy = O_p\Big(\frac{1}{nh^{5/2}}\Big) + O_p\Big(h^{1/2}\Big) + O_p\Big(\frac{1}{nh^{9/2}}\Big). \tag{11.16}$$

(b) *Assume the error density to be* f_0 *and Assumptions* (11.8)–(11.10) *with* $f = f_0$, (11.11) *and* (11.12) *hold. Then, under* H_0*,*

$$\begin{aligned}&nh^{1/2}\Big|\int \big(\widehat{f}_n(y) - f_n(y)\big)\big(f_n(y) - \mu_n(y)\big)dy\Big| \\ &= O_p\Big(\frac{1}{n^{1/2}h^{3/2}}\Big) + O_p\Big(\frac{1}{n^{1/2}h^{5/2}}\Big) + O_p\Big(h^{1/2}\Big).\end{aligned} \tag{11.17}$$

Proof We first prove (11.16). Let

$$\delta_1 := \widehat{\alpha} - \alpha, \qquad \delta_2 := \widehat{\beta} - \beta, \qquad \xi_i := \delta_1 + \delta_2' x_i, \quad 1 \le i \le n.$$

By (11.8) and (11.9),

$$\begin{aligned}n^{1/2}\max_{1\le i\le n}\big|\delta_2' x_i\big| &= n^{1/2}\max_{1\le i\le n}\big|(\widehat{\beta} - \beta)' A A^{-1} x_i\big| \\ &\le \|A(\widehat{\beta} - \beta)\| n^{1/2}\max_{1\le i\le n}\big\|A^{-1}x_i\big\| = O_p(1), \\ n^{1/2}\max_{1\le i\le n}\big|\xi_i\big| &\le n^{1/2}\big|\delta_1\big| + n^{1/2}\max_{1\le i\le n}\big|\delta_2' x_i\big| = O_p(1).\end{aligned} \tag{11.18}$$

□

Moreover,

$$\varepsilon_i - \widehat{\varepsilon}_i = (\widehat{\alpha} - \alpha) + (\widehat{\beta} - \beta)' x_i = \xi_i, \qquad 1 \le i \le n.$$

These facts, the Cauchy–Schwarz (C-S) inequality and Fubini's theorem, are often used in the proofs below, sometimes without mentioning them.

By the Taylor expansion with integral remainder,

$$K\big(\frac{y-\widehat{\varepsilon}_i}{h}\big) - K\big(\frac{y-\varepsilon_i}{h}\big) = \frac{\xi_i}{h}\dot{K}\big(\frac{y-\varepsilon_i}{h}\big) + \int_{(y-\varepsilon_i)/h}^{(y-\widehat{\varepsilon}_i)/h} \ddot{K}(t)\Big(\frac{y-\varepsilon_i}{h} - t\Big)dt.$$

Hence,

$$\begin{aligned} \widehat{f}_n(y) - f_n(y) &= \frac{1}{nh}\sum_{i=1}^n \Big[K\big(\frac{y-\widehat{\varepsilon}_i}{h}\big) - K\big(\frac{y-\varepsilon_i}{h}\big)\Big] \\ &= \frac{1}{nh^2}\sum_{i=1}^n \xi_i \dot{K}\big(\frac{y-\varepsilon_i}{h}\big) + \frac{1}{nh}\sum_{i=1}^n \int_{(y-\varepsilon_i)/h}^{(y-\widehat{\varepsilon}_i)/h} \ddot{K}(t)\Big(\frac{y-\varepsilon_i}{h} - t\Big)dt. \end{aligned} \tag{11.19}$$

Let

$$\begin{aligned} C_{n1} &:= \int \Big(\sum_{i=1}^n \xi_i \dot{K}\big(\frac{y-\varepsilon_i}{h}\big)\Big)^2 dy, \\ C_{n2} &:= \int \Big(\sum_{i=1}^n \int_{(y-\varepsilon_i)/h}^{(y-\widehat{\varepsilon}_i)/h} \ddot{K}(t)\Big(\frac{y-\varepsilon_i}{h} - t\Big)dt\Big)^2 dy. \end{aligned}$$

Then,

$$\int \big(\widehat{f}_n(y) - f_n(y)\big)^2 dy \le 2\Big[\frac{1}{n^2h^4}C_{n1} + \frac{1}{n^2h^2}C_{n2}\Big]. \tag{11.20}$$

Let $\Delta := A\delta_2$, $c_i := A^{-1}x_i$, $1 \le i \le n$. Then $\delta_2' x_i \equiv \delta_2' A A^{-1} x_i = \Delta' c_i$ and $\xi_i = \delta_1 + \delta_2' x_i = \delta_1 + \Delta' c_i$, for all $1 \le i \le n$. By (11.9), $\|\Delta\| = O_p(1)$. Moreover,

$$\begin{aligned} &\sum_{i=1}^n \|c_i\|^2 = \sum_{i=1}^n \big\|A^{-1}x_i\big\|^2 = \sum_{i=1}^n x_i' A^{-1} A^{-1} x_i = \operatorname{trace}\big(A^{-1} X' X A^{-1}\big) = p, \\ &\Big\|\sum_{i=1}^n c_i\Big\|^2 \le n \sum_{i=1}^n \|c_i\|^2 = np. \end{aligned} \tag{11.21}$$

To analyze C_{n1}, let

$$C_{n11} := \int \Big(\sum_{i=1}^{n} \dot{K}\big(\frac{y-\varepsilon_i}{h}\big)\Big)^2 dy, \qquad C_{n12} := \int \Big\| \sum_{i=1}^{n} c_i\, \dot{K}\big(\frac{y-\varepsilon_i}{h}\big)\Big\|^2 dy. \tag{11.22}$$

Then,

$$C_{n1} = \int \Big(\sum_{i=1}^{n} \big(\delta_1 + \Delta' c_i\big)\, \dot{K}\big(\frac{y-\varepsilon_i}{h}\big)\Big)^2 dy \le 2\delta_1^2 C_{n11} + 2\|\Delta\|^2 C_{n12}.$$

Center the summands inside the square in C_{n11} to write

$$\begin{aligned} C_{n11} &= \int \Big(\sum_{i=1}^{n} \Big\{\dot{K}\big(\frac{y-\varepsilon_i}{h}\big) - E\dot{K}\big(\frac{y-\varepsilon_i}{h}\big)\Big\} + \sum_{i=1}^{n} E\dot{K}\big(\frac{y-\varepsilon_i}{h}\big)\Big)^2 dy \\ &\le 2G_{n1} + 2G_{n2}, \end{aligned} \tag{11.23}$$

where

$$\begin{aligned} G_{n1} &:= \int \Big(\sum_{i=1}^{n} \Big\{\dot{K}\big(\frac{y-\varepsilon_i}{h}\big) - E\dot{K}\big(\frac{y-\varepsilon_i}{h}\big)\Big\}\Big)^2 dy, \\ G_{n2} &:= \int \Big(\sum_{i=1}^{n} E\dot{K}\big(\frac{y-\varepsilon_i}{h}\big)\Big)^2 dy. \end{aligned}$$

By Fubini's theorem,

$$\begin{aligned} EG_{n1} &= n \int \mathrm{Var}\Big(\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big) dy \le n \int E\Big(\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big)^2 dy \\ &= n\,h \int \dot{K}^2(v) \int f(y-hv)\,dy\,dv = n\,h \int \dot{K}^2(v)\,dv. \end{aligned} \tag{11.24}$$

Note that

$$\int \Big(E\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big)^2 dy = h^2 \int \Big(\int \dot{K}(v) f(y-hv)\,dv\Big)^2 dy.$$

Because $\int \dot{K}(v) f(y-hv)dv = \int f(y-hv)dK(v)$, the integration by parts shows that

$$\begin{aligned} \int \Big(\int \dot{K}(v) f(y-hv)\,dv\Big)^2 dy &= h^2 \int \Big(\int K(v) \dot{f}(y-hv)\,dv\Big)^2 dy \\ &= h^2 \int \dot{f}^2(y)\,dy + O\big(h^2\big) = O\big(h^2\big). \end{aligned}$$

The above fact follows from the following result which uses $\dot{f}, \ddot{f}$ being square integrable:

$$\begin{aligned}
&\int \Big(\int K(v) \Big\{ \dot{f}(y-hv) - \dot{f}(y) \Big\} dv \Big)^2 dy \\
&= \int \Big(\int v \Big\{ \int_0^h \ddot{f}(y-sv) ds \Big\} K(v) dv \Big)^2 dy \\
&\le \int v^2 K(v) dv \int \int \Big\{ \int_0^h \ddot{f}(y-sv) ds \Big\}^2 K(v) dv dy \\
&\le h \int v^2 K(v) dv \int \int \int_0^h \ddot{f}^2(y-sv) ds K(v) dv dy \\
&= h^2 \int v^2 K(v) dv \int \ddot{f}^2(y) dy.
\end{aligned}$$

Therefore,

$$\int \Big(E\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big)^2 dy = h^4 \int \dot{f}^2(y) dy + O\big(h^4\big) = O(h^4),$$

$$G_{n2} = n^2 \int \Big(E\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big)^2 dy = O(n^2 h^4). \tag{11.25}$$

Because $\delta_1^2 = O_p(n^{-1})$, by the Markov inequality, (11.23)–(11.25),

$$G_{n1} = O_p\big(nh\big), \;\; C_{n11} = O_p\big(nh\big) + O\big(n^2 h^4\big), \;\; \delta_1^2 C_{n11} = O_p\big(h\big) + O_p\big(nh^4\big). \tag{11.26}$$

Similarly we obtain

$$\begin{aligned}
C_{n12} &\le 2 \int \Big\| \sum_{i=1}^n c_i \Big[\dot{K}\big(\frac{y-\varepsilon_i}{h}\big) - E\dot{K}\big(\frac{y-\varepsilon_i}{h}\big)\Big]\Big\|^2 dy \\
&\quad + 2 \int \Big\| \sum_{i=1}^n c_i E\dot{K}\big(\frac{y-\varepsilon_i}{h}\big)\Big\|^2 dy, \qquad (11.27) \\
EC_{n12} &\le 2 \sum_{i=1}^n \|c_i\|^2 \int E\dot{K}^2\big(\frac{y-\varepsilon}{h}\big) dy + 2 \Big\| \sum_{i=1}^n c_i \Big\|^2 \int \Big(E\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big)^2 dy \\
&= 2\, p\, h \int \dot{K}^2(v) dv + np\, O(h^4).
\end{aligned}$$

Therefore, by the Markov inequality and Assumptions (11.9), (11.11) and the fact (11.26),

$$C_{n12} = O_p(h) + O(nh^4) = \|\Delta\|^2 C_{n12}, \qquad C_{n1} = O_p(h) + O(nh^4),$$
$$\frac{1}{n^2h^4} C_{n1} = O_p\Big(\frac{1}{n^2h^3}\Big) + O_p\Big(\frac{1}{n}\Big), \quad nh^{1/2}\frac{1}{n^2h^4} C_{n1} = O_p\Big(\frac{1}{nh^{5/2}}\Big) + O_p\big(h^{1/2}\big). \tag{11.28}$$

To analyze C_{n2}, by (11.8), (11.9) and (11.18),

$$\sum_{i=1}^n \xi_i^2 \le 2n\delta_1^2 + 2\|\Delta\|^2 \sum_{i=1}^n \|c_i\|^2 = 2\big(n\delta_1^2 + p\|\Delta\|^2\big) = O_p(1),$$
$$n\sum_{i=1}^n \xi_i^4 \le n \max_{1\le i\le n} \big|\xi_i\big|^2 \sum_{i=1}^n \xi_i^2 = O_p(1).$$

Use these bounds, the C-S inequality on the sum and then on the integral w.r.t. t, Fubini's theorem and the fact $(y-\varepsilon_i)/h \le t \le (y-\widehat{\varepsilon}_i)/h$, for all $1\le i\le n$, to obtain

$$\begin{aligned}
C_{n2} &= \int \Big(\sum_{i=1}^n \int_{(y-\varepsilon_i)/h}^{(y-\widehat{\varepsilon}_i)/h} \ddot{K}(t)\Big(\frac{y-\varepsilon_i}{h} - t\Big)dt\Big)^2 dy \\
&\le n\int \sum_{i=1}^n \Big(\int_{(y-\varepsilon_i)/h}^{(y-\widehat{\varepsilon}_i)/h} \ddot{K}(t)\Big(\frac{y-\varepsilon_i}{h} - t\Big)dt\Big)^2 dy \\
&\le n\sum_{i=1}^n \frac{\big|\xi_i\big|^3}{h^3} \int\int_{(y-\varepsilon_i)/h}^{(y-\widehat{\varepsilon}_i)/h} \ddot{K}^2(t)dtdy \\
&= n\sum_{i=1}^n \frac{\big|\xi_i\big|^3}{h^3} \int\int_{\widehat{\varepsilon}_i+ht}^{\varepsilon_i+ht} dy\ddot{K}^2(t)dt = n\sum_{i=1}^n \frac{\big|\xi_i\big|^4}{h^3}\int \ddot{K}^2(t)dt = O_p\Big(\frac{1}{h^3}\Big).
\end{aligned}$$

Hence

$$C_{n2} = O_p\Big(\frac{1}{h^3}\Big), \qquad nh^{1/2}\frac{1}{n^2h^2} C_{n2} = O_p\Big(\frac{1}{nh^{9/2}}\Big). \tag{11.29}$$

The proof of (11.16) is completed by combining (11.29) with (11.28) and (11.20).

Proof of (11.17). Now assume H_0 is true and Assumptions (11.9) and (11.10) hold under H_0, i.e., when $f = f_0$. Let E_0 denote the expectation under H_0. From (11.6) and $h\to 0$, we obtain

$$T_n = \frac{1}{nh}\int K^2(v)dv + O_p\Big(\frac{1}{nh^{1/2}}\Big) = O_p\Big(\frac{1}{nh}\Big). \tag{11.30}$$

Let $\rho_n(y) := f_n(y) - \mu_n(y)$. Use (11.19) to rewrite

$$\int \big(\widehat{f}_n(y) - f_n(y)\big)\big(f_n(y) - \mu_n(y)\big)dy$$
$$= \frac{1}{nh^2} \int \sum_{i=1}^{n} \xi_i \dot{K}\big(\frac{y - \varepsilon_i}{h}\big)\rho_n(y)dy$$
$$+ \frac{1}{nh} \int \sum_{i=1}^{n} \int_{(y-\varepsilon_i)/h}^{(y-\widehat{\varepsilon}_i)/h} \ddot{K}(t)\Big(\frac{y - \varepsilon_i}{h} - t\Big)dt\rho_n(y)dy$$
$$= \frac{1}{nh^2} D_{n1} + \frac{1}{nh} D_{n2}, \qquad \text{say.} \tag{11.31}$$

By the C-S inequality, (11.29) and (11.30),

$$\big|D_{n2}\big| := \Big| \int \Big(\sum_{i=1}^{n} \int_{(y-\varepsilon_i)/h}^{(y-\widehat{\varepsilon}_i)/h} \ddot{K}(t)\Big(\frac{y - \varepsilon_i}{h} - t\Big)dt\Big)\rho_n(y)dy\Big|$$
$$\leq C_{n2}^{1/2} T_n^{1/2} = O_p\big(\frac{1}{h^{3/2}}\big) O_p\big(\frac{1}{n^{1/2}h^{1/2}}\big) = O_p\big(\frac{1}{n^{1/2}h^2}\big), \tag{11.32}$$
$$\frac{1}{nh}\big|D_{n2}\big| = O_p\big(\frac{1}{n^{3/2}h^3}\big), \qquad nh^{1/2}\frac{1}{nh}\big|D_{n2}\big| = O_p\big(\frac{1}{n^{1/2}h^{5/2}}\big).$$

To analyze D_{n1}, let

$$D_{n11} := \int \sum_{i=1}^{n} \dot{K}\big(\frac{y - \varepsilon_i}{h}\big)\rho_n(y)dy, \quad D_{n12} := \int \sum_{i=1}^{n} c_i \dot{K}\big(\frac{y - \varepsilon_i}{h}\big)\rho_n(y)dy. \tag{11.33}$$

Then,

$$D_{n1} := \int \sum_{i=1}^{n} \xi_i \dot{K}\big(\frac{y - \varepsilon_i}{h}\big)\rho_n(y)dy = \delta_1 D_{n11} + \Delta' D_{n12}. \tag{11.34}$$

Write

$$D_{n11} := \int \sum_{i=1}^{n} \Big[\dot{K}\big(\frac{y - \varepsilon_i}{h}\big) - E_0 \dot{K}\big(\frac{y - \varepsilon}{h}\big)\Big]\rho_n(y)dy$$
$$+ n \int E_0\Big(\dot{K}\big(\frac{y - \varepsilon}{h}\big)\Big)\rho_n(y)dy$$
$$= \Gamma_{n1} + n\Gamma_{n2}, \quad \text{say.} \tag{11.35}$$

Note that (11.26) is true for any f satisfying (11.10), so it is true also when $f = f_0$. By the C-S inequality, (11.26) and (11.30),

$$\left|\Gamma_{n1}\right| \leq G_{n1}^{1/2} T_n^{1/2} = O_p(1), \qquad nh^{1/2}\frac{1}{nh^2}\left|\delta_1\Gamma_{n1}\right| = O_p(\frac{1}{n^{1/2}h^{3/2}}). \quad (11.36)$$

To analyze Γ_{n2}, let $\gamma_n(y) := \int f_0(y-vh)\dot{K}(v)dv$ and $\zeta_{ni}(y) := h^{-1}K\big((y-\varepsilon_i)/h\big) - \mu_n(y)$. Then,

$$\begin{aligned}
E_0\dot{K}\big(\frac{y-\varepsilon}{h}\big) &= \int \dot{K}\big(\frac{y-x}{h}\big) f_0(x)dx = h\int f_0(y-hv)\dot{K}(v)dv = h\gamma_n(y),\\
\rho_n(y) &= n^{-1}\sum_{i=1}^n \Big[\frac{1}{h}K\big(\frac{y-\varepsilon_i}{h}\big) - \mu_n(y)\Big] = n^{-1}\sum_{i=1}^n \zeta_{ni}(y),\\
\Gamma_{n2} &:= \int E_0\Big(\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big)\rho_n(y)dy = h\int \gamma_n(y)\rho_n(y)dy\\
&= hn^{-1}\sum_{i=1}^n \int \gamma_n(y)\zeta_{ni}(y)dy.
\end{aligned}$$

Therefore,

$$n\Gamma_{n2} = h\sum_{i=1}^n \int \gamma_n(y)\zeta_{ni}(y)dy. \quad (11.37)$$

By (11.11), $\int \left|v\dot{K}(v)\right|dv < \infty$, $\int v^2K(v)dv < \infty$ and $\int vK(v)dv = 0$. Hence, $vK(v) \to 0$ and $v^2K(v) \to 0$, as $v \to \pm\infty$, and by the integration by parts, $\int v\dot{K}(v)dv = \int vdK(v) = -\int K(v)dv = -1$ and $\int v^2\dot{K}(v)dv = \int v^2dK(v) = -2\int vK(v)dv = 0$.

Using the Taylor expansion for f_0 with integral remainder and by (11.11), which guarantees $\int \dot{K}(v)dv = 0$, we obtain

$$f_0(y-hv) = f_0(y) - hv\dot{f}_0(y) + v^2\int_0^h \ddot{f}_0(y-sv)(h-s)ds, \quad (11.38)$$

$$\begin{aligned}
\gamma_n(y) &= \int \Big[f_0(y) - hv\dot{f}_0(y) + v^2\int_0^h \ddot{f}_0(y-sv)(h-s)ds\Big]\dot{K}(v)dv\\
&= h\dot{f}_0(y) + \int \psi_n(y,v)v^2\dot{K}(v)dv,
\end{aligned}$$

where $\psi_n(y,v) := \int_0^h \ddot{f}_0(y-sv)(h-s)ds$. Let $\ell_n(y) := \int \psi_n(y,v)v^2\dot{K}(v)dv$ and let

$$S_{n1} := \frac{1}{n^{1/2}}\sum_{i=1}^n \int \dot{f}_0(y)\zeta_{ni}(y)dy, \quad S_{n2} := \frac{1}{n^{1/2}}\sum_{i=1}^n \int \ell_n(y)\zeta_{ni}(y)dy.$$

Then, by (11.37) and (11.38),

$$n\Gamma_{n2} = h\sum_{i=1}^{n}\int\Big[h\dot f_0(y) + \int \psi_n(y,v)v^2\dot K(v)dv\Big]\zeta_{ni}(y)dy$$
$$= h^2n^{1/2}S_{n1} + hn^{1/2}S_{n2}, \tag{11.39}$$
$$nh^{1/2}\frac{1}{nh^2}\delta_1 n\Gamma_{n2} = n^{1/2}\delta_1\Big(h^{1/2}S_{n1} + \frac{1}{h^{1/2}}S_{n2}\Big).$$

Note that $E_0S_{n1} = 0$ and by Fubini's theorem,

$$E_0S_{n1}^2 = E_0\Big(\int \dot f_0(y)\Big[\frac{1}{h}K\big(\frac{y-\varepsilon}{h}\big) - \mu_n(y)\Big]dy\Big)^2$$
$$= E_0\int\int \dot f_0(x)\dot f_0(y)\Big[\frac{1}{h^2}K\big(\frac{x-\varepsilon}{h}\big)K\big(\frac{y-\varepsilon}{h}\big) - \mu_n(x)\mu_n(y)\Big]dxdy$$
$$= E_0\int\int \dot f_0(\varepsilon+uh)\dot f_0(\varepsilon+vh)K\big(u\big)K\big(v\big)dudv$$
$$- h^2E_0\int\int \dot f_0(\varepsilon+uh)\dot f_0(\varepsilon+vh)\mu_n(\varepsilon+uh)\mu_n(\varepsilon+vh)dudv$$
$$= E_0\Big(\int \dot f_0(\varepsilon+uh)K\big(u\big)du\Big)^2 - h^2E_0\Big(\int \dot f_0(\varepsilon+uh)\mu_n(\varepsilon+uh)du\Big)^2$$
$$\le E_0\Big(\int \dot f_0(\varepsilon+uh)K\big(u\big)du\Big)^2 \to E_0\dot f_0^2(\varepsilon).$$

Hence, by the Markov inequality and (11.9),

$$\big|S_{n1}\big| = O_p\big(1\big), \qquad \big|h^{1/2}S_{n1}\big| = O_p\big(h^{1/2}\big). \tag{11.40}$$

To analyze S_{n2}, use the C-S inequality and Fubini's theorem to obtain

$$\psi_n^2(y,v) = \Big(\int_0^h \ddot f_0(y-sv)(h-s)ds\Big)^2 \le h^3\int_0^h \ddot f_0^2(y-sv)ds,$$
$$\ell_n^2(y) = \Big(\int \psi_n(y,v)v^2\dot K(v)dv\Big)^2 = \Big(\int \psi_n(y,v)\big|v^2\dot K(v)\big|^{1/2}\big|v^2\dot K(v)\big|^{1/2}dv\Big)^2$$
$$\le \int \big|v^2\dot K(v)\big|dv\int \psi_n^2(y,v)v^2\big|\dot K(v)\big|dv$$
$$\le h^3\int v^2\big|\dot K(v)\big|dv\int\int_0^h \ddot f^2(y-sv)dsv^2\big|\dot K(v)\big|dv.$$

Therefore,

$$\int \ell_n^2(y)dy \le h^3\int v^2\big|\dot K(v)\big|dv\int_0^h\int\int \ddot f_0^2(y-sv)dy\,v^2|\dot K(v)|ds$$
$$= h^4\Big(\int v^2\big|\dot K(v)\big|dv\Big)^2\int \ddot f_0^2(y)dy. \tag{11.41}$$

Also, $E_0 S_{n2} = 0$ and

$$E_0\big(S_{n2}^2\big) = E_0\Big(\int \ell_n(y)\Big[\frac{1}{h}K\big(\frac{y-\varepsilon}{h}\big) - \mu_n(x)\Big]dy\Big)^2$$
$$\le \int \ell_n^2(y)dy\,\frac{1}{h^2}\int E_0\Big(K^2\big(\frac{y-\varepsilon}{h}\big)\Big)dy$$
$$\le h^4\Big(\int v^2\big|\dot{K}(v)\big|dv\Big)^2\int \ddot{f}_0^2(y)dy\,\frac{1}{h}\int K^2(v)dv$$
$$= h^3\Big(\int v^2\big|\dot{K}(v)\big|dv\Big)^2\int \ddot{f}_0^2(y)dy\int K^2(v)dv.$$

Therefore,

$$\big|S_{n2}\big| = O_p\big(h^{3/2}\big), \qquad \big|\frac{1}{h^{1/2}}S_{n2}\big| = O_p\big(h\big). \tag{11.42}$$

In view of (11.39), (11.40), and (11.42) and Assumption (11.9), which guarantees $n^{1/2}|\delta_1| = O_p(1)$, we obtain

$$\big|\Gamma_{n2}\big| = O_p\Big(\frac{h^2}{n^{1/2}}\Big), \qquad nh^{1/2}\frac{1}{nh^2}\big|\delta_1 n\Gamma_{n2}\big| = O_p(h^{1/2}). \tag{11.43}$$

Combine (11.43) with (11.36) and (11.35) to obtain

$$nh^{1/2}\frac{1}{nh^2}\big|\delta_1 D_{n11}\big| = O_p\Big(\frac{1}{n^{1/2}h^{3/2}}\Big) + O_p\big(h^{1/2}\big). \tag{11.44}$$

We next consider D_{n12} of (11.34), where

$$D_{n12} := \int\sum_{i=1}^n c_i\Big[\dot{K}\big(\frac{y-\varepsilon_i}{h}\big) - E_0\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big]\rho_n(y)dy$$
$$+\sum_{i=1}^n c_i\int E_0\Big(\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big)\rho_n(y)dy = B_{n1} + B_{n2}, \quad \text{say}.$$

Note that with Γ_{n2} as in (11.37) and in view of (11.21) and (11.43),

$$\big\|B_{n2}\big\| = \Big\|\sum_{i=1}^n c_i\Big\|\big|\Gamma_{n2}\big| = O_p\big(h^2\big).$$

By the C-S inequality, arguing as for the C_{n12}, we obtain, in view of (11.30),

$$\|B_{n1}\| \le \Big(\int \Big\| \sum_{i=1}^{n} c_i \Big[\dot{K}\big(\frac{y-\varepsilon_i}{h}\big) - E_0 \dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big]\Big\|^2 dy\Big)^{1/2} T_n^{1/2}$$
$$= O_p\big(h^{1/2}\big) O_p\Big(\frac{1}{(nh)^{1/2}}\Big) = O_p\Big(\frac{1}{n^{1/2}}\Big).$$

Therefore,

$$nh^{1/2}\frac{1}{nh^2}\big|\Delta' D_{n12}\big| \le \|\Delta\| \frac{1}{h^{3/2}} \|D_{n12}\| = \frac{1}{h^{3/2}}\Big[O_p\Big(\frac{1}{n^{1/2}}\Big) + O_p\big(h^2\big)\Big]$$
$$= O_p\Big(\frac{1}{n^{1/2}h^{3/2}}\Big) + O_p\big(h^{1/2}\big).$$

Upon combining this bound with (11.34) and (11.44) we obtain

$$nh^{1/2}\frac{1}{nh^2}\big|D_{n1}\big| = O_p\Big(\frac{1}{n^{1/2}h^{3/2}}\Big) + O_p\big(h^{1/2}\big).$$

The proof of (11.17) is completed by combining this fact with (11.32) and (11.31). This also completes the proof of the lemma. □

11.3 Asymptotic Distribution of $\widehat{T}_n$ Under H_1

We shall now discuss the asymptotic distribution of $\widehat{T}_n$ under the global L_2 alternatives considered in [2], $H_1 : f \ne f_0,\ \int \big(f(y) - f_0(y)\big)^2 dy > 0$. As in the case of H_0, the first step useful for this purpose is to analyze the difference $\widehat{T}_n - T_n$ under H_1. The starting point is again the decomposition (11.15). Recall that (11.16) is true for any f satisfying (11.10), so it is a priori true for any such f under H_1. Hence, by (11.12) and (11.16), under H_1,

$$n^{1/2}\int \big(\widehat{f}_n(y) - f_n(y)\big)^2 dy = o_p(1). \tag{11.45}$$

The following lemma shows the behavior of the cross-product term. Unlike (11.17), it shows that the $n^{1/2} \times$ (the cross-product term) is approximated by a sequence of non-vanishing random variables involving $n^{1/2}\delta_1$, Δ and $\mathcal{D} := \int \dot{f}(y)\big(f(y) - f_0(y)\big)dy$.

Lemma 11.2 *Assume H_1 and Assumptions* (11.9)–(11.12) *hold. Then,*

$$\Big| n^{1/2}\int \big(\widehat{f}_n(y) - f_n(y)\big)\big(f_n(y) - \mu_n(y)\big)dy - \Big(n^{1/2}\delta_1 + n^{-1/2}\Delta' \sum_{i=1}^{n} c_i\Big)\mathcal{D}\Big| = o_p(1). \tag{11.46}$$

Proof Assume H_1 holds and let f be a density satisfying (11.10) and H_1. Let E denote the expectation when density of ε is f. From (11.7), it follows that

$$T_n = \int K * (f - f_0)(v)dv + O_p\Big(\frac{1}{n^{1/2}}\Big) = O_p\big(1\big). \tag{11.47}$$

Many details of the proof below are similar to those of the proof of (11.17), where we use (11.47) instead of (11.30). □

Recall the decomposition (11.31), i.e., with $\rho_n = f_n - \mu_n$,

$$\int \big(\widehat{f}_n(y) - f_n(y)\big)\rho_n(y)dy = \frac{1}{nh^2}D_{n1} + \frac{1}{nh}D_{n2}, \tag{11.48}$$

where D_{n1}, D_{n2} are as in (11.32)–(11.34). Note that the inequalities (11.29) and (11.32) are valid for any f satisfying (11.10). By (11.29) and (11.47), we obtain

$$\big|D_{n2}\big| \le C_{n2}^{1/2}T_n^{1/2} = O_p\Big(h^{-3/2}\Big), \qquad \frac{1}{nh}\big|n^{1/2}D_{n2}\big| = O_p\Big(\frac{1}{n^{1/2}h^{5/2}}\Big). \tag{11.49}$$

To analyze D_{n1}, recall the decomposition (11.34), viz., $D_{n1} = \delta_1 D_{n11} + \Delta' D_{n12}$ so that

$$n^{1/2}D_{n1} = n^{1/2}\delta_1\, D_{n11} + n^{1/2}\Delta' D_{n12}. \tag{11.50}$$

Write

$$\begin{aligned} D_{n11} &:= \int \sum_{i=1}^{n}\Big[\dot{K}\big(\frac{y-\varepsilon_i}{h}\big) - E\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big]\rho_n(y)dy \\ &\quad + n\int E\Big(\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big)\rho_n(y)dy = J_{n1} + nJ_{n2}, \quad \text{say.} \end{aligned} \tag{11.51}$$

By the C-S inequality, (11.9), (11.24), and (11.47),

$$|J_{n1}| \le G_{n1}^{1/2}T_n^{1/2} = O_p\big((nh)^{1/2}\big), \qquad \frac{1}{nh^2}|J_{n1}| = O_p\big(\frac{1}{n^{1/2}h^{3/2}}\big). \tag{11.52}$$

As in the proof of (11.17), to analyze J_{n2} term, we introduce $\lambda_n(y) := \int f(y - vh)\dot{K}(v)dv$. The role played by $\lambda_n(y)$ here is analogous to that of $\gamma_n(y)$ under H_0. Note that

$$\begin{aligned} E\dot{K}\big(\frac{y-\varepsilon}{h}\big) &= \int \dot{K}\big(\frac{y-z}{h}\big)f(z)dz = h\int f(y-hv)\dot{K}(v)dv = h\lambda_n(y), \\ nJ_{n2} &= h\sum_{i=1}^{n}\int \lambda_n(y)\zeta_{ni}(y)dy, \end{aligned}$$

where $\zeta_{ni}(y) \equiv h^{-1}K\big((y-\varepsilon_i)/h\big) - \mu_n(y)$. Because now $E\zeta_{ni}(y) \neq 0$, the effect of alternatives is manifested in the asymptotic behavior of J_{n2} as is evidenced by the following fact:

$$\left|\frac{1}{nh^2} nJ_{n2} - \mathcal{D}\right| = o_p(1), \quad (H_1). \tag{11.53}$$

Proof of (11.53). Let $\nu_n(y) := \int f(y-hv)K(v)dv$, $g_n(y) := \nu_n(y) - \mu_n(y)$, and $\eta_{ni}(y) := h^{-1}K\big((y-\varepsilon_i)/h\big) - \nu_n(y)$. Clearly, $E\eta_{ni}(y) \equiv 0$, $\zeta_{ni}(y) \equiv \eta_{ni}(y) + g_n(y)$, and hence

$$\begin{aligned} J_{n2} &:= \frac{h}{n}\sum_{i=1}^n \int \lambda_n(y)\zeta_{ni}(y)dy = \frac{h}{n}\sum_{i=1}^n \int \lambda_n(y)\eta_{ni}(y)dy + h\int \lambda_n(y)g_n(y)dy \\ &= \frac{h}{n}J_{n21} + hJ_{n22}, \qquad \text{say.} \end{aligned}$$

Let $\varphi_n(y,v) := \int_0^h \ddot{f}(y-sv)(h-s)ds$ and $L_n(y) := \int \varphi_n(y,v)v^2\dot{K}(v)dv$. Arguing as for (11.38), we have

$$\lambda_n(y) = \int f(y-hv)\dot{K}(v)dv = h\dot{f}(y) + \int \varphi_n(y,v)v^2\dot{K}(v)dv = h\dot{f}(y) + L_n(y).$$

Let

$$\begin{aligned} V_{n1} &:= \frac{1}{n}\sum_{i=1}^n \int \dot{f}(y)\eta_{ni}(y)dy, \qquad V_{n2} := \frac{1}{n}\sum_{i=1}^n \int L_n(y)\eta_{ni}(y)dy, \\ \mathcal{D}_n &:= \int \dot{f}(y)g_n(y)dy, \qquad W_n := \int L_n(y)g_n(y)dy. \end{aligned}$$

Therefore,

$$\begin{aligned} J_{n21} &:= \sum_{i=1}^n \int \lambda_n(y)\eta_{ni}(y) \\ &= h\sum_{i=1}^n \int \dot{f}(y)\eta_{ni}(y)dy + \sum_{i=1}^n \int L_n(y)\eta_{ni}(y)dy = nhV_{n1} + nV_{n2}, \end{aligned} \tag{11.54}$$

$$\begin{aligned} J_{n22} &:= \int \lambda_n(y)g_n(y)dy = h\int \dot{f}(y)g_n(y)dy + \int L_n(y)g_n(y)dy, \\ &= h\mathcal{D}_n + W_n. \end{aligned} \tag{11.55}$$

Note that

$$\frac{1}{nh^2} nJ_{n2} - \mathcal{D} = V_{n1} + \frac{1}{h}\big(V_{n2} + W_n\big) + \mathcal{D}_n - \mathcal{D}.$$

Arguing as for S_{n1}, we have $EV_{n1} = 0$ and $nE\big(V_{n1}^2\big) \le E\Big(\int \dot{f}(\varepsilon + uh)K\big(u\big)du\Big)^2$ $\to E\dot{f}^2(\varepsilon)$. Hence, by the Markov inequality,

$$\big|V_{n1}\big| = O_p\big(\frac{1}{n^{1/2}}\big) = o_p(1). \tag{11.56}$$

Analogous to (11.41), we have

$$\int L_n^2(y)dy \le h^4\Big(\int v^2\big|\dot{K}(v)\big|dv\Big)^2 \int \ddot{f}^2(y)dy. \tag{11.57}$$

Also, arguing as for S_{n2}, we obtain $EV_{n2} = 0$ and

$$nE\big(V_{n2}^2\big) \le h^3\Big(\int v^2\big|\dot{K}(v)\big|dv\Big)^2 \int \ddot{f}^2(y)dy \int K^2(v)dv.$$

Therefore,

$$\big|V_{n2}\big| = O_p\Big(\frac{h^{3/2}}{n^{1/2}}\Big), \quad h\big|V_{n2}\big| = O_p\Big(\frac{h^{5/2}}{n^{1/2}}\Big) = o_p(1). \tag{11.58}$$

On the other hand, we have

$$\begin{aligned}
\int g_n^2(y)dy &= \int \big(\nu_n(y) - \mu_n(y)\big)^2 dy \\
&= \int \Big(\int \big[f(y-hv) - f_0(y-hv)\big]K(v)dv\Big)^2 dy \\
&\le \int\int \big[f(y-hv) - f_0(y-hv)\big]^2 K(v)dvdy \\
&= \int \big[f(y) - f_0(y)\big]^2 dy < \infty.
\end{aligned}$$

Hence, by (11.57),

$$W_n^2 \le \int L_n^2(y)dy \int g_n^2(y)dy = O(h^4), \quad h\big|W_n\big| = O(h^3) = o(1). \tag{11.59}$$

We shall shortly prove

$$\big|\mathcal{D}_n - \mathcal{D}\big| = O(h) = o(1). \tag{11.60}$$

From (11.54)–(11.56) and (11.58)–(11.60), we thus obtain

$$\begin{aligned}
J_{n2} &= h^2 V_{n1} + hV_{n2} + hW_n + h^2\mathcal{D}_n, \qquad (11.61)\\
|J_{n2}| &\le h^2|V_{n1}| + h|V_{n2}| + h|W_n| + h^2|\mathcal{D}_n|\\
&= O_p\Big(\frac{h^2}{n^{1/2}}\Big) + O_p\Big(\frac{h^{5/2}}{n^{1/2}}\Big) + O\Big(h^3\Big) + O_p\Big(h^2\Big) = O_p\Big(h^2\Big),\\
\Big|\frac{1}{nh^2} nJ_{n2} - \mathcal{D}\Big| &= O_p\Big(\frac{1}{n^{1/2}}\Big) + O_p\Big(\frac{h^{1/2}}{n^{1/2}}\Big) + O\Big(h\Big) + o_p(1) = o_p(1),
\end{aligned}$$

thereby proving (11.53).

Proof of (11.60). We can see that

$$\begin{aligned}
\mathcal{L}_n(f) &:= \int\Big(\int\big[f(y-hv) - f(y)\big]K(v)dv\Big)^2 dy\\
&= \int\Big(\int\Big\{\int_0^h \dot{f}(y-sv)ds\Big\}(-v)K(v)dv\Big)^2 dy\\
&\le \int\Big(\int|vK(v)|^{1/2}\Big\{\int_0^h|\dot{f}(y-sv)|ds\Big\}|vK(v)|^{1/2}dv\Big)^2 dy\\
&\le \Big(\int|v|K(v)dv\Big)\int\int\Big(\int_0^h|\dot{f}(y-sv)|ds\Big)^2|v|K(v)dvdy\\
&\le h\Big(\int|v|K(v)dv\Big)\int\int\Big(\int_0^h\dot{f}^2(y-sv)ds\Big)|v|K(v)dvdy\\
&= h\Big(\int|v|K(v)dv\Big)\int_0^h\int\Big(\int\dot{f}^2(y-sv)dy\Big)|v|K(v)dvds\\
&= h^2\Big(\int|v|K(v)dv\Big)^2\int\dot{f}^2(y)dy = O(h^2) \to 0.
\end{aligned}$$

Similarly $\mathcal{L}_n(f_0) = O(h^2) \to 0$. Therefore,

$$\begin{aligned}
&\int\Big(g_n(y) - \big(f(y) - f_0(y)\big)\Big)^2 dy\\
&= \int\Big\{\int\Big(f(y-hv) - f_0(y-hv)\Big)K(v)dv - \Big(f(y) - f_0(y)\Big)\int K(v)dv\Big\}^2 dy\\
&= \int\Big\{\int\Big(f(y-hv) - f(y)\Big)K(v)dv - \int\Big(f_0(y-hv) - f_0(y)\Big)K(v)dv\Big\}^2 dy\\
&\le 2\mathcal{L}_n(f) + 2\mathcal{L}_n(f_0) = O(h^2) \to 0.
\end{aligned}$$

This fact in turn implies

$$\int g_n^2(y)dy \to \int \big(f(y) - f_0(y)\big)^2 dy,$$
$$\big|\mathcal{D}_n - \mathcal{D}\big| = \Big|\int \dot{f}(y)\Big(g_n(y) - \big(f(y) - f_0(y)\big)\Big)dy\Big|$$
$$\le \Big\{\int \dot{f}^2(y)dy \int \Big(g_n(y) - \big(f(y) - f_0(y)\big)\Big)^2 dy\Big\}^{1/2}$$
$$= O(h) = o(1),$$

thereby proving (11.60).

Summarizing, in view of (11.9) and (11.51)–(11.53), we have

$$\Big|\frac{n^{1/2}}{nh^2}\delta_1 D_{n11} - n^{1/2}\delta_1\mathcal{D}\Big| = o_p(1). \tag{11.62}$$

To analyze D_{n12} of (11.34), recall the definition of J_{n2} from (11.51) and write

$$D_{n12} := \int \sum_{i=1}^n c_i\Big[\dot{K}\big(\frac{y-\varepsilon_i}{h}\big) - E\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big]\rho_n(y)dy + \sum_{i=1}^n c_i J_{n2}$$
$$= \mathcal{B}_{n1} + \mathcal{B}_{n2}, \quad \text{say.}$$

By the C-S inequality and by arguing as for the C_{n12}, we obtain, in view of (11.47),

$$\big\|\mathcal{B}_{n1}\big\| \le \Big(\int \Big\|\sum_{i=1}^n c_i\Big[\dot{K}\big(\frac{y-\varepsilon_i}{h}\big) - E\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big]\Big\|^2 dy\Big)^{1/2} T_n^{1/2} = O_p\big(h^{1/2}\big). \tag{11.63}$$

Hence,

$$\frac{n^{1/2}}{nh^2}\big|\Delta'\mathcal{B}_{n1}\big| = O_p\big(\frac{1}{n^{1/2}h^{3/2}}\big).$$

In view of (11.61),

$$\frac{n^{1/2}}{nh^2}\mathcal{B}_{n2} := \frac{1}{n^{1/2}h^2}\sum_{i=1}^n c_i J_{n2} = \frac{1}{n^{1/2}h^2}\sum_{i=1}^n c_i\Big(h^2 V_{n1} + h\big(V_{n2} + W_n\big) + h^2\mathcal{D}_n\Big)$$
$$= \frac{1}{n^{1/2}}\sum_{i=1}^n c_i\Big(V_{n1} + \frac{1}{h}\big(V_{n2} + W_n\big) + \mathcal{D}_n\Big).$$

Therefore, by using (11.9), (11.21), (11.56), and (11.58)–(11.60), we obtain

$$\frac{n^{1/2}}{nh^2}\Delta'\mathcal{B}_{n2} - \frac{\Delta'}{n^{1/2}}\sum_{i=1}^{n} c_i\mathcal{D} = \frac{\Delta'}{n^{1/2}}\sum_{i=1}^{n} c_i\Big\{V_{n1} + \frac{1}{h}\big(V_{n2} + W_n\big) + \mathcal{D}_n - \mathcal{D}\Big\},$$

$$\begin{aligned}
&\Big|\frac{n^{1/2}}{nh^2}\Delta'\mathcal{B}_{n2} - \frac{\Delta'}{n^{1/2}}\sum_{i=1}^{n} c_i\mathcal{D}\Big| \\
&\quad \le \|\Delta'\|\Big\|\frac{1}{n^{1/2}}\sum_{i=1}^{n} c_i\Big\|\Big[\big|V_{n1}\big| + \frac{1}{h}\big|V_{n2}\big| + \frac{1}{h}\big|W_n\big| + \big|\mathcal{D}_n - \mathcal{D}\big|\Big] \\
&\quad = O_p(1)\Big[O_p\big(\frac{1}{n^{1/2}}\big) + O_p\big(\frac{h^{1/2}}{n^{1/2}}\big) + O_p\big(h\big) + O_p\big(h\big)\Big] \\
&\quad = O_p(h) = o_p(1).
\end{aligned} \tag{11.64}$$

By (11.64), (11.63) and the triangle inequality, we now readily obtain

$$\begin{aligned}
&\Big|\frac{n^{1/2}}{nh^2}\Delta' D_{n12} - \frac{\Delta'}{n^{1/2}}\sum_{i=1}^{n} c_i\mathcal{D}\Big| \\
&\quad = \Big|\frac{1}{n^{1/2}h^2}\Delta'\mathcal{B}_{n1} + \frac{n^{1/2}}{nh^2}\Delta'\mathcal{B}_{n2} - \frac{\Delta'}{n^{1/2}}\sum_{i=1}^{n} c_i\mathcal{D}\Big| = o_p(1).
\end{aligned}$$

Combine this fact with (11.62) and (11.50) to obtain

$$\begin{aligned}
&\Big|\frac{n^{1/2}}{nh^2}D_{n1} - \Big(n^{1/2}\delta_1 + \frac{\Delta'}{n^{1/2}}\sum_{i=1}^{n} c_i\Big)\mathcal{D}\Big| \\
&\quad = \Big|\frac{n^{1/2}}{nh^2}\delta_1 D_{n11} - n^{1/2}\delta_1\mathcal{D} + \frac{n^{1/2}}{nh^2}\Delta' D_{n12} - \frac{\Delta'}{n^{1/2}}\sum_{i=1}^{n} c_i\mathcal{D}\Big| = o_p(1).
\end{aligned}$$

In turn, this fact combined with (11.49) and (11.48) now readily yields (11.46). This also completes the proof of Lemma 11.2. □

The asymptotic normality result (11.7), decompositions (11.15), (11.45), and (11.46) readily yield the following theorem.

Theorem 11.2 *Assume the model* (11.1) *with f as in H_1 and Assumptions* (11.8)–(11.12) *hold. Then, under H_1,*

$$\Big|n^{1/2}(\widehat{T}_n - T_n) - \Big(n^{1/2}\delta_1 + \frac{\Delta'}{n^{1/2}}\sum_{i=1}^{n} c_i\Big)\mathcal{D}\Big| = o_p(1).$$

If, in addition, K is bounded and compactly supported and $\ddot{f}$ is bounded, then, under H_1,

$$n^{1/2}\Big(\widehat{T}_n - \int \big\{K_h * (f - f_0)(y)\big\}^2 dy\Big) - \Big(n^{1/2}\delta_1 + \frac{\Delta'}{n^{1/2}}\sum_{i=1}^{n} c_i\Big)\mathcal{D} \to_D N(0, 4\omega^2),$$

where $\omega^2 := Var_f\big(f(\varepsilon) - f_0(\varepsilon)\big)$.

11.4 Random Covariates

In this section, we shall discuss the analogs of the above results when the predicting variables in the model (11.1) are random. Accordingly, our model now is

$$Y_i = \alpha + Z_i'\beta + \varepsilon_i, \qquad 1 \le i \le n, \tag{11.65}$$

where $Z, Z_i \in \mathbb{R}^p$, $1 \le i \le n$ are i.i.d. random vectors and independent of $\{\varepsilon, \varepsilon_i\}$. We need to replace Assumptions (11.8) and (11.9) by the following two assumptions. Let $\widetilde{\alpha}, \widetilde{\beta}$ be estimates of α, β of the model (11.65), respectively. Assume

$$E\|Z\|^4 < \infty, \tag{11.66}$$

$$n^{1/2}\big|\widetilde{\alpha} - \alpha\big| + n^{1/2}\|\widetilde{\beta} - \beta\| = O_p(1). \tag{11.67}$$

Let $\mathcal{Z}$ denote the $n \times p$ matrix whose ith row is Z_i', $1 \le i \le n$. Note that $\sum_{i=1}^n Z_i Z_i' = \mathcal{Z}'\mathcal{Z}$. By the Law of Large Numbers (LLNs), under (11.66), $n^{-1}\mathcal{Z}'\mathcal{Z} = n^{-1}\sum_{i=1}^n Z_i Z_i' \to_p \Sigma := E\big(ZZ'\big)$.

Now the residuals are $\widetilde{\varepsilon}_i := Y_i - \widetilde{\alpha} - \widetilde{\beta}' Z_i$. Let $\widetilde{T}_n$ denote the analog of $\widehat{T}_n$ with $\widehat{\varepsilon}_i$ replaced by $\widetilde{\varepsilon}_i$. This is the test statistic in the case of random covariates. The asymptotic distributions of $\widetilde{T}_n$ under H_0 and H_1 are described in the next two subsections.

11.4.1 Asymptotic Null Distribution of $\widetilde{T}_n$

The following theorem describes the asymptotic null distribution of $\widetilde{T}_n$.

Theorem 11.3 *Suppose Model* (11.65) *holds and Assumption* (11.10) *with* $f = f_0$, (11.11), (11.12), (11.66), *and* (11.67) *when* $f = f_0$ *hold. Then, under* H_0, *the analog of* (11.13) *with* $\widehat{T}_n$ *replaced by* $\widetilde{T}_n$ *continues to hold.*

If, in addition, $\ddot{f}_0$ *is bounded and* K *is compactly supported and bounded, then, under* H_0, (11.14) *holds with* $\widehat{T}_n$ *replaced by* $\widetilde{T}_n$.

Proof The proof is quite similar to that of Theorem 11.1. We shall thus be brief, indicating only the differences.

Proof of the analog of (11.16). Let $\widetilde{\delta}_1 := \widetilde{\alpha} - \alpha$, $\widetilde{\delta}_2 := \widetilde{\beta} - \beta$ and let $\eta_i = \widetilde{\delta}_1 + \widetilde{\delta}_2' Z_i$. Then $\varepsilon_i - \widetilde{\varepsilon}_i = \widetilde{\delta}_1 + \widetilde{\delta}_2' Z_i = \eta_i$, $1 \le i \le n$. The role played by η_i here is similar to that played by ξ_i in the previous sections.

By (11.66) and (11.67), $n^{-1}\sum_{i=1}^n \|Z_i\|^k \to_p E\|Z\|^k < \infty$, $k = 1, \ldots, 4$ and

$$\max_{1\le i\le n}|\eta_i| \le |\widetilde{\delta}_1| + n^{1/2}\|\widetilde{\delta}_2\| \max_{1\le i\le n} n^{-1/2}\|Z_i\| = o_p(1),$$

$$n^{-1/2}\sum_{i=1}^n |\eta_i| \le n^{1/2}|\widetilde{\delta}_1| + n^{1/2}\|\widetilde{\delta}_2\| n^{-1}\sum_{i=1}^n \|Z_i\| = O_p(1).$$

$$\sum_{i=1}^n \eta_i^2 \le 2n\widetilde{\delta}_1^2 + 2n\|\widetilde{\delta}_2\|^2 n^{-1}\sum_{i=1}^n \|Z_i\|^2 = O_p(1),$$

$$n\sum_{i=1}^n \eta_i^4 = n\sum_{i=1}^n \Big(\widetilde{\delta}_1 + \widetilde{\delta}_2' Z_i\Big)^4 \le 8n\Big(n\widetilde{\delta}_1^4 + \sum_{i=1}^n (\widetilde{\delta}_2' Z_i)^4\Big)$$

$$\le 8n^2\widetilde{\delta}_1^4 + 8n^2\|\widetilde{\delta}_2\|^4 n^{-1}\sum_{i=1}^n \|Z_i\|^4 = O_p(1).$$

Recall the bound (11.20). Let $\widetilde{C}_{n1}, \widetilde{C}_{n2}$ be the analogs of C_{n1}, C_{n2} with ξ_i and x_i replaced by η_i and Z_i, respectively. That is, with $\widetilde{\varepsilon}_i = Y_i - \widetilde{\alpha} - \widetilde{\beta}' Z_i$,

$$\widetilde{C}_{n1} := \int \Big(\sum_{i=1}^n \eta_i \dot{K}\big(\frac{y-\varepsilon_i}{h}\big)\Big)^2 dy,$$

$$\widetilde{C}_{n2} := \int \Big(\sum_{i=1}^n \int_{(y-\varepsilon_i)/h}^{(y-\widetilde{\varepsilon}_i)/h} \ddot{K}(t)\Big(\frac{y-\varepsilon_i}{h} - t\Big)dt\Big)^2 dy.$$

Then the bound (11.20) continues to hold with C_{n1}, C_{n2} replaced by $\widetilde{C}_{n1}, \widetilde{C}_{n2}$, respectively.

Let $\widetilde{C}_{n12}$ denote the C_{n12} of (11.22) with c_i replaced by Z_i, i.e.,

$$\widetilde{C}_{n12} := \int \Big\|\sum_{i=1}^n Z_i \dot{K}\big(\frac{y-\varepsilon_i}{h}\big)\Big\|^2 dy.$$

Then, with C_{n11} as in (11.22), we have $\widetilde{C}_{n1} \le 2\widetilde{\delta}_1^2 C_{n11} + 2\|\widetilde{\delta}_2\|^2 \widetilde{C}_{n12}$. Since C_{n11} does not involve Z_i's, the bound (11.26) holds here also, i.e., $\widetilde{\delta}_1^2 C_{n11} = O_p\big(h\big) + O_p\big(nh^4\big)$. Then,

$$\widetilde{C}_{n12} = \int \Big\|\sum_{i=1}^n Z_i\Big(\dot{K}\big(\frac{y-\varepsilon_i}{h}\big) - E\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big) + \sum_{i=1}^n Z_i E\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big\|^2 dy$$

$$\le 2\int \Big\|\sum_{i=1}^n Z_i\Big(\dot{K}\big(\frac{y-\varepsilon_i}{h}\big) - E\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big)\Big\|^2 dy$$

$$+ 2\int \Big\|\sum_{i=1}^n Z_i E\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big\|^2 dy.$$

By the independence of $\{Z_i\}$ and $\{\varepsilon_i\}$ and Fubini's theorem, we have

$$\begin{aligned}
&E\int \Big\| \sum_{i=1}^{n} Z_i\Big(\dot{K}\big(\frac{y-\varepsilon_i}{h}\big) - E\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big)\Big\|^2 dy \\
&= nE\|Z\|^2 \int \mathrm{Var}_0\big(\dot{K}\big(\frac{y-\varepsilon}{h}\big)\big) \le nE\|Z\|^2 E\dot{K}^2\big(\frac{y-\varepsilon}{h}\big) \\
&= nhE\|Z\|^2 \int \dot{K}^2(v)dv = O(nh). \qquad (11.68)
\end{aligned}$$

Hence,

$$\begin{aligned}
E\widetilde{C}_{n12} &\le 2nE\|Z\|^2 \int E\dot{K}^2\big(\frac{y-\varepsilon}{h}\big) + E\Big\|\sum_{i=1}^{n} Z_i\Big\|^2 \int \Big(E\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big)^2 \\
&\le 2nE\|Z\|^2 h \int \dot{K}^2(v)dv + 2n^2 E\|Z\|^2 O\big(h^4\big), \\
\big\|\delta_2\big\|^2 \widetilde{C}_{n12} &= n\big\|\delta_2\big\|^2 n^{-1}\widetilde{C}_{n12} = O_p(1)\Big(O_p\big(h\big) + O_p\big(nh^4\big)\Big), \\
nh^{1/2}\frac{1}{n^2h^4}\widetilde{C}_{n1} &= \frac{1}{nh^{7/2}}\Big(O_p\big(h\big) + O_p\big(nh^4\big)\Big) = O_p\big(\frac{1}{nh^{5/2}}\big) + O_p\big(h^{1/2}\big).
\end{aligned}$$

This bound is similar to the rate bound given at (11.28).

Analogous to (11.29), we have

$$\widetilde{C}_{n2} \le n\sum_{i=1}^{n} \frac{\eta_i^4}{h^3} \int \ddot{K}^2(v)dv = O_p\big(\frac{1}{h^3}\big), \quad nh^{1/2}\frac{1}{n^2h^4}\widetilde{C}_{n2} = O_p\big(\frac{1}{nh^{9/2}}\big). \qquad (11.69)$$

These bounds combined with (11.20) complete the proof of (11.16) in the random covariates case.

Proof of (11.17). Let $\widetilde{D}_{n1}$ and $\widetilde{D}_{n2}$ denote the analogs of D_{n1}, D_{n2} of the identity (11.31) in the current setup, respectively. Then, similar to (11.31), we have the following decomposition in the current setup:

$$\int \big(\widehat{f}_n(y) - f_n(y)\big)\rho_n(y)dy = \frac{1}{nh^2}\widetilde{D}_{n1} + \frac{1}{nh}\widetilde{D}_{n2}. \qquad (11.70)$$

Similar to (11.32),

$$\big|\widetilde{D}_{n2}\big| \le \widetilde{C}_{n2}^{1/2} T_n^{1/2} = O_p\Big(\frac{1}{n^{1/2}h^2}\Big), \quad nh^{1/2}\frac{1}{nh}\big|\widetilde{D}_{n2}\big| = O_p\Big(\frac{1}{n^{1/2}h^{5/2}}\Big).$$

To analyze $\widetilde{D}_{n1}$, let

$$\widetilde{D}_{n12} := \int \sum_{i=1}^{n} Z_i \dot{K}\big(\frac{y-\varepsilon_i}{h}\big)\rho_n(y)dy.$$

Then, with D_{n11} as in (11.33), analogous to (11.34), we have

$$\widetilde{D}_{n1} = \widetilde{\delta}_1 D_{n11} + \widetilde{\delta}_2' \widetilde{D}_{n12}. \tag{11.71}$$

Since D_{n11} does not involve Z_i's, the result (11.44) is applicable in the current setup also, i.e., we have

$$nh^{1/2}\frac{1}{nh^2}\big|\widetilde{\delta}_1 D_{n11}\big| = O_p\Big(\frac{1}{n^{1/2}h^{3/2}}\Big) + O_p\Big(h^{1/2}\Big).$$

Also, we write

$$\begin{aligned}\widetilde{D}_{n12} &= \int \sum_{i=1}^{n} Z_i \Big(\dot{K}\big(\frac{y-\varepsilon_i}{h}\big) - E_0\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big)\rho_n(y)dy \\ &\quad + \int \sum_{i=1}^{n} Z_i E_0\Big(\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big)\rho_n(y)dy \\ &= \widetilde{B}_{n1} + \widetilde{B}_{n2}, \qquad \text{say.}\end{aligned}$$

Note that $\widetilde{B}_{n2} = \sum_{i=1}^{n} Z_i\Gamma_{n2}$, where Γ_{n2} is as in (11.35). Thus, in view of (11.67) and (11.43),

$$\begin{aligned}\big|\widetilde{\delta}_2'\widetilde{B}_{n2}\big| &= \big|\sum_{i=1}^{n}\widetilde{\delta}_2' Z_i\big|\big|\Gamma_{n2}\big| \le \sum_{i=1}^{n}\big|\widetilde{\delta}_2' Z_i\big|\big|\Gamma_{n2}\big| \\ &\le n^{1/2}\|\widetilde{\delta}_2\| n^{-1/2}\sum_{i=1}^{n}\|Z_i\|\big|\Gamma_{n2}\big| = O_p(1)O_p\big(n^{1/2}\big)O_p\Big(\frac{h^2}{n^{1/2}}\Big) = O_p(h^2).\end{aligned}$$

Therefore, by (11.67) and (11.30), similar to (11.27),

$$\begin{aligned}\big|\widetilde{\delta}_2'\widetilde{B}_{n1}\big| &\le \|\widetilde{\delta}_2\|\,\|\widetilde{B}_{n1}\| \\ &\le \|\widetilde{\delta}_2\|\Big(\int\Big\|\sum_{i=1}^{n} Z_i\Big(\dot{K}\big(\frac{y-\varepsilon_i}{h}\big) - E_0\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big)\Big\|^2 dy\Big)^{1/2} T_n^{1/2} \\ &= O_p\big(n^{-1/2}\big)O_p\big((nh)^{1/2}\big)O_p\big((nh)^{-1/2}\big) = O_p\big(n^{-1/2}\big).\end{aligned}$$

Hence,

$$nh^{1/2}\frac{1}{nh^2}\big|\widetilde{\delta}_2' D_{n12}\big| \le \frac{1}{h^{3/2}}\Big(\big|\widetilde{\delta}_2'\widetilde{B}_{n1}\big| + \big|\widetilde{\delta}_2'\widetilde{B}_{n2}\big|\Big) = O_p\Big(\frac{1}{n^{1/2}h^{3/2}}\Big) + O_p\big(h^{1/2}\big),$$
$$nh^{1/2}\frac{1}{nh^2}\big|\widetilde{D}_{n1}\big| = O_p\Big(\frac{1}{n^{1/2}h^{3/2}}\Big) + O_p\Big(h^{1/2}\Big),$$

just as in non-random covariates case. This also completes the proof of (11.17) in the random covariates case, and also of Theorem 11.3. □

Remark 11.1 To test the composite hypothesis $H_\sigma : f(x) = \sigma^{-1} f_0(x/\sigma)$, for all $x \in \mathbb{R}$ and some $\sigma > 0$, modify the above test as follows. Let $\hat{\sigma}$ ($\widetilde{\sigma}$) be an estimator of σ in the non-random (random) predictors case such that $\big|n^{1/2}(\hat{\sigma} - \sigma)\big| = O_p(1)$, ($\big|n^{1/2}(\widetilde{\sigma} - \sigma)\big| = O_p(1)$), under H_σ. Let

$$\widehat{\mu}_n(y) := \frac{1}{\hat{\sigma}}\int\int f_0\Big(\frac{y - hv}{\hat{\sigma}}\Big)K(v)dvdy,$$
$$\widetilde{\mu}_n(y) := \frac{1}{\widetilde{\sigma}}\int\int f_0\Big(\frac{y - hv}{\widetilde{\sigma}}\Big)K(v)dvdy,$$
$$T_n^* := \int \big(\widehat{f}_n(y) - \widehat{\mu}_n(y)\big)^2 dy \quad \text{or} \quad \int \big(\widetilde{f}_n(y) - \widetilde{\mu}_n(y)\big)^2 dy.$$

Then, arguing as in [9], one shows that

$$nh^{1/2}\Big(T_n^* - \frac{1}{nh}\int K^2(v)dv\Big) \to_D N\big(0, \tau^2\big),$$

in both non-random and random predictor cases.

11.4.2 Asymptotic Distribution of $\widetilde{T}_n$ Under H_1

In this subsection, we shall prove the analog of Lemma 11.2 and Theorem 11.2 in the linear regression model (11.65) with random predictors. First we shall prove the following lemma.

Lemma 11.3 *Assume H_1 and Assumptions* (11.11)*,* (11.12)*,* (11.66)*, and* (11.67) *hold. Then,*

$$\Big|n^{1/2}\int \big(\widetilde{f}_n(y) - f_n(y)\big)\big(f_n(y) - \mu_n(y)\big)dy - \Big(n^{1/2}\widetilde{\delta}_1 + n^{1/2}\widetilde{\delta}_2' E(Z)\Big)\mathcal{D}\Big| = o_p(1).$$

Proof Again, the proof of this lemma is similar to that of Lemma 11.2. We shall be brief, indicating only the differences due to random covariates.

To start with recall the decomposition (11.70). In view of (11.47) and (11.69), we have

$$\big|\widetilde{D}_{n2}\big| \le \widetilde{C}_{n2}^{1/2} T_n^{1/2} = O_p\big(h^{-3/2}\big), \qquad n^{1/2}\frac{1}{nh}\big|\widetilde{D}_{n2}\big| = O_p\Big(\frac{1}{n^{1/2}h^{5/2}}\Big).$$

To analyze $\widetilde{D}_{n1}$ recall the identity (11.71). Since D_{n11} does not involve Z_i's, the analog of the result (11.62) is applicable in the current setup also, i.e.,

$$\big|\frac{n^{1/2}}{nh^2}\widetilde{\delta}_1 D_{n11} - n^{1/2}\widetilde{\delta}_1 \mathcal{D}\big| = o_p(1).$$

To analyze $\widetilde{D}_{n12}$, we write

$$\begin{aligned}
\widetilde{D}_{n12} &= \int \sum_{i=1}^n Z_i\Big(\dot{K}\big(\frac{y-\varepsilon_i}{h}\big) - E\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big)\rho_n(y)dy \\
&\quad + \int \sum_{i=1}^n Z_i E\Big(\dot{K}\big(\frac{y-\varepsilon}{h}\big)\Big)\rho_n(y)dy \\
&= \widetilde{G}_n + \sum_{i=1}^n Z_i J_{n2}, \qquad \text{say.}
\end{aligned}$$

Note that $\widetilde{G}_n$ is like $\mathcal{B}_{n1}$ defined in the proof of (11.46). In view of (11.47) and (11.68),

$$\|\widetilde{G}_n\| = O_p\big((nh)^{1/2}\big), \quad n^{1/2}\frac{1}{nh^2}\big|\widetilde{\delta}_2'\widetilde{G}_n\big| \le n^{1/2}\|\widetilde{\delta}_2\|\frac{1}{nh^2}\|\widetilde{G}_n\| = O_p\big(\frac{1}{n^{1/2}h^{3/2}}\big).$$

In view of (11.56) and (11.58)–(11.61),

$$\begin{aligned}
&\Big|n^{1/2}\frac{1}{nh^2}\widetilde{\delta}_2'\sum_{i=1}^n Z_i\, J_{n2} - n^{1/2}\widetilde{\delta}_2' n^{-1}\sum_{i=1}^n Z_i\mathcal{D}\Big| \\
&= \Big|\frac{1}{nh^2}n^{1/2}\widetilde{\delta}_2'\sum_{i=1}^n Z_i\Big(h^2 V_{n1} + hV_{n2} + hW_n + h^2\mathcal{D}_n\Big) - n^{1/2}\widetilde{\delta}_2' n^{-1}\sum_{i=1}^n Z_i\mathcal{D}\Big| \\
&= \Big|n^{1/2}\widetilde{\delta}_2' n^{-1}\sum_{i=1}^n Z_i\Big(V_{n1} + h^{-1}\big(V_{n2} + W_n\big) + \mathcal{D}_n - \mathcal{D}\Big)\Big| \\
&\le \|n^{1/2}\widetilde{\delta}_2\|\;\|n^{-1}\sum_{i=1}^n Z_i\|\Big\{O_p\big(n^{-1/2}\big) + O_p\big(\frac{h^{1/2}}{n^{1/2}}\big) + O_p\big(h^3\big) + O(h)\Big\} \\
&= o_p(1),
\end{aligned}$$

by (11.12), (11.67) and the fact that $n^{-1}\sum_{i=1}^n Z_i \to_p E(Z)$. This completes the proof of Lemma 11.3. □

Analogous to Theorem 11.2 we have the following theorem in the random predictors case.

Theorem 11.4 *Assume the model* (11.65) *with the error density* f *as in* H_1 *holds. In addition, assume that* (11.10)–(11.12), (11.66) *and* (11.67) *hold. Under* H_1,

$$\left| n^{1/2}\big(\widetilde{T}_n - T_n\big) - \Big(n^{1/2}\widetilde{\delta}_1 + n^{1/2}\widetilde{\delta}_2' E(Z)\Big)\mathcal{D}\right| = o_p(1).$$

If, in addition, $\ddot{f}$ *is bounded and* K *is compactly supported and bounded, then*

$$n^{1/2}\Big(\widetilde{T}_n - \int \{K_h * (f - f_0)(y)\}^2 dy\Big) - \Big(n^{1/2}\widetilde{\delta}_1 + n^{1/2}\widetilde{\delta}_2' E(Z)\Big)\mathcal{D} \to_D N(0, 4\omega^2).$$

Remark 11.2 The papers [2, 9] dealt with the zero intercept linear autoregressive stationary time series $\{Z_i, i = 0, \pm 1, \cdots\}$ having $E(Z_0) = 0$. In our case, we have nonzero intercept parameter in both non-random and random covariates linear regression models. This is the reason for the difference between the limiting distribution of $\widehat{T}_n$ in the present setup and that of its analog in [2] under the global L_2 alternatives H_1. It is also the reason for us to assume that $h \to 0, nh^5 \to \infty$ compared to the assumption $h \to 0, nh^4 \to \infty$ needed by [2, 9].

However, if both f_0 and $f(\neq f_0)$ satisfy (11.10) and are such that $\mathcal{D} = \int \dot{f}(y)\big(f(y) - f_0(y)\big)dy = 0$, then the results of Theorems 11.2 and 11.4 are the same as those obtained in [2] in the linear autoregressive models with zero intercept. In particular, if both f, f_0 are symmetric around zero, then $\dot{f}(-y) \equiv -\dot{f}(y)$ and

$$\begin{aligned}\mathcal{D} &= \int_{-\infty}^{0} \dot{f}(y)\big(f(y) - f_0(y)\big)dy + \int_0^{\infty} \dot{f}(y)\big(f(y) - f_0(y)\big)dy \\ &= -\int_0^{\infty} \dot{f}(y)\big(f(y) - f_0(y)\big)dy + \int_0^{\infty} \dot{f}(y)\big(f(y) - f_0(y)\big)dy = 0.\end{aligned}$$

11.5 Finite Sample Simulations and A Real Data Example

This section contains findings of a simulation study and some real data examples. The finite sample performance of the goodness-of-fit tests based on $\widetilde{T}_n$ and T_n^* of Remark 11.1 is described in Sect. 11.5.1 while the two real data examples appear in Sect. 11.5.2.

11.5.1 Finite Sample Simulations

In this subsection, we describe the findings of a finite sample study using Monte Carlo and asymptotic distributions of $\widetilde{T}_n$ in the random design case.

For each value of $n = 250, 500$, and 1000, we generated Z_i's randomly from Uniform$(-3, 3)$ distribution, and ε_i's randomly from the density f_0, independent of Z_i's. Then the observations $\{Y_1, \ldots, Y_n\}$ were generated using the model (11.65), with $\alpha = 10$, $\beta = 5$. We used the three densities: standard normal density φ_N, Student-t_5 density ψ_{t_5} with degrees of freedom 5 and logistic density $\psi_L(x) = e^{-x}/(1+e^{-x})^2$, $x \in \mathbb{R}$. Let φ_{t_5}, φ_L denote the standardized ψ_{t_5}, ψ_L densities, each having unit variance. Recall that the variance of ψ_{t_5} is $5/3$ and that of ψ_L is $\pi^2/3$. Thus $\varphi_{t_5}(x) = (3/5)^{1/2}\psi_{t_5}(x/(5/3)^{1/2})$ and $\varphi_L(x) := (\pi/\sqrt{3})\psi_L\big(\pi x/\sqrt{3}\big)$, $x \in \mathbb{R}$.

We first consider the following two simple null hypotheses:

$$H_{0N}^* : f(x) = \varphi_N(x), \quad \forall x \in \mathbb{R}; \qquad H_{0L}^* : f(x) = \varphi_L(x), \quad \forall x \in \mathbb{R},$$

each against the alternative that the given null hypothesis is not true.

We used the kernel $K(x) = (1/0.4439938)\exp(-1/(1-x^2))I(|x| \le 1)$. The properties that for any $r > 0$, $\sup_{y>0}\big(y^r e^{-y}\big) < \infty$ and that $y^r e^{-y} \to 0$, as $y \to \infty$, are used to show that this K satisfies the assumption (11.11). Moreover,

$$\begin{aligned}
&\int K^2(v)dv = 0.675117, \qquad \int (K*K)^2(v)dv = 0.487467,\\
&\int v^2 K(v)dv = 0.020155, \qquad \sigma^{-2}\int \varphi_N^2(x/\sigma)dy = 0.2820948/\sigma,\\
&c^{-2}\int \varphi_L^2(x/c)dx = 1\big/(6c), \quad \forall \sigma > 0, c > 0.
\end{aligned} \tag{11.72}$$

For the Monte Carlo study, the simulation was carried out as follows. For each iteration, the parameters α and β were estimated by the least squares estimators $\widetilde{\alpha}$, $\widetilde{\beta}$, respectively. Let $\widetilde{\varepsilon}_i = Y_i - \widetilde{\alpha} - \widetilde{\beta} Z_i$, $1 \le i \le n$ and

$$\widetilde{f}_n(y) = \frac{1}{nh}\sum_{i=1}^n K\Big(\frac{y - \widetilde{\varepsilon}_i}{h}\Big).$$

Let μ_{nN}, μ_{nL} denote μ_n of (11.2) when $f_0 = \varphi_N$, $f_0 = \varphi_L$, respectively, and let

$$\begin{aligned}
T_{nN} &:= \int \Big(\widetilde{f}_n(y) - \mu_{nN}(y)\Big)^2 dy, \qquad T_{nL} := \int \Big(\widetilde{f}_n(y) - \mu_{nL}(y)\Big)^2 dy,\\
\tau_{nN}^2 &= 2\int \varphi_N^2(z)dz \int \big(K*K(y)\big)^2 dy, \ \ \tau_{nL}^2 = 2\int \varphi_L^2(z)dz \int \big(K*K(y)\big)^2 dy,\\
M_{nN} &= \frac{n\sqrt{h}}{\tau_{nN}}\Big(T_{nN} - \frac{1}{nh}\int K^2(v)dv\Big), \quad M_{nL} = \frac{n\sqrt{h}}{\tau_{nL}}\Big(T_{nL} - \frac{1}{nh}\int K^2(v)dv\Big).
\end{aligned}$$

As in [9], for testing the above simple hypotheses we used the bandwidth $h = 1/(3n^{1/5.1})$. The large values of $|M_{nN}|$ ($|M_{nL}|$) are significant for H_{0N}^* (H_{0L}^*).

Next, we consider the composite hypotheses

$$H_{0N} : f(x) = \frac{1}{\sigma}\varphi_N\Big(\frac{x}{\sigma}\Big), \quad \exists\, \sigma > 0,\ \forall\, x \in \mathbb{R};$$
$$H_{0L} : f(x) = \frac{1}{c}\psi_L\Big(\frac{x}{c}\Big), \quad \exists\, c > 0,\ \forall\, x \in \mathbb{R},$$

each against the alternative that the null is not true.

Let s denote the sample standard deviation of $\widetilde{\varepsilon}_i$, $1 \le i \le n$, $\widetilde{c} = \sqrt{3}s/\pi$ and $h_{nN} = s\big/(3n^{1/5.1})$ and $h_{nL} = \widetilde{c}\big/(3n^{1/5.1})$. These choices of the bandwidth sequences are justified from some optimality considerations, see, e.g., [11]. Let

$$\widetilde{f}_{nN}(y) := \frac{1}{nh_{nN}}\sum_{i=1}^{n} K\Big(\frac{y-\widetilde{\varepsilon}_i}{h_{nN}}\Big), \quad \widetilde{f}_{nL}(y) := \frac{1}{nh_{nL}}\sum_{i=1}^{n} K\Big(\frac{y-\widetilde{\varepsilon}_i}{h_{nL}}\Big),$$
$$\widetilde{\mu}_{nN}(y) := \frac{1}{h_{nN}}\int K\Big(\frac{y-z}{h_{nN}}\Big)\frac{1}{s}\varphi_N\Big(\frac{z}{s}\Big)dz,$$
$$\widetilde{\tau}^2_{nN} := 2\int \frac{1}{s^2}\varphi^2_N\Big(\frac{z}{s}\Big)dz \int \big(K * K(y)\big)^2 dy,$$
$$\widetilde{\mu}_{nL}(y) := \frac{1}{h_{nL}}\int K\Big(\frac{y-z}{h_{nL}}\Big)\frac{1}{\widetilde{c}}\psi_L\Big(\frac{z}{\widetilde{c}}\Big)dz,$$

and

$$\widetilde{\tau}^2_{nL} := 2\int \frac{1}{\widetilde{c}^2}\psi^2_L\Big(\frac{z}{\widetilde{c}}\Big)dz \int \big(K * K(y)\big)^2 dy,$$
$$\widetilde{T}_{nN} := \int \Big(\widetilde{f}_{nN}(y) - \widetilde{\mu}_{nN}(y)\Big)^2 dy, \quad \widetilde{T}_{nL} := \int \Big(\widetilde{f}_{nL}(y) - \widetilde{\mu}_{nL}(y)\Big)^2 dy,$$
$$\widetilde{M}_{nN} := \frac{n\sqrt{h_{nN}}}{\widetilde{\tau}_{nN}}\Big(\widetilde{T}_{nN} - \frac{1}{nh_{nN}}\int K^2(v)dv\Big),$$
$$\widetilde{M}_{nL} := \frac{n\sqrt{h_{nL}}}{\widetilde{\tau}_{nL}}\Big(\widetilde{T}_{nL} - \frac{1}{nh_{nL}}\int K^2(v)dv\Big). \tag{11.73}$$

The large values of $|\widetilde{M}_{nN}|$ $(|\widetilde{M}_{nL}|)$ are significant for H_{0N} (H_{0L}).

Let M_n be the generic notation for any one of the four statistics M_{nN}, M_{nL}, $\widetilde{M}_{nN}$, and $\widetilde{M}_{nL}$. For each iteration the statistic M_n was computed. The procedure was repeated 5,000 times collecting replication of M_n, denoted by $M_{n,i}$, $1 \le i \le 5000$. Also, denote the ordered statistics of $M_{n,i}$ by $M_{n,(i)}$. The asymptotic test with the asymptotic size α rejects the given null hypothesis when the corresponding $|M_n| > z_{\alpha/2}$, where z_α is the $100(1-\alpha)$th percentile of the standard normal distribution. The empirical test with size α rejects the null hypothesis when $M_n < M_{n,(5000[\alpha/2])}$ or $M_n > M_{n,(5000[1-\alpha/2])}$.

Tables 11.1 and 11.2 contain the empirical sizes and powers of these tests for the null hypotheses H_{0N}^* and H_{0L}^*, respectively. Tables 11.3 and 11.4 contain empirical sizes and powers of these tests for the composite null hypotheses H_{0N}, H_{0L}, respectively. In these tables, test based on the asymptotic (empirical) critical values of the corresponding M_n is denoted by $M_{n,a}$ ($M_{n,e}$). The left (right) half of the table is for the nominal level of $\alpha = 0.05$ (0.10). Overall, it is seen that the large sample tests tend to have conservative finite sample level when testing simple or composite hypotheses.

11.5.2 Real Data Examples

Here we shall illustrate the proposed test by applying it to two datasets. In the first example on energy, the objective was to regress the energy requirements (in

Table 11.1 Empirical sizes and powers of the tests for H_{0N}^*

n	$f\backslash M_n$	$\alpha = 0.05$		$\alpha = 0.10$	
		$M_{n,a}$	$M_{n,e}$	$M_{n,a}$	$M_{n,e}$
250	H_{0N}^*	0.030	0.050	0.074	0.100
	φ_{t_5}	0.226	0.256	0.301	0.351
	φ_L	0.076	0.096	0.128	0.164
500	H_{0N}^*	0.035	0.050	0.080	0.100
	φ_{t_5}	0.523	0.537	0.602	0.652
	φ_L	0.158	0.169	0.220	0.262
1000	H_{0N}^*	0.032	0.050	0.084	0.100
	φ_{t_5}	0.906	0.916	0.935	0.953
	φ_L	0.332	0.359	0.417	0.480

Table 11.2 Empirical sizes and powers of the tests for H_{0L}^*

n	$f\backslash M_n$	$\alpha = 0.05$		$\alpha = 0.10$	
		$M_{n,a}$	$M_{n,e}$	$M_{n,a}$	$M_{n,e}$
250	φ_N	0.064	0.082	0.112	0.152
	φ_{t_5}	0.047	0.065	0.091	0.120
	H_{0L}^*	0.029	0.050	0.074	0.100
500	φ_N	0.127	0.137	0.185	0.223
	φ_{t_5}	0.070	0.080	0.115	0.139
	H_{0L}^*	0.032	0.050	0.080	0.100
1000	φ_N	0.328	0.324	0.408	0.451
	φ_{t_5}	0.103	0.105	0.169	0.199
	H_{0L}^*	0.036	0.050	0.075	0.100

Table 11.3 Empirical sizes and powers of the tests for H_{0N}

n	$f\backslash\widetilde{M}_n$	$\alpha = 0.05$		$\alpha = 0.10$	
		$\widetilde{M}_{n,a}$	$\widetilde{M}_{n,e}$	$\widetilde{M}_{n,a}$	$\widetilde{M}_{n,e}$
250	H_{0N}	0.035	0.050	0.082	0.100
	ψ_{t_5}	0.229	0.266	0.292	0.361
	ψ_L	0.068	0.090	0.116	0.162
500	H_{0N}	0.033	0.050	0.088	0.100
	ψ_{t_5}	0.477	0.533	0.554	0.640
	ψ_L	0.132	0.168	0.191	0.253
1000	H_{0N}	0.035	0.050	0.092	0.100
	ψ_{t_5}	0.864	0.883	0.902	0.928
	ψ_L	0.332	0.365	0.410	0.476

Table 11.4 Empirical sizes and powers of the tests for H_{0L}

n	$f\backslash\widetilde{M}_n$	$\alpha = 0.05$		$\alpha = 0.10$	
		$\widetilde{M}_{n,a}$	$\widetilde{M}_{n,e}$	$\widetilde{M}_{n,a}$	$\widetilde{M}_{n,e}$
250	φ_N	0.049	0.053	0.093	0.111
	ψ_{t_5}	0.060	0.067	0.109	0.126
	H_{0L}	0.033	0.050	0.086	0.100
500	φ_N	0.081	0.089	0.134	0.164
	ψ_{t_5}	0.069	0.079	0.122	0.145
	H_{0L}	0.039	0.050	0.088	0.100
1000	φ_N	0.200	0.222	0.280	0.309
	ψ_{t_5}	0.102	0.118	0.158	0.176
	H_{0L}	0.035	0.050	0.087	0.100

Mcal/day) (Y) on the body weights (in kg) (Z) of 64 grazing merino sheep in Australia. Identifying a relation between the energy requirement and the weight plays an important role in predicting meat production in grazing sheep systems. The data for this example is from [12], which is also available on p. 64 of [1]. We fitted a simple linear regression model by the method of least squares and computed the residuals to compute the proposed test. The fitted model is $\widetilde{Y} = 13.98 + 12.97Z$.

In the second example, the problem is to see if there is a relationship between the two methods of measuring iron content of crushed blast furnace slag. The variable Y is the measurement based on a chemical test conducted in the lab, which is time-consuming and expensive. The variable Z is the measurement based on a quicker and cheaper magnetic test. Dataset consisting of 53 observations was initially analyzed in [10]. It is also available on p. 62 of [1]. The simple linear regression model fitted to the data by the ordinary least squares is $\widetilde{Y} = 8.9565 + 0.5866Z$.

After fitting least squares regression lines to both datasets, we performed the goodness-of-fit tests for the error density. In both of these examples, the variable Z

is random, $\widetilde{\varepsilon}_i := Y_i - \widetilde{Y}_i$. Let φ_N and ψ_L denote the density of $N(0, 1)$ and logistic distributions, respectively, as in Sect. 11.5.1. In each data example mentioned above, we tested the two hypotheses H_{0N} and H_{0L} of Sect. 11.5.1 against the alternatives that the given null hypothesis is not true.

Recall the notation (11.73). With $\widetilde{c}$, K and s as in Sect. 11.5.1. From (11.72), we obtain

$$\widetilde{\tau}_{nN}^2 := 2s^{-2} \int \varphi_N^2(x/s)dx \int (K * K)^2(v)dv = 0.2750235\, s^{-1}, \quad \text{for } H_{0N},$$

$$\widetilde{\tau}_{nL}^2 := 2\widetilde{c}^{-2} \int \varphi_L^2\big(x/\widetilde{c}\big)dx \int (K * K)^2(v)dv = 0.1624888\, \widetilde{c}^{-1}, \quad \text{for } H_{0L}.$$

The test statistics to be used are $|\widetilde{M}_{nN}|$ and $|\widetilde{M}_{nL}|$ of (11.73). The test that rejects H_{0N} (H_{0L}) whenever $\big|\widetilde{M}_{nN}\big| > z_{\alpha/2}$ $\big(\big|\widetilde{M}_{nL}\big| > z_{\alpha/2}\big)$ is of the asymptotic size $0 < \alpha < 1$.

Example 11.1 Sheep energy versus weight. First, consider testing for H_{0N}. Here $n = 64$, The residual s.d. $s = 6.27965$, the bandwidth $h_{nN} = 0.9261072$, $\widetilde{T}_{nN} = 0.009118$, and $\widetilde{M}_{nN} = -0.6687283$. Since $|\widetilde{M}_{nN}|$ is small compared to any reasonable standard normal cut-off value, we do not reject the null hypothesis H_{0N} at any reasonable level of significance α. For example, if $\alpha = 0.05$, then $z_{\alpha/2} = 1.96$ and hence we can't reject the null hypothesis as $|M_n| < 1.96$ at $\alpha = 0.05$.

In the case of testing for H_{0L}, $\widetilde{c} = 3.462153$, the bandwidth $h_{nL} = 0.5105897$, $\widetilde{T}_{nL} = 0.01504718$, and $\widetilde{M}_{nL} = -1.184803$. Again we do not reject H_{0L} at any reasonable level.

Example 11.2 Iron content. Here $n = 53$. When testing for H_{0N}, $s = 3.4301$, $h_{nN} = 0.5249225$, $\widetilde{T}_{nN} = 0.01371428$, and $\widetilde{M}_{nN} = -1.430998$. When testing for H_{0L}, $\widetilde{c} = 1.8911$, $h_{nL} = 0.289405$, $\widetilde{T}_{nL} = 0.03934203$, and $\widetilde{M}_{nL} = -0.4545036$. Again, based on this data, neither of the two null hypotheses can be rejected at any reasonable level of significance.

No conflict of interest statement: On behalf of all authors, the corresponding author states that there is no conflict of interest.

Acknowledgements Authors thank the referee for detailed comments that helped to improve the presentation.

References

1. Abraham, B. and Ledolter, J. (2006). *Introduction to Regression Modeling*. Thomson Brooks/Cole, USA.
2. Bachmann, D. and Dette, H. (2005). A note on the Bickel–Rosenblatt test in autoregressive time series. *Statist. Probab. Lett.* **74** 221–234.
3. Bickel, P.J. and Rosenblatt, M. (1973). On some global measures of the deviations of density function estimates. *Ann. Statist.* **6** 1071–1095.

4. Cheng, F. and Sun, S. (2008). A goodness-of-fit test of the errors in nonlinear autoregressive time series models. *Statist. Probab. Lett.* **78** 50–59.
5. Hall, P. G. (1984). Central limit theorem for integrated square error of multivariate nonparametric density estimators. *J. Multivariate Anal.* **14** 1–16.
6. Khmaladze, E. V. and Koul, H. L. (2004). Martingale transforms goodness-of-fit tests in regression models. *Ann. Statist.* **32** 995–1034.
7. Khmaladze, E. V. and Koul, H. L. (2009). Goodness-of-fit problem for errors in nonparametric regression: distribution free approach. *Ann. Statist.* **37** 3165–3185.
8. Koul, H. L. and Mimoto, N. (2012). A goodness-of-fit test for GARCH innovation density. *Metrika* **75** 127–149.
9. Lee, S. and Na, S. (2002). On the Bickel–Rosenblatt test for first-order autoregressive models. *Statist. Probab. Lett.* **56** 23–35.
10. Roberts, H. V. and Ling, R. F. (1982). *Conversational Statistics with IDA*. Scientific Press/McGraw-Hill, New York.
11. Silverman B. W. (1998). *Density Estimation for Statistics and Data Analysis*. Chapman & Hall/CRC, Boca Raton, FL.
12. Wallach, D. and Goffinet, B. (1987). Mean square error of prediction in models for studying ecological and agronomic system. *Biometrics* **43** 561–573.

Chapter 12
A Minimum Contrast Estimation for Spectral Densities of Multivariate Time Series

Yan Liu

Abstract We propose a minimum contrast estimator for multivariate time series in the frequency domain. This extension has not been thoroughly investigated, although the minimum contrast estimator for univariate time series has been studied for a long time. The proposal in this paper is based on the prediction errors of parametric time series models. The properties of the proposed contrast estimation function are explained in detail. We also derive the asymptotic normality of the proposed estimator and compare the asymptotic variance with the existing results. The asymptotic efficiency of the proposed minimum contrast estimation is also considered. The theoretical results are illustrated by some numerical simulations.

12.1 Introduction

The estimation of stationary processes is an elementary problem in time series analysis. In the basic statistics course, the parameter estimation of time series is done in the time domain with the famous Yule–Walker estimator derived from the autoregressive models. As an exploration into the probabilistic structure, the maximum likelihood estimator is also introduced for the Gaussian autoregressive models. In view of its usefulness, the quasi-Gaussian maximum likelihood estimator plays a central role in the parameter estimation problem of time series. The limitations of all these approaches are realized when the Whittle likelihood is introduced in an intermediate course.

An innovative criterion between a parameterized model and a nonparametric estimator of the spectral density of stationary processes, concerning the stabilization of estimators under small perturbations, was introduced in [12] by Professor Taniguchi. This approach opened up a new horizon of the minimum contrast estimation in a series of works [13, 14, 16]. Under his supervision, we interpret the prediction errors, interpolation errors, and extrapolation errors as a measure of discrepancy between the model and a nonparametric estimator. It is no longer a surprise now that this type

Y. Liu (✉)
Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
e-mail: liu@waseda.jp

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_12

of measures works well for parameter estimation [6, 7, 11]. The minimum contrast estimation approach does not stop at its theory, but has applications in discriminant analysis [17], and empirical likelihood [9], just to name a few.

The only surprise to the author is that the approach was only developed for the scalar-valued linear processes. In this paper, we propose a minimum contrast estimator for the parameter estimation of spectral densities of multivariate time series based on the prediction errors [8]. The proposal is new but the mathematics is unexpectedly challenging. We focus on the main and intriguing cases in practice and establish the theoretical properties in asymptotics.

The rest of the paper is organized as follows. In Sect. 12.2, we propose our new measure for the multivariate time series, and establish its fundamental properties. In Sect. 12.3, we derive the asymptotic distribution of the proposed minimum contrast estimators. Section 12.4 discusses the asymptotic efficiency for multivariate Gaussian processes. The numerical study is presented in Sect. 12.5. Throughout this paper, we denote the set of all integers by $\mathbb{Z}$, and Kronecker's delta by

$$\delta(a, b) = \begin{cases} 1, & \text{if } a = b, \\ 0, & \text{if } a \neq b. \end{cases}$$

12.2 Contrast Function for Multivariate Time Series

Suppose $\{\boldsymbol{z}(t);\ t \in \mathbb{Z}\}$ is a vector-valued linear process generated as

$$\boldsymbol{z}(t) = \sum_{j=0}^{\infty} G(j)\boldsymbol{e}(t-j), \qquad t \in \mathbb{Z}, \tag{12.1}$$

where $\boldsymbol{z}(t), t \in \mathbb{Z}$, has q real components; $\boldsymbol{e}(t), t \in \mathbb{Z}$, has s real components such that $\mathbb{E}[\boldsymbol{e}(t)] = \boldsymbol{0}$ and $\mathbb{E}[\boldsymbol{e}(m)\boldsymbol{e}(n)^\top] = \delta(n, m)K$, with K a nonsingular $(s \times s)$-matrix. Notice that under the setting (12.1), $G(j)$, $j \in \mathbb{Z}$, is a $(q \times s)$-matrix; we assume all components of $G(j)$ are also real.

We suppose that the process $\{\boldsymbol{z}(t)\}$ is second-order stationary process. Assuming $\sum_{j=0}^{\infty} \mathrm{tr}\big[G(j)KG(j)^\top\big] < \infty$, the process has a spectral density matrix

$$\boldsymbol{f}(\omega) = \frac{1}{2\pi}\boldsymbol{k}(\omega)K\boldsymbol{k}(\omega)^*, \qquad \omega \in \Lambda := (-\pi, \pi],$$

where $\boldsymbol{k}(\omega) = \sum_{j=0}^{\infty} G(j)\mathrm{e}^{\mathrm{i}j\omega}$. Denote the (α, β)-th component of $G(j)$ by $G_{\alpha\beta}(j)$, and denote the αth component of $\boldsymbol{z}(t)$ and $\boldsymbol{e}(t)$ by $z_\alpha(t)$ and $e_\alpha(t)$, respectively. Under this setting, we assume that the process $\{\boldsymbol{e}(t)\}$ is fourth-order stationary with

$$\sum_{t_1,t_2,t_3=-\infty}^{\infty} |Q^e_{\alpha_1,\alpha_2,\alpha_3,\alpha_4}(t_1, t_2, t_3)| < \infty,$$

where $Q^e_{\alpha_1,\alpha_2,\alpha_3,\alpha_4}(t_1, t_2, t_3)$ is the joint fourth-order cumulant of $e_{\alpha_1}(t)$, $e_{\alpha_2}(t + t_1)$, $e_{\alpha_3}(t + t_2)$ and $e_{\alpha_4}(t + t_3)$, $1 \leq \alpha_1, \alpha_2, \alpha_3, \alpha_4 \leq s$. Henceforth, the process $\{\boldsymbol{e}(t)\}$ has a fourth-order spectral density

$$\begin{aligned}\tilde{Q}^e_{\alpha_1,\alpha_2,\alpha_3,\alpha_4}&(\omega_1, \omega_2, \omega_3)\\ &= \frac{1}{(2\pi)^3} \sum_{t_1,t_2,t_3=-\infty}^{\infty} \exp\Big(-\mathrm{i}(\omega_1 t_1 + \omega_2 t_2 + \omega_3 t_3)\Big) Q^e_{\alpha_1,\alpha_2,\alpha_3,\alpha_4}(t_1, t_2, t_3).\end{aligned}$$

Similarly, denote by $Q^z_{\beta_1,\beta_2,\beta_3,\beta_4}(t_1, t_2, t_3)$ and $\tilde{Q}^z_{\beta_1,\beta_2,\beta_3,\beta_4}(\omega_1, \omega_2, \omega_3)$, $1 \leq \beta_1, \beta_2, \beta_3, \beta_4 \leq q$, the fourth-order cumulant and fourth-order spectral density of the process $\{\boldsymbol{z}(t)\}$, respectively.

We introduce a new contrast function $\mathscr{D}$ for two spectral density matrices as follows:

$$\mathscr{D}(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}) = \int_{-\pi}^{\pi} a(\boldsymbol{\theta})\mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega)]\, \mathrm{d}\omega, \tag{12.2}$$

where $|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)| := \sqrt{\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)^* \boldsymbol{f}_\theta(\omega)}$ is the square root of the matrix $\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)^* \boldsymbol{f}_{\boldsymbol{\theta}}(\omega)$, and

$$a(\boldsymbol{\theta}) = \begin{cases} \left(\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p+1}]\, \mathrm{d}\omega\right)^{-p/(p+1)}, & p \neq -1, \\ 1, & p = -1. \end{cases}$$

The proposal of this new contrast function (12.2) is based on the observation of the following lemma.

Lemma 12.1 *Let* $p > 0$. *Suppose the matrices* $\boldsymbol{f}_{\boldsymbol{\theta}}$ *and* $\boldsymbol{g}$ *are both Hermitian and semi-definite in any* $\omega \in \Lambda$. *The following inequality holds:*

$$\mathscr{D}(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}) \leq \left(\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{g}(\omega)|^{p+1}]\Big)\, \mathrm{d}\omega\right)^{1/(p+1)}. \tag{12.3}$$

Proof We only have to show that

$$\mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega)] \leq \Big(\mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p+1}]\Big)^{p/(p+1)} \Big(\mathrm{tr}[|\boldsymbol{g}(\omega)|^{p+1}]\Big)^{1/(p+1)}, \qquad p > 0. \tag{12.4}$$

Denoting the eigenvalues of a $q \times q$-matrix A by $\Xi_1(A) \geq \cdots \geq \Xi_q(A)$, we first prove that

$$\sum_{j=1}^{q} \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega)) \leq \sum_{j=1}^{q} \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p)\Xi_j(\boldsymbol{g}(\omega)). \tag{12.5}$$

From the fundamental property that the determinant is the product of eigenvalues,

$$\begin{aligned}\prod_{j=1}^{q} \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega)) &= \det\big(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega)\big) \\ &= \det\big(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p\big) \det\big(\boldsymbol{g}(\omega)\big) \\ &= \prod_{j=1}^{q} \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p) \prod_{k=1}^{q} \Xi_k(\boldsymbol{g}(\omega)) \\ &= \prod_{j=1}^{q} \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p) \Xi_j(\boldsymbol{g}(\omega)).\end{aligned}$$

Equivalently, we have

$$\sum_{j=1}^{q} \log \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega)) = \sum_{j=1}^{q} \log\big\{\Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p) \Xi_j(\boldsymbol{g}(\omega))\big\}.$$

Let us define two vectors $\boldsymbol{x} := (x_1, \dots, x_{q+1})$ and $\boldsymbol{y} := (y_1, \dots, y_{q+1})$, where the components of $\boldsymbol{x}$ are

$$\begin{aligned} x_j &= \log \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega)), \qquad j = 1, \dots, q, \\ x_{q+1} &= \min\big(\log \Xi_q(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p), \log \Xi_q(\boldsymbol{g}(\omega))\big), \end{aligned}$$

and the components of $\boldsymbol{y}$ are

$$\begin{aligned} y_j &= \log \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p) \Xi_j(\boldsymbol{g}(\omega)), \qquad j = 1, \dots, q, \\ y_{q+1} &= \sum_{j=1}^{q+1} x_j - \sum_{j=1}^{q} y_j. \end{aligned}$$

If it holds that

$$\sum_{j=1}^{l} x_j \le \sum_{j=1}^{l} y_j, \qquad l = 1, \dots, q+1, \tag{12.6}$$

then applying Karamata's inequality (e.g., [1, p. 125]) with the convex function $\exp(\cdot)$ to each element of $\boldsymbol{x}$ and $\boldsymbol{y}$, we obtain

$$\begin{aligned} &\sum_{j=1}^{q} \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega)) \\ &\le \sum_{j=1}^{q} \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p) \Xi_j(\boldsymbol{g}(\omega)) + \exp(y_{q+1}) - \exp(x_{q+1}) \end{aligned}$$

$$\leq \sum_{j=1}^{q} \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p)\Xi_j(\boldsymbol{g}(\omega)) + \exp(x_{q+1})\left(\exp\Big(\sum_{i=1}^{q}(x_q - y_q)\Big) - 1\right)$$
$$\leq \sum_{j=1}^{q} \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p)\Xi_j(\boldsymbol{g}(\omega)),$$

since $\exp\Big(\sum_{i=1}^{q}(x_q - y_q)\Big) \leq 1$. This shows the inequality (12.5).

Now let us prove the inequality (12.6), i.e.,

$$\sum_{j=1}^{l} \log \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega)) \leq \sum_{j=1}^{l} \log\big\{\Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p)\Xi_j(\boldsymbol{g}(\omega))\big\}, \tag{12.7}$$

for any $1 \leq l \leq q$, since it is trivial when $l = q + 1$. This proof depends on the singular value decomposition, since the product $|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega)$ is not necessarily Hermitian. Concretely, there exist two unitary matrices $\boldsymbol{U}, \boldsymbol{V} \in \mathbb{C}^{q \times q}$ such that

$$|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega) = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^*,$$

where $\boldsymbol{\Sigma} = \mathrm{diag}(\Xi_1(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega)), \ldots, \Xi_q(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega)))$. With the first l columns $\boldsymbol{U}_l$ and $\boldsymbol{V}_l$ of the matrices $\boldsymbol{U}$ and $\boldsymbol{V}$, we have

$$\mathrm{diag}(\Xi_1(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega)), \ldots, \Xi_l(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega))) = \boldsymbol{U}_l^*|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega)\boldsymbol{V}_l,$$

for any $l = 1, \ldots, q$. Thus,

$$\begin{aligned}\prod_{j=1}^{l} \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega)) &= \det\big(\boldsymbol{U}_l^*|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega)\boldsymbol{V}_l\big) \\ &\leq \det\big(\boldsymbol{U}_l^*|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p\big)\det\big(\boldsymbol{g}(\omega)\boldsymbol{V}_l\big) \\ &\leq \prod_{j=1}^{l} \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p)\Xi_j(\boldsymbol{g}(\omega)), \end{aligned} \tag{12.8}$$

which implies (12.7) after taking the logarithm of both sides in (12.8). This completes the proof of (12.5).

Also, applying Hölder's inequality yields

$$
\begin{aligned}
\mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega)] &= \sum_{j=1}^{q} \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega)) \\
&\le \sum_{j=1}^{q} \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p)\,\Xi_j(\boldsymbol{g}(\omega)) \\
&\le \left\{\sum_{j=1}^{q} \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p)^{(p+1)/p}\right\}^{p/(p+1)} \left\{\sum_{j=1}^{q} \Xi_j(\boldsymbol{g}(\omega))^{p+1}\right\}^{1/(p+1)} \\
&= \left\{\sum_{j=1}^{q} \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p+1})\right\}^{p/(p+1)} \left\{\sum_{j=1}^{q} \Xi_j(|\boldsymbol{g}(\omega)|^{p+1})\right\}^{1/(p+1)} \\
&= \left(\mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p+1}]\right)^{p/(p+1)} \left(\mathrm{tr}[|\boldsymbol{g}(\omega)|^{p+1}]\right)^{1/(p+1)},
\end{aligned}
$$

which shows (12.4). Further, applying Hölder's inequality again, we can see that

$$
\begin{aligned}
&\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \boldsymbol{g}(\omega)]\,\mathrm{d}\omega \\
\le &\int_{-\pi}^{\pi} \left(\mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p+1}]\right)^{p/(p+1)} \left(\mathrm{tr}[|\boldsymbol{g}(\omega)|^{p+1}]\right)^{1/(p+1)} \mathrm{d}\omega \\
\le &\left(\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p+1}]\,\mathrm{d}\omega\right)^{p/(p+1)} \left(\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{g}(\omega)|^{p+1}]\,\mathrm{d}\omega\right)^{1/(p+1)}, \qquad (12.9)
\end{aligned}
$$

which is equivalent to (12.3). Thus, the proof of Lemma 12.1 is complete. □

Corollary 12.1 *Let $p < 0$. Suppose the matrices $\boldsymbol{f}_{\boldsymbol{\theta}}$ and $\boldsymbol{g}$ are both Hermitian and semi-definite in any $\omega \in \Lambda$. The following inequality holds:*

$$
\mathscr{D}(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{g}) \ge \left(\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{g}(\omega)|^{p+1}]\,\mathrm{d}\omega\right)^{1/(p+1)}. \qquad (12.10)
$$

Proof We consider two cases: (i) $p = -1$ and (ii) $p \ne -1$.

For the case (i), we refer the readers to [2, 4]. For the case (ii), let

$$
r = \frac{p}{p+1}, \qquad -1 < p < 0.
$$

Note that $1 - 1/r < 0, r(p+1) = p$ and $(p+1)/p = 1/r$. Applying the inequality (12.9) to $\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p+1} |\boldsymbol{g}(\omega)|^r |\boldsymbol{g}(\omega)|^{-r}]\,\mathrm{d}\omega$ yields

$$\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p+1}|\boldsymbol{g}(\omega)|^{1/r}|\boldsymbol{g}(\omega)|^{-1/r}]\,\mathrm{d}\omega$$
$$\leq \left(\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p}|\boldsymbol{g}(\omega)|]\,\mathrm{d}\omega\right)^{1/r}\left(\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{g}(\omega)|^{-p/r}]\,\mathrm{d}\omega\right)^{1/(p+1)}.$$

Equivalently, this shows that

$$\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p}|\boldsymbol{g}(\omega)|]\,\mathrm{d}\omega$$
$$\geq\left(\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p+1}|\boldsymbol{g}(\omega)|^{1/r}|\boldsymbol{g}(\omega)|^{-1/r}]\,\mathrm{d}\omega\right)^{r}\left(\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{g}(\omega)|^{-p/r}]\,\mathrm{d}\omega\right)^{-r/(p+1)}$$
$$=\left(\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p+1}\,\mathrm{d}\omega\right)^{p/(p+1)}\left(\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{g}(\omega)|]^{-(p+1)}\,\mathrm{d}\omega\right)^{-1/(p+1)}.$$

Thus, we obtain

$$\mathscr{D}(\boldsymbol{f}_{\boldsymbol{\theta}},\boldsymbol{g}) \geq \left(\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{g}(\omega)|^{p+1}]\,\mathrm{d}\omega\right)^{1/(p+1)}.$$

When $p < -1$, the inequality (12.10) can be shown in a similar way. □

12.3 Statistical Inference

In this section, we discuss the parameter estimation problem based on the new contrast function (12.2). The following assumptions are imposed:

Assumption 12.1

(i) The parameter space Θ is a compact subset of $\mathbb{R}^d$.
(i) If $\boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$, then $\boldsymbol{f}_{\boldsymbol{\theta}_1} \neq \boldsymbol{f}_{\boldsymbol{\theta}_2}$ on a set of positive Lebesgue measure.
(i) The spectral density matrix $\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)$ is three times continuously differentiable with respect to $\boldsymbol{\theta}$ and the second derivative $\frac{\partial^2}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^\top}\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)$ is continuous in ω.

Lemma 12.2
Under Assumption 12.1, we have the following:

(i) When $p > 0$, $\boldsymbol{\theta}_0$ maximizes the disparity $\mathscr{D}(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{f}_{\boldsymbol{\theta}_0})$ if $\boldsymbol{\theta}_0 \in \Theta$.
(ii) When $p < 0$, $\boldsymbol{\theta}_0$ minimizes the disparity $\mathscr{D}(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{f}_{\boldsymbol{\theta}_0})$ if $\boldsymbol{\theta}_0 \in \Theta$.

Proof The equality in (12.3) or (12.10) only holds if and only if the equality in Karamata's inequality holds. In other words,

$$\Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p\boldsymbol{f}_{\boldsymbol{\theta}_0}(\omega)) = \Xi_j(|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p)\,\Xi_j(\boldsymbol{f}_{\boldsymbol{\theta}_0}(\omega)), \tag{12.11}$$

for all $j = 1, \ldots, q$ and $\omega \in \Lambda$. If the spectral density matrices $\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)$ and $\boldsymbol{f}_{\boldsymbol{\theta}_0}(\omega)$ commute for any $\omega \in \Lambda$, i.e.,

$$\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)\boldsymbol{f}_{\boldsymbol{\theta}_0}(\omega) = \boldsymbol{f}_{\boldsymbol{\theta}_0}(\omega)\boldsymbol{f}_{\boldsymbol{\theta}}(\omega),$$

for all $\omega \in \Lambda$, then (12.11) holds. Thus, the statements (i) and (ii) are true if $\boldsymbol{\theta} = \boldsymbol{\theta}_0$. □

In the following, let $p < 0$. Following Lemma 12.2, we defined the true value $\boldsymbol{\theta}_0$ as the minimizer of the contrast function, i.e.,

$$\boldsymbol{\theta}_0 = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathscr{D}(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{f}_{\boldsymbol{\theta}_0}). \tag{12.12}$$

We now turn to prepare the regularity conditions for the process (12.1). See, e.g., [4], for details.

Let $\mathcal{F}(t)$ denote the σ-field generated by random vectors $\{\boldsymbol{e}(s); -\infty < s \le t\}$.

Assumption 12.2

(i) For each $1 \le \beta_1, \beta_2 \le s$, nonnegative integer m, and real $\eta_1 > 0$,

$$\mathrm{Var}[\mathbb{E}\big(e_{\beta_1}(t)e_{\beta_2}(t+m)|\mathcal{F}(t-\tau)\big)] = O(\tau^{-2-\eta_1})$$

uniformly in t.

(ii) For each $1 \le \beta_1, \beta_2, \beta_3, \beta_4 \le s$, and any real $\eta_2 > 0$,

$$\begin{aligned}\mathbb{E}|\mathbb{E}\{&e_{\beta_1}(t_1)e_{\beta_2}(t_2)e_{\beta_3}(t_3)e_{\beta_4}(t_4)|\mathcal{F}(t_1-\tau)\}\\ &- \mathbb{E}(e_{\beta_1}(t_1)e_{\beta_2}(t_2)e_{\beta_3}(t_3)e_{\beta_4}(t_4))| = O(\tau^{-1-\eta_2}),\end{aligned}$$

uniformly in t_1, where $t_1 \le t_2 \le t_3 \le t_4$.

(iii) For each $1 \le \beta_1, \beta_2 \le s$, any real $\eta_3 > 0$, and for any fixed integer $L \ge 0$, there exists $B_{\eta_3} > 0$ such that

$$\mathbb{E}[T(n,s)^2 \mathbb{1}\{T(n,s) > B_{\eta_3}\}] < \eta_3$$

uniformly in n and s, where

$$T(n,s) = \left[\frac{1}{n}\sum_{r=0}^{L}\left\{\sum_{t=1}^{n} e_{\beta_1}(t+s)e_{\beta_2}(t+s+r) - K_{\beta_1\beta_2}\delta(0,r)\right\}^2\right]^{1/2}.$$

(iv) The spectral densities f_{jj}, $1 \le j \le q$, are square-integrable, and $\boldsymbol{f}$ is in the Lipschitz class of degree k, $k > 1/2$.

Let $\boldsymbol{I}_n^z$ be the periodogram constructed from a partial realization $\{z(1), \ldots, z(n)\}$,

$$\boldsymbol{I}_n^z(\omega) = \boldsymbol{d}_n^z(\omega)\boldsymbol{d}_n^z(\omega)^*, \qquad \boldsymbol{d}_n^z(\omega) = \frac{1}{\sqrt{2\pi n}}\sum_{t=1}^{n} \boldsymbol{z}(t)\mathrm{e}^{\mathrm{i}t\omega}, \qquad \omega \in \Lambda.$$

Denote the (α, β)-th component of $\boldsymbol{I}_n^z$ by $I_{\alpha\beta}^z$. The parameter estimator based on (12.12) can be defined as

$$\hat{\boldsymbol{\theta}}_n = \arg\min_{\boldsymbol{\theta}\in\Theta} \mathscr{D}(\boldsymbol{f}_{\boldsymbol{\theta}}, \boldsymbol{I}_n^z). \tag{12.13}$$

Before stating the next assumption, we prepare a lemma for reference.

Lemma 12.3
([10]) *For any $r \in \mathbb{R}$ and any positive definite matrix A, it holds that*

$$\frac{\mathrm{d}A^r}{\mathrm{d}\boldsymbol{\theta}} = rA^{\alpha}\left(\frac{\mathrm{d}A}{\mathrm{d}\boldsymbol{\theta}} + H_{\alpha,r}^A\right)A^{r-1-\alpha}, \qquad \alpha \in \mathbb{R},$$

where

$$H_{\alpha,r}^A := \frac{1}{r}A^{-\alpha}F^A A^{\alpha+1} - \frac{1}{r}A^{r-\alpha}F^A A^{\alpha-r+1} - F^A A + AF^A,$$

and

$$F^A = \frac{\mathrm{d}\Gamma_A}{\mathrm{d}\boldsymbol{\theta}}\Gamma_A^\top,$$

where Γ_A is the orthogonal matrix diagonalizing the matrix A.

This lemma generalizes the derivative of the inverse of a matrix. It can be summarized in the following as a corollary.

Corollary 12.2 *For any positive definite matrix A,*

$$\frac{\mathrm{d}A^{-1}}{\mathrm{d}\boldsymbol{\theta}} = -A^{-1}\frac{\mathrm{d}A}{\mathrm{d}\boldsymbol{\theta}}A^{-1}.$$

It is technically difficult to develop the theoretical argument about the parameter estimation in the full scope of Lemma 12.3. We impose the following assumption.

Assumption 12.3 For some $\eta \in \mathbb{R}$, it holds that $H_{\eta,p}^{|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|} = O$ for any $\lambda \in \Lambda$. In other words, there exists $\eta \in \mathbb{R}$ such that

$$\frac{\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^p}{\partial\boldsymbol{\theta}} = p|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{\eta}\frac{\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|}{\partial\boldsymbol{\theta}}|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p-1-\eta}.$$

This assumption is restrictive but it is worth investigating. The typical example of Assumption 12.3 is given in Corollary 12.2, i.e., $p = -1$. The constant η can be always chosen as $\eta = -1$, which results in

$$\frac{\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{-1}}{\partial\boldsymbol{\theta}} = -|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{-1}\frac{\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|}{\partial\boldsymbol{\theta}}|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{-1}.$$

Other examples are listed as follows:

- The eigenvectors of $|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|$ do not depend on the parameter $\boldsymbol{\theta}$.
- The matrices $|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|$ and $F^{|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|}$ commute.

These examples are not explicit, but they are quite possible to be used in the modeling.

To keep the brevity, we let θ be a real-valued parameter[1] and ∂ denotes the (elementwise) derivative with respect to θ.

Theorem 12.1 *Under Assumptions 12.1–12.3, if* $\boldsymbol{\theta}_0 \in \Theta \subset \mathbb{R}$, *then the following holds:*

$$(\mathrm{a})\hat{\boldsymbol{\theta}}_n \xrightarrow{p} \boldsymbol{\theta}_0.$$
$$(\mathrm{b})\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathcal{N}(\mathbf{0}, H(\boldsymbol{\theta}_0)^{-1} V(\boldsymbol{\theta}_0) H(\boldsymbol{\theta}_0)^{-1}),$$

where

$$\begin{aligned} H(\boldsymbol{\theta}) = &\Big(\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^p \partial \boldsymbol{f}_{\boldsymbol{\theta}}(\omega)]\,\mathrm{d}\omega\Big)^2 - \Big(\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p+1}]\,\mathrm{d}\omega\Big) \\ &\times\Bigg(\int_{-\pi}^{\pi} \mathrm{tr}\Big[\partial\Big\{|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{\eta}\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)| \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p-1-\eta}\Big\}\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)\Big]\,\mathrm{d}\omega \\ &\qquad - \int_{-\pi}^{\pi} \mathrm{tr}\Big[\partial\Big\{|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p}\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|\Big\}\Big]\,\mathrm{d}\omega\Bigg), \end{aligned}$$

and

$$\begin{aligned} V(\boldsymbol{\theta}) = 4\pi \int_{-\pi}^{\pi} \mathrm{tr}\Bigg[&\Big\{\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p+1}]\,\mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{\eta}\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)| \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p-1-\eta} \\ &- \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p}\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|]\,\mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p}\Big\}\boldsymbol{f}_{\boldsymbol{\theta}}(\omega) \\ &\times \Big\{\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p+1}]\,\mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{\eta}\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)| \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p-1-\eta} \\ &- \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p}\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|]\,\mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p}\Big\}\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)\Bigg]\,\mathrm{d}\omega \\ + 2\pi \sum_{r,t,u,v=1}^{s} \iint_{-\pi}^{\pi} &\Big\{\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p+1}]\,\mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega_1)|^{\eta}\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega_1)| \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega_1)|^{p-1-\eta} \\ &- \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p}\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|]\,\mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega_1)|^{p}\Big\}_{rt} \\ &\times \Big\{\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p+1}]\,\mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega_2)|^{\eta}\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega_2)| \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega_2)|^{p-1-\eta} \end{aligned}$$

[1] The extension to the vector parameter $\boldsymbol{\theta} \in \Theta \subset \mathbb{R}^d$, $d > 2$, is straightforward, but the formula for $V(\boldsymbol{\theta})$ in Theorem 12.1 will be lengthy but not sufficiently fruitful for this paper. This leads to the thought of only presenting the case $\theta \in \mathbb{R}$.

$$- \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p} \partial |\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|] \, \mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega_2)|^{p} \Big\}_{uv}$$
$$\times \tilde{Q}^{z}_{rtuv}(-\omega_1, \omega_2, -\omega_2) \, \mathrm{d}\omega_1 \, \mathrm{d}\omega_2,$$

(A_{rt} denotes the (r, t)-th element of the matrix A) with

$$\begin{aligned}
&\tilde{Q}^{z}_{\beta_1,\beta_2,\beta_3,\beta_4}(-\omega_1, \omega_2, \omega_3) \\
&= \sum_{\alpha_1,\alpha_2,\alpha_3,\alpha_4=1}^{q} k_{\beta_1\alpha_1}(\omega_1 + \omega_2 + \omega_3) k_{\beta_2\alpha_2}(-\omega_1) k_{\beta_3\alpha_3}(-\omega_2) k_{\beta_4\alpha_4}(-\omega_3) \\
&\qquad \times \tilde{Q}^{e}_{\alpha_1,\alpha_2,\alpha_3,\alpha_4}(\omega_1 + \omega_2 + \omega_3, \omega_2, \omega_3).
\end{aligned}$$

Proof This can be shown in the same way as [6, p. 128], if we replace $A_1(\boldsymbol{\theta})$, $B_1(\boldsymbol{\theta})$, $A_2(\boldsymbol{\theta})$, and $B_2(\boldsymbol{\theta})$ in [6] with the following quantities:

$$\begin{aligned}
A_1(\boldsymbol{\theta}) &:= \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p+1}] \, \mathrm{d}\lambda, \\
B_1(\boldsymbol{\theta}) &:= |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{\eta} \partial |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)| \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p-1-\eta}, \\
A_2(\boldsymbol{\theta}) &:= \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p} \partial |\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|] \, \mathrm{d}\lambda, \\
B_2(\boldsymbol{\theta}) &:= |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p}.
\end{aligned}$$

The details are omitted. □

Remark 12.1 Under the condition

$$\mathrm{cum}(e_{\alpha_1}(t_1), e_{\alpha_2}(t_2), e_{\alpha_3}(t_3), e_{\alpha_4}(t_4)) = \begin{cases} \kappa_{\alpha_1\alpha_2\alpha_3\alpha_4}, & \text{if } t_1 = t_2 = t_3 = t_4, \\ 0, & \text{otherwise,} \end{cases}$$

$V(\boldsymbol{\theta})$ can be simplified as

$$\begin{aligned}
V(\boldsymbol{\theta}) = 4\pi \int_{-\pi}^{\pi} \mathrm{tr}\Big[\Big\{ &\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p+1}] \, \mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{\eta} \partial |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)| \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p-1-\eta} \\
&- \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p} \partial |\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|] \, \mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p} \Big\} \boldsymbol{f}_{\boldsymbol{\theta}}(\omega) \\
&\times \Big\{ \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p+1}] \, \mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{\eta} \partial |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)| \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p-1-\eta} \\
&- \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p} \partial |\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|] \, \mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p} \Big\} \boldsymbol{f}_{\boldsymbol{\theta}}(\omega) \Big] \mathrm{d}\omega \\
&+ 2\pi \sum_{r,t,u,v=1}^{s} \kappa_{rtuv} \iint_{-\pi}^{\pi} \Big\{ \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p+1}] \, \mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega_1)|^{\eta} \partial |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega_1)| \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega_1)|^{p-1-\eta} \\
&- \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p} \partial |\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|] \, \mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega_1)|^{p} \Big\}_{rt} \mathrm{d}\omega_1
\end{aligned}$$

$$\times \left\{ \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p+1}]\,\mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega_2)|^{\eta}\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega_2)| \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega_2)|^{p-1-\eta} - \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p}\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|]\,\mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega_2)|^{p} \right\}_{uv} \mathrm{d}\omega_2.$$

In addition, if the process (12.1) is Gaussian, then

$$\begin{aligned} V(\boldsymbol{\theta}) = 4\pi \int_{-\pi}^{\pi} \mathrm{tr}\Bigg[& \Bigg\{ \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p+1}]\,\mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{\eta}\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)| \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p-1-\eta} \\ & - \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p}\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|]\,\mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p} \Bigg\} \boldsymbol{f}_{\boldsymbol{\theta}}(\omega) \\ & \times \Bigg\{ \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p+1}]\,\mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{\eta}\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)| \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p-1-\eta} \\ & - \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p}\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|]\,\mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p} \Bigg\} \boldsymbol{f}_{\boldsymbol{\theta}}(\omega) \Bigg] \mathrm{d}\omega \\ := \tilde{V}(\boldsymbol{\theta}) \quad & \text{(say)}. \end{aligned}$$

12.4 Asymptotic Efficiency under Gaussianity

In this section, we consider the asymptotic efficiency for multivariate Gaussian processes (12.1). The Gaussian–Fisher information matrix is known as

$$\mathscr{F}(\boldsymbol{\theta}) = \frac{1}{4\pi} \int_{-\pi}^{\pi} \mathrm{tr}\Big[\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)^{-1}\partial \boldsymbol{f}_{\boldsymbol{\theta}}(\lambda) \boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)^{-1}\partial \boldsymbol{f}_{\boldsymbol{\theta}}(\lambda) \Big] \mathrm{d}\lambda.$$

We refer the readers to [15, p. 58] for details.

Remark 12.2 For the estimation of innovation-free parameters, we have

$$\begin{aligned} \mathbf{0} &= \frac{\partial}{\partial \boldsymbol{\theta}} \log \det(K) \\ &= \frac{\partial}{\partial \boldsymbol{\theta}} \int_{-\pi}^{\pi} \log \det(\boldsymbol{f}_{\boldsymbol{\theta}}(\omega))\,\mathrm{d}\omega = \int_{-\pi}^{\pi} \mathrm{tr}\Big[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{-1} \frac{\partial}{\partial \boldsymbol{\theta}} |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)| \Big] \mathrm{d}\omega. \end{aligned}$$

Comparing Theorem 12.1 with the established one for the Whittle likelihood in [4], we substitute $p = -1$, and thus, from Corollary 12.2, $\eta = -1$, so that

$$\begin{aligned} \mathrm{tr}\Bigg[& \Bigg\{ \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p+1}]\,\mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{\eta}\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)| \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p-1-\eta} \\ & - \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p}\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|]\,\mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p} \Bigg\} \boldsymbol{f}_{\boldsymbol{\theta}}(\omega) \end{aligned}$$

$$\times \left\{ \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p+1}]\, \mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{\eta} \partial |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)| \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p-1-\eta} \right.$$
$$\left. - \int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{p} \partial |\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|]\, \mathrm{d}\lambda \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p} \right\} \boldsymbol{f}_{\boldsymbol{\theta}}(\omega) \Bigg]$$
$$= q \,\mathrm{tr}\left[\left\{ |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{-1} \partial |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)| |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{-1} \partial |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)| \right] \right],$$

which shows that $\tilde{V}(\boldsymbol{\theta}) = (4\pi)^2 q \cdot \mathscr{F}(\boldsymbol{\theta})$.

Before investigating the asymptotic efficiency, we first introduce a useful inequality in the following.

Lemma 12.4 *Suppose the matrix products $A(\omega)A(\omega)^{\top}\boldsymbol{g}(\omega)$, $A(\omega)B(\omega)^{\top}$, and $B(\omega)B(\omega)^{\top}\boldsymbol{g}(\omega)^{-1}$ are all well-defined and square matrices. If*

$$\int_{-\pi}^{\pi} \mathrm{tr}\big[B(\omega)\boldsymbol{g}(\omega)^{-1}B(\omega)^{\top}\boldsymbol{g}(\omega)^{-1}\big]\,\mathrm{d}\omega \neq 0,$$

then it holds that

$$\left(\int_{-\pi}^{\pi} \mathrm{tr}\big[B(\omega)\boldsymbol{g}(\omega)^{-1}B(\omega)^{\top}\boldsymbol{g}(\omega)^{-1}\big]\,\mathrm{d}\omega\right)^{-1}$$
$$\leq \left(\int_{-\pi}^{\pi} \mathrm{tr}\big[A(\omega)B(\omega)^{\top}\big]\,\mathrm{d}\omega\right)^{-1} \int_{-\pi}^{\pi} \mathrm{tr}\big[A(\omega)\boldsymbol{g}(\omega)A(\omega)^{\top}\boldsymbol{g}(\omega)\big]\,\mathrm{d}\omega$$
$$\times \left(\int_{-\pi}^{\pi} \mathrm{tr}\big[A(\omega)B(\omega)^{\top}\big]\,\mathrm{d}\omega\right)^{-1}.$$

Especially, the equality holds if and only if there exists a constant matrix C such that

$$\mathrm{tr}\big[\boldsymbol{g}(\omega)A(\omega) + CB(\omega)\big] = 0, \qquad a.e. \quad \omega \in [-\pi, \pi].$$

Proof The proof follows the same line with the proof for the Cauchy–Schwarz inequality. The inequality without the trace of the matrices can be found in [3, 5]. □

To compare the asymptotic variance of the estimator in (12.12), we need a stronger assumption in the following:

Assumption 12.4

(i) The Gaussian–Fisher information matrix is invertible.
(ii) The eigenvectors of $\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)$ do not depend on the parameter $\boldsymbol{\theta}$.

Note that Assumption 12.4 (ii) implies Assumption 12.3. This assumption allows the matrices $|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|$ and $\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|$ to commute. More general, under this assumption, $|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{\xi_1}$ and $\partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\lambda)|^{\xi_2}$ commute for any real powers $\xi_1, \xi_2 \in \mathbb{R}$. Thus, it alleviates the computation of the second derivative of the contrast functions and enables the application of the inequality in Lemma 12.4.

Theorem 12.2 *Under Assumptions 12.1, 12.2, and 12.4, if $\boldsymbol{\theta}_0 \in \Theta \subset \mathbb{R}$, then we obtain*

$$\mathscr{F}(\boldsymbol{\theta}_0)^{-1} \leq H(\boldsymbol{\theta}_0)^{-1}\tilde{V}(\boldsymbol{\theta}_0)H(\boldsymbol{\theta}_0)^{-1}.$$

The equality holds then $\alpha = -1$, or the spectral density matrix of the process (12.1) *does not depend on ω.*

Proof Take the matrices $A(\omega)$, $B(\omega)$, and $\boldsymbol{g}(\omega)$ as

$$\begin{aligned} A(\omega) &= \partial\big(A_1(\boldsymbol{\theta})B_1(\boldsymbol{\theta}) - A_2(\boldsymbol{\theta})B_2(\boldsymbol{\theta})\big), \\ B(\omega) &= \partial|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|, \\ \boldsymbol{g}(\omega) &= |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|. \end{aligned}$$

It is not difficult to see that

$$\begin{aligned} \mathscr{F}(\boldsymbol{\theta}) &= \int_{-\pi}^{\pi} \mathrm{tr}\big[B(\omega)\boldsymbol{g}(\omega)^{-1}B(\omega)^{\top}\boldsymbol{g}(\omega)^{-1}\big]\,\mathrm{d}\omega, \\ \tilde{V}(\boldsymbol{\theta}) &= \int_{-\pi}^{\pi} \mathrm{tr}\big[A(\omega)\boldsymbol{g}(\omega)A(\omega)^{\top}\boldsymbol{g}(\omega)\big]\,\mathrm{d}\omega. \end{aligned}$$

Thus, the only equality we have to show is

$$H(\boldsymbol{\theta}) = \int_{-\pi}^{\pi} \mathrm{tr}\big[A(\omega)B(\omega)^{\top}\big]\,\mathrm{d}\omega.$$

In fact, we have

$$\begin{aligned} \int_{-\pi}^{\pi} \mathrm{tr}\Big[\big\{\partial A_1(\boldsymbol{\theta})\big\}B_1(\boldsymbol{\theta})B(\omega)^{\top}\Big]\,\mathrm{d}\omega - \int_{-\pi}^{\pi} \mathrm{tr}\Big[A_2(\boldsymbol{\theta})\big\{\partial B_2(\boldsymbol{\theta})\big\}B(\omega)^{\top}\Big]\,\mathrm{d}\omega \\ = \Big(\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p}\partial\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)]\,\mathrm{d}\omega\Big)^2. \end{aligned}$$

Noting that

$$\int_{-\pi}^{\pi} \mathrm{tr}\Big[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p}\frac{\partial^2}{\partial\theta_i\theta_j}|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|\Big] = \int_{-\pi}^{\pi} \mathrm{tr}\Big[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{\eta}\frac{\partial^2}{\partial\theta_i\theta_j}|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)||\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p-\eta}\Big],$$

$1 \leq i, j \leq d$, we obtain

$$\begin{aligned} &\int_{-\pi}^{\pi} \mathrm{tr}\Big[\big\{\partial B_1(\boldsymbol{\theta})\big\}A_1(\boldsymbol{\theta})B(\omega)^{\top}\Big]\,\mathrm{d}\omega - \int_{-\pi}^{\pi} \mathrm{tr}\Big[B_2(\boldsymbol{\theta})\big\{\partial A_2(\boldsymbol{\theta})\big\}B(\omega)^{\top}\Big]\,\mathrm{d}\omega \\ =&\Big(\int_{-\pi}^{\pi} \mathrm{tr}[|\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p+1}]\,\mathrm{d}\omega\Big) \end{aligned}$$

$$\times \left(\int_{-\pi}^{\pi} \operatorname{tr}\left[\partial \left\{ |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{\eta} \partial |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)| \cdot |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p-1-\eta} \right\} \boldsymbol{f}_{\boldsymbol{\theta}}(\omega) \right] \mathrm{d}\omega \right.$$
$$\left. - \int_{-\pi}^{\pi} \operatorname{tr}\left[\partial \left\{ |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)|^{p} \partial |\boldsymbol{f}_{\boldsymbol{\theta}}(\omega)| \right\} \right] \mathrm{d}\omega \right)$$

under Assumption 12.4, which completes the proof. □

12.5 Numerical Study

In this section, we present the numerical performance of our minimum contrast estimator (12.12) for multivariate time series. The data are generated from the following two models:

Model (i)

$$\boldsymbol{X}_t = \begin{pmatrix} 0.2 & \theta_1 \\ 0 & 0.4 \end{pmatrix} \boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{2\times 2}),$$

where $\boldsymbol{I}_{2\times 2}$ denotes the identity matrix of size 2.

Model (ii)

$$\boldsymbol{X}_t = \begin{pmatrix} 0.2 & \theta_2 \\ 0 & 0.4 \end{pmatrix} \boldsymbol{X}_{t-1} + \begin{pmatrix} 0 & 0 \\ \theta_2/3 & 0 \end{pmatrix} \boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t, \qquad \boldsymbol{\epsilon}_t \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_{2\times 2}).$$

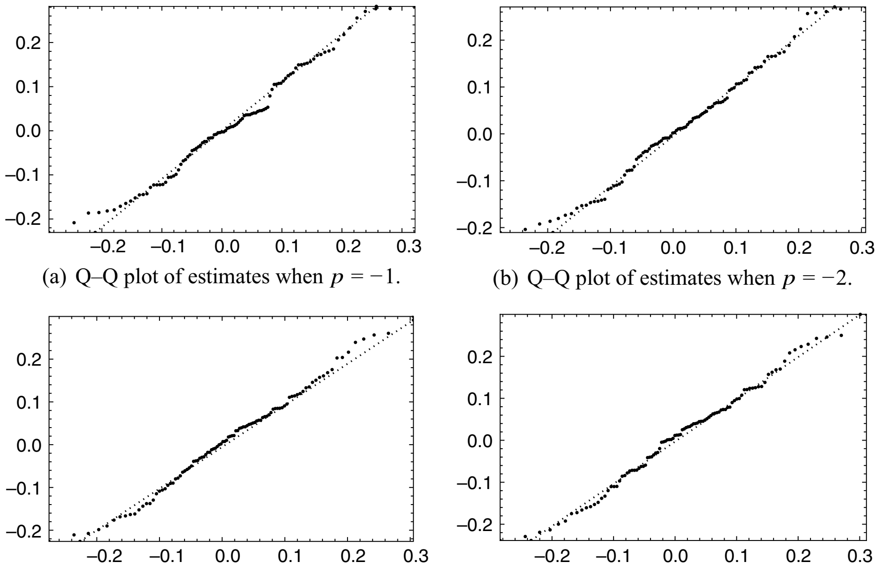
We simulated the estimation with the true parameters $\theta_1 = 0, 0.2, 0.4, 0.6, 0.8, 1$ and $\theta_2 = 0, 0.2, 0.4, 0.6, 0.8, 1$. All cases with the true parameters are stationary. The size of the data is $n = 100$. The simulations are repeated 100 times.

In our simulations, we used the contrast function (12.2) with $p = -1, -2, -3, -4$. Also, we suppose the parameter $\theta \in [\theta_0 - \zeta, \theta_0 + \zeta]$, where $\theta_0 = \theta_1$ for **Model (i)** and $\theta_0 = \theta_2$ for **Model (ii)**, and $\zeta = 0.3$.

We adopt **Model (i)** with $\theta_1 = 0$ and **Model (ii)** with $\theta_2 = 0.8$ as the illustration of the estimates in numerical simulations, and show them in Figs. 12.1 and 12.2 by the Q–Q plot.

We report the empirical relative efficiency (ERE) of each estimators when $p = -1, -2, -3, -4$, compared with the estimator when $p = -1$. To be specific, the ERE is defined as

$$\mathrm{ERE} = \frac{\sum_{i=1}^{100} (\hat{\theta}^i_{(-1)} - \theta_0)^2}{\sum_{i=1}^{100} (\hat{\theta}^i_{(p)} - \theta_0)^2},$$



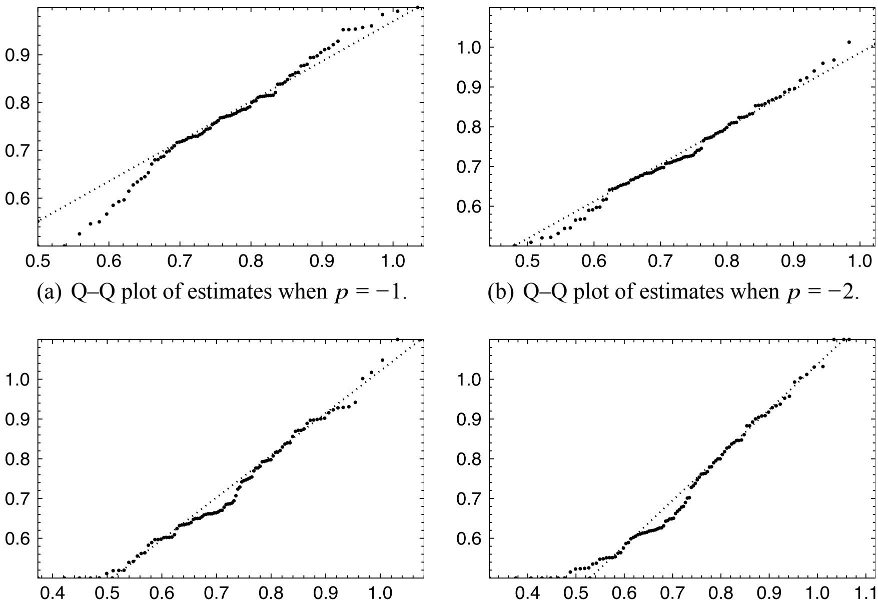
(a) Q–Q plot of estimates when $p = -1$.

(b) Q–Q plot of estimates when $p = -2$.

(c) Q–Q plot of estimates when $p = -3$.

(d) Q–Q plot of estimates when $p = -4$.

Fig. 12.1 Q–Q plots for **Model (i)** with $\theta_1 = 0$



(a) Q–Q plot of estimates when $p = -1$.

(b) Q–Q plot of estimates when $p = -2$.

(c) Q–Q plot of estimates when $p = -3$.

(d) Q–Q plot of estimates when $p = -4$.

Fig. 12.2 Q–Q plots for **Model (ii)** with $\theta_2 = 0.8$

Table 12.1 The ERE of the estimators with $p = -1, -2, -3, -4$ for **Model (i)**

θ_1	$p = -1$	$p = -2$	$p = -3$	$p = -4$
$\theta_1 = 0.0$	1.000	1.099	1.106	1.058
$\theta_1 = 0.2$	1.000	1.076	1.051	0.952
$\theta_1 = 0.4$	1.000	1.020	0.852	0.675
$\theta_1 = 0.6$	1.000	0.916	0.631	0.534
$\theta_1 = 0.8$	1.000	0.795	0.546	0.451
$\theta_1 = 1.0$	1.000	0.733	0.523	0.451

Table 12.2 The ERE of the estimators with $p = -1, -2, -3, -4$ for **Model (ii)**

θ_2	$p = -1$	$p = -2$	$p = -3$	$p = -4$
$\theta_2 = 0.0$	1.000	1.090	1.105	1.068
$\theta_2 = 0.2$	1.000	1.071	1.033	0.932
$\theta_2 = 0.4$	1.000	0.977	0.828	0.611
$\theta_2 = 0.6$	1.000	0.838	0.603	0.466
$\theta_2 = 0.8$	1.000	0.706	0.499	0.397
$\theta_2 = 1.0$	1.000	0.508	0.360	0.319

where $\hat{\theta}^i_{(p)}$ denotes the estimate in the ith simulation. By definition, if the ERE is greater than 1, then the estimator $\hat{\theta}^i_{(p)}$ has smaller mean squared error. On the contrary, if the ERE is smaller than 1, then the estimator $\hat{\theta}^i_{(p)}$ has larger mean squared error.

From both Tables 12.1 and 12.2, we find that the contrast estimators with $p = -2, -3, -4$ have smaller mean squared error when $\theta_1 = 0$ or $\theta_2 = 0$. This new finding was unexpected, compared with the results in [6] for the scalar-valued time series. We will leave the interpretation of this new finding as our future work, but we conclude that the estimators when $p = -2, -3, -4$ concentrate around 0 more than the Whittle estimator (the estimate when $p = -1$).

Acknowledgements The author was supported by JSPS Grant-in-Aid for Scientific Research (C) 20K11719. He also would like to express his thanks to the Institute for Mathematical Science (IMS), Waseda University, for their kind hospitality.

References

1. Cvetkovski, Z (2012). *Inequalities: Theorems, Techniques and Selected Problems*. Springer Science & Business Media.
2. Dunsmuir, W. (1979). A central limit theorem for parameter estimation in stationary vector time series and its application to models for a signal observed with noise. *The Annals of Statistics* **7** 490–506.

3. Grenander, U. and Rosenblatt, M. (1957). *Statistical Analysis of Stationary Time Series*. Johm Wiley & Sons, New York.
4. Hosoya, Y. and Taniguchi, M. (1982). A central limit theorem for stationary processes and the parameter estimation of linear processes. *The Annals of Statistics* **10** 132–153.
5. Kholevo, A. (1969). On estimates of regression coefficients. *Theory of Probability & Its Applications* **14** 79–104.
6. Liu, Y. (2017). Robust parameter estimation for stationary processes by an exotic disparity from prediction problem. *Statistics & Probability Letters* **129** 120–130.
7. Liu, Y., Akashi, F. and Taniguchi, M. (2018). *Empirical Likelihood and Quantile Methods for Time Series: Efficiency, Robustness, Optimality, and Prediction*. Springer.
8. Liu, Y., Xue, Y. and Taniguchi, M. (2020). Robust linear interpolation and extrapolation of stationary time series in L_p. *Journal of Time Series Analysis* **41** 229–248.
9. Ogata, H. and Taniguchi, M. (2010). An empirical likelihood approach for non-Gaussian vector stationary processes and its application to minimum constrast estimation. *Australian and New Zealand Jounal of Statistics* **52** 451–468.
10. Sebastiani, P. (1996). On the derivatives of matrix powers. *SIAM Journal on Matrix Analysis and Applications* **17** 640–648.
11. Suto, Y., Liu, Y. and Taniguchi, M. (2016). Asymptotic theory of parameter estimation by a contrast function based on interpolation error. *Statistical Inference for Stochastic Processes* **19** 93–110.
12. Taniguchi, M. (1979). On estimation of parameters of Gaussian stationary processes. *Journal of Applied Probability* **16** 575–591.
13. Taniguchi, M. (1981). An estimation procedure of parameters of a certain spectral density model. *Journal of the Royal Statistical Society Series B* **43** 34–40.
14. Taniguchi, M. (1987). Minimum contrast estimation for spectral densities of stationary processes. *Journal of the Royal Statistical Society Series B* **49** 315–325.
15. Taniguchi, M. and Kakizawa, Y. (2000). *Asymptotic Theory of Statistical Inference for Time Series*. Springer.
16. Taniguchi, M., van Garderen, K. J. and Puri, M. L. (2003). Higher order asymptotic theory for minimum contrast estimators of spectral parameters of stationary processes. *Econometric Theory* **19** 984–1007.
17. Zhang, G. and Taniguchi, M. (1995). Nonparametric approach for discriminant analysis in time series. *Journaltitle of Nonparametric Statistics* **5** 91–101.

Chapter 13
Generalized Linear Spectral Models for Locally Stationary Processes

Tommaso Proietti, Alessandra Luati, and Enzo D'Innocenzo

Abstract A class of parametric models for locally stationary processes is introduced. The class depends on a power parameter that applies to the time-varying spectrum so that it can be locally represented by a (finite low dimensional) Fourier polynomial. The coefficients of the polynomial have an interpretation as time-varying autocovariances, whose dynamics are determined by a linear combination of smooth transition functions, depending on some static parameters. Frequency domain estimation is based on the generalized Whittle likelihood and the pre-periodogram, while model selection is performed through information criteria. Change points are identified via a sequence of score tests. Consistency and asymptotic normality are proved for the parametric estimators considered in the paper, under weak assumptions on the time-varying parameters.

13.1 Introduction

Locally stationary processes are stochastic processes that locally behave like stationary processes, but gradually change in time, with slowly varying autocovariances and spectral density functions. In the frequency domain, a characterization of the properties of locally stationary processes is provided by the time-varying spectrum. The theory of evolutionary spectra dates back to [37, 44], and has later received interest due to the development of a rigorous asymptotic theory, based on infill asymptotics. Infill asymptotics [8, 10] essentially consists in rescaling the time to the unit interval

T. Proietti
Università di Roma Tor Vergata, Via Columbia 2, 00133 Roma, Italy
e-mail: tommaso.proietti@uniroma2.it

A. Luati (✉)
Imperial College London, 180 Queen's Gate, SW7 2AZ London, UK
e-mail: a.luati@imperial.ac.uk

E. D'Innocenzo
Vrije Universiteit Amsterdam, De Boelelaan 1105, 1081 HV Amsterdam, Netherlands
e-mail: e.dinnocenzo@vu.nl

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_13

and letting the interval grow dense, so that, as long as the time dimension increases, more and more observations eventually become available at each local interval.

The estimation of the time-varying spectrum has attracted a great deal of attention. Some references are concerned with the optimal partitioning of the sample time series, as the paper by [1], who proposed an adaptive segmentation method based on binary trees for piecewise stationary processes. The paper [34] considered the Smooth Localized Complex Exponential or SLEX transform, based on dyadically partitioning the series into overlapping segments; the best basis algorithm was applied to determine the optimal segmentation. The SLEX approach was extended to the multivariate setting by [35]. A related method, based on fitting piecewise autoregressive processes, was proposed by [16], where the optimal number and locations of break points were developed according to a minimum description length criterion. Other localized methods were based on wavelets, as the model by [32], who used wavelets as stochastic building blocks. A test of stationarity based on wavelets was developed by [31]. Further references on testing stationarity include the contribution by [43], that discusses testing composite hypothesis in locally stationary processes, and the contribution by [17], that, with a different perspective, relies on a minimum distance approach by [17]. A smoothing spline ANOVA model for the log spectrum was proposed by [24], while a Bayesian approach to model nonstationary spectra was used by [41, 42].

Statistical inference in locally stationary processes, based on Whittle likelihood estimation and empirical process theory, has been largely developed by [10, 11, 13, 14]. An overview on locally stationary processes with a focus both on the theoretical aspects and on the applications that range from neuroscience [34, 35, 40, 46] to finance [2, 15] was provided by [12]. A bootstrap procedure for locally stationary time series that combines a time domain wild bootstrap approach with a non-parametric frequency domain approach was considered in [26].

This paper introduces a class of parametric models for locally stationary processes. The models are based on the Box–Cox transform of the spectral density function of a locally stationary process and therefore involve powers of the spectral density function as well as its log transform. Both log transforms and power transforms of the spectral density function (sdf) have a consolidated tradition in the analysis of stationary time series in the frequency domain, dating back to [4, 45]. Several specifications are nested in the class defined in the paper, such as time-varying autoregressive moving average (ARMA) models with smoothly varying coefficients and the time-varying extension of Bloomfield's exponential model.

Likelihood inference is carried out in the frequency domain. To enforce the constraints needed to guarantee the positivity of the spectral density, we reparameterize the model based on a set of inverse partial autocorrelations. Consistency and asymptotic normality of the Whittle estimators based on the pre-periodogram are proved based on results on infill asymptotics and empirical processes as in [13, 14]. A sequence of score tests allows us to account for change points. The results are illustrated through the empirical analysis of the annualized quarterly growth rate of the U.S. Gross Domestic Product.

13.2 A Generalized Linear Model for the Time-Varying Spectrum

Let $\{X_{tn}, t = 1, \ldots n\}_{n\in\mathbb{N}}$ denote a triangular array of random variables generated according to the linear causal locally stationary process

$$X_{tn} = \mu_{tn} + \sigma_{tn} \sum_{j=0}^{\infty} \psi_{tn,j}\epsilon_{t-j},$$

where $\psi_{tn,0} = 1$ and $\epsilon_t \sim \text{iid}(0, 1)$, i.e., $\{\epsilon_t\}$ is independently and identically distributed with zero mean and unit variance. Denoting with $\psi_{tn}(\omega) = \sum_{j=0}^{\infty} \psi_{tn,j} e^{-\imath\omega j}$ and with $\overline{\psi_{tn}(\omega)}$ its complex conjugate for $\omega \in [-\pi, \pi]$, there exist functions $\mu(t/n)$, $\sigma(t/n)$, and $\psi(t/n, \omega)$, such that $\sup_t |\mu_{tn} - \mu(t/n)| = O(n^{-1})$ and

$$\sup_{t,\omega} \left| \sigma_{tn}^2 \psi_{tn}(\omega)\overline{\psi_{tn}(\omega)} - 2\pi f\left(\frac{t}{n}, \omega\right) \right| = O(n^{-1}),$$

where $f(u, \omega)$, $u \in [0, 1]$ is the instantaneous sdf, where $u = t/n$ denotes rescaled time in [0, 1] and ω denotes the angular frequency in $[-\pi, \pi]$. The main assumption concerning the instantaneous sdf is given in the following.

Assumption 13.1 There exist two positive constants $\underline{C}$ and $\overline{C}$, such that $\forall(u, \omega)$, $0 < \underline{C} \leq f(u, \omega) \leq \overline{C} < \infty$.

Assumption 13.1 rules out locally non-invertible and long memory behavior.

We propose a class of locally stationary processes formulated in the frequency domain in terms of the instantaneous sdf, $f(u, \omega) > 0$. The latter is defined as the inverse [6] (Box–Cox) transformation of a trigonometric polynomial $g_\lambda(u, \omega)$, with varying coefficients

$$f(u, \omega) = \begin{cases} \frac{1}{2\pi}(1 + \lambda g_\lambda(u, \omega))^{\frac{1}{\lambda}} & \lambda \neq 0, \\ \frac{1}{2\pi}\exp(g_\lambda(u, \omega)) & \lambda = 0, \end{cases} \tag{13.1}$$

where $\lambda \in \mathbb{R}$ is the power transformation parameter, and

$$g_\lambda(u, \omega) = c_\lambda(u, 0) + 2\sum_{k=1}^{K} c_\lambda(u, k)\cos(\omega k). \tag{13.2}$$

The function $g_\lambda(u, \omega)$ is periodic in ω with period 2π and continuous in u; for a fixed u it is a Fourier polynomial of order K. The Fourier coefficients $c_\lambda(u, k)$ are referred to as the generalized cepstral coefficients (GCC), borrowing the terminology from [5].

Remark 13.1 The complexity of (13.1) is only apparent: it provides a generalized linear model for the sdf, resulting from the transformation, depending on λ, of a trigonometric polynomial, which is linear in the generalized cepstral coefficients. The link function is the Box–Cox inverse link, as $g_\lambda(u,\omega)$ is the Box–Cox transformation of the sdf, i.e.,

$$g_\lambda(u,\omega) = \begin{cases} \frac{[2\pi f(u,\omega)]^\lambda - 1}{\lambda} & \lambda \neq 0, \\ \ln[2\pi f(u,\omega)] & \lambda = 0. \end{cases} \tag{13.3}$$

The model generalizes the specification by [39] to the locally stationary case. Alternative specifications are encompassed: a time-varying autoregressive (AR) model of order K arises in the case $\lambda = -1$ (inverse link), whereas $\lambda = 1$ (identity link) amounts to fitting a time-varying moving average (MA) model of order K to the series. For $\lambda = 0$ we obtain the locally stationary exponential model proposed by [4] in the globally stationary case, and considered by [25].

Remark 13.2 The coefficients $c_\lambda(u,k)$ are related to the local generalized autocovariances, see [38], which are obtained by the Fourier inversion of $[2\pi f(u,\omega)]^\lambda$, namely, $\gamma_\lambda(u,k) = \frac{1}{2\pi}\int_{-\pi}^{\pi}[2\pi f(u,\omega)]^\lambda e^{\imath\omega k}\mathrm{d}\omega$. When $\lambda \neq 0$,

$$\begin{aligned} [2\pi f(u,\omega)]^\lambda &= 1 + \lambda g_\lambda(u,\omega) \\ &= 1 + \lambda c_\lambda(u,0) + 2\sum_{k=1}^{K}\lambda c_\lambda(u,k)\cos(\omega k), \end{aligned} \tag{13.4}$$

which, combined with the above definition of generalized autocovariances, yields

$$\begin{aligned} 1 + \lambda c_\lambda(u,0) &= \gamma_\lambda(u,0), \\ \lambda c_\lambda(u,k) &= \gamma_\lambda(u,k), \quad k \neq 0, \end{aligned}$$

i.e., the coefficients $c_\lambda(u,k)$ are linear functions of the time-varying generalized autocovariances. When $\lambda = 0$, $c_0(u,k)$ is directly interpretable as a time-varying cepstral coefficient.

For a given λ, the variation of the sdf over time depends solely on the functions $c_\lambda(u,k)$, $k = 0,\ldots,K$. For $\lambda = 0$, the only restriction is that they are continuous in u. For $\lambda \neq 0$, they need to satisfy the positivity constraint

$$1 + \lambda g_\lambda(u,\omega) > 0. \tag{13.5}$$

Moreover, Assumption 13.1 is satisfied provided that $\{1 + \lambda g_\lambda(u,\omega)\}^{1/\lambda} \leq M$ for some $0 < M < \infty$. The restrictions on $g_\lambda(u,\omega)$ will be enforced by a reparameterization, discussed in Sect. 13.2.1.

Note that the instantaneous sdf admits the factorization

$$f(u,\omega) = \frac{1}{2\pi}\sigma^2(u)\psi(u,\omega)\overline{\psi(u,\omega)},$$

where $\psi(u,\omega) = \sum_{j=0}^{\infty} \psi_j(u) e^{-\imath\omega j}$ and $\psi_0(u) = 1$. Under the condition (13.5), for $\lambda \neq 0$, the time-varying coefficients $\psi_j(u)$ are obtained using the recursive formula in [36], i.e.,

$$\psi_j(u) = j^{-1} \sum_{k=1}^{j\wedge K} k c_\lambda(u,k)\psi_{j-k}(u), \quad j = 1, 2, \ldots \tag{13.6}$$

and $\psi_0\,(u) = 1$. When $\lambda = 0$, $\sigma^2 = \exp(c_0(u))$ and

$$\psi(u, e^{-\imath\omega}) = \exp(\sum_k c_\lambda(u,k)\exp(-\imath\omega k)).$$

13.2.1 Reparameterization

The positivity constraint (13.5), for $\lambda \neq 0$, is enforced by a reparameterization based of the representation theorem of Fejér and Riesz for non-negative trigonometric polynomials, see [23, pp. 20 and 21]. In particular, we consider the factorization

$$1 + \lambda g_\lambda(u,\omega) = \sigma_\lambda^2(u)|\beta_\lambda(u,\omega)|^2,$$

where $|\beta_\lambda(u,\omega)|^2 = \beta_\lambda(u, e^{-\imath\omega})\beta_\lambda(u, e^{\imath\omega})$, and $\beta_\lambda(u, e^{-\imath\omega}) = 1 + \beta_\lambda(u,1)e^{-\imath\omega} + \cdots + \beta_\lambda(u,K)e^{-\imath\omega K}$. As a result, for $\lambda \neq 0$, the sdf is parameterized as

$$2\pi f(u,\omega) = \left[\sigma_\lambda^2(u)|\beta_\lambda(u,\omega)|^2\right]^{\frac{1}{\lambda}}. \tag{13.7}$$

The coefficients $\beta_\lambda^2(u,1)$ can be obtained by the coefficients of the infinite MA representation of the original process X_{tn}, through the recursion

$$\beta_\lambda(u,k) = \frac{1}{k}\sum_{h=1}^{k}[h(\lambda+1) - k]\psi_h(u)\beta_\lambda(u,k-h), \quad k = 1, \ldots, K,$$

with $\beta_\lambda(u,0) = 1$, and the normalizing constant $\sigma_\lambda^{2/\lambda}(u)$ is interpreted as the local prediction error variance of the locally stationary process, as $\sigma_\lambda^{2/\lambda}(u) = \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln[2\pi f(u,\omega)]d\omega\right\}$. The local prediction error variance was discussed in detail by [28], while a further discussion on prediction in L^p was provided by [29]. The relation between the time-varying parameters $c_\lambda(u,k)$ and $\beta_\lambda(u,k)$ is obtained by Eqs. (13.4) and (13.7),

$$\begin{aligned} c_\lambda(u,0) &= \tfrac{1}{\lambda}\left[\sigma_\lambda^2(u)(1 + \beta_\lambda^2(u,1) + \cdots + \beta_\lambda^2(u,K)) - 1\right], \\ c_\lambda(u,k) &= \tfrac{1}{\lambda}\sigma_\lambda^2(u)\textstyle\sum_{j=k}^{K}\beta_\lambda(u,j)\beta_\lambda(u,j-k). \end{aligned} \tag{13.8}$$

The coefficients of the infinite MA representation are obtained by applying the Gould formula [22]:

$$\psi_j(u) = \frac{1}{\lambda j}\sum_{k=1}^{j\wedge K}\{k(\lambda+1)-\lambda j\}\beta_\lambda(u,k)\psi_{j-k}(u),\ j>0, \quad \psi_0(u)=1, \quad \lambda\in\mathbb{R}. \tag{13.9}$$

Example 13.1 Let $K = 1$ in (13.1). Setting $\lambda = 1$,

$$f(u,\omega) = (2\pi)^{-1}\{1 + c_1(u,0) + 2c_1(u,1)\cos\omega\}$$

is the time-varying sdf of a locally stationary MA process of order 1. The generalized cepstral coefficients are continuous in u and need to satisfy the following two conditions: $c_1(u,0) > -1$ and $|c_1(u,1)/(1+c_1(u,0))| < 0.5$. To guarantee these conditions we can reparameterize them in terms of $\sigma^2(u) > 0$ and $|\beta_1(u,1)| < 1$, so that $1 + c_1(u,0) = \sigma^2(u)\left\{1+\beta_1^2(u,1)\right\}$, and $c_1(u,1) = \sigma^2(u)\beta_1(u,1)$.

If $\lambda = -1$, under the same restrictions, (13.1) provides the instantaneous sdf of a locally stationary AR process of order 1. The coefficients of the MA representation are $\psi_j(u) = (-1)^j\beta_{-1}^j(u,1)$.

The case $\lambda = 0$ generates the time-varying exponential process for the sdf by [4]: the cepstral coefficients are now unrestricted and the time-varying MA coefficients are $\psi_j(u) = (j!)^{-1}c_0^j(u,1)$.

In the general case, $c_\lambda(u,0) > -\lambda^{-1}$ and $|\lambda c_\lambda(u,1)| < (1+\lambda c_\lambda(u,0))/2$ guarantee the positivity condition. A simple reparameterization in terms of $\sigma^2(u) > 0$ and $|\beta_\lambda(u,1)| < 1$ will ensure this. Finally, the MA coefficients are $\psi_j(u) = (-1)^j\beta_\lambda^j(u,1)\prod_{k=1}^j\left(1-\frac{1+\lambda}{\lambda k}\right)$.

13.2.2 Modeling Structural Change: Local Generalized Cepstral Coefficients

The generalized spectral coefficients $c_\lambda(u,k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} g_\lambda(u,\omega)\cos(\omega k)\mathrm{d}\omega,\ k = 0,1,\ldots,K$, are assumed to vary smoothly according to the following dynamics:

$$c_\lambda(u,k) = G_\lambda\left(\theta_{k0} + \sum_{i=1}^{q}\theta_{ki}\Psi(u;\gamma_i,\tau_i)\right), \tag{13.10}$$

where, for $\lambda = 0$, G_0 is the identity function, and, for $\lambda \neq 0$, G_λ, is a non-linear smooth function of the structural parameters θ_{ki}, $k = 0,1,\ldots,K$, $i = 0,1,\ldots,q$, constrained so as to guarantee that the sdf is positive.

The functions $\Psi(u;\gamma_i,\tau_i)$ are the logistic functions

$$\Psi(u; \gamma_i, \tau_i) = \frac{1}{1 + \exp(-\gamma_i(u - \tau_i))}, \tag{13.11}$$

taking values in (0, 1) and depending on the location parameter $\tau_i \in (0, 1)$, and the shape, or smoothness, parameter $\gamma_i > 0$. The function describes a smooth transition occurring at time τ_i. The location parameters, $\tau_i = t_i/n$, $0 < \tau_1 < \tau_2 < \cdots < \tau_q < 1$, regulate the timing of the transition to a new regime, while the parameters γ_i determine the speed of the transition: if $\gamma_i \to \infty$, then the transition function tends to the step function $I(u \geq \tau_i)$. On the other hand, if $\gamma_i \to 0$, $\Psi(u; \gamma_i, \tau_i) \to 1/2$. If $q = 1$, the coefficients vary in the range $(\theta_{k0}, \theta_{k0} + \theta_{k1})$; if in addition γ_1 is small and τ_1 is close to 0.5, the structural change takes the form of a trend in the time-varying parameters.

The q logistic functions described in Eq. (13.11) form a flexible basis for the dynamics of $c_\lambda(u, k)$ and thus for the variation of the sdf. This basis has a long tradition in the time series literature, see [2, 21, 47].

13.2.3 Structural Changes Preserving Local Stationarity

The non-linear mapping G_λ defined in Eq. (13.14) preserves the local stationarity of the process when allowing for structural changes. In particular, we need to guarantee that $\sigma^2_\lambda(u) > 0$ and $0 < b \leq |\beta_\lambda(u, \omega)|^2 \leq B < \infty$.

The second condition is enforced by a convenient transformation of the coefficients $\beta_\lambda(u, k)$ in terms of the generalized partial autocorrelation coefficients $\zeta_\lambda(u, k)$, in the range $(-1, 1)$. By means of a recursive procedure, due to [3], and further extended by [30], the coefficients $\beta_\lambda(u, k)$, $k = 1, \ldots, K$, are obtained from the last iteration (the $k-1$-th) of the following Durbin–Levinson recursions [18, 27]:

$$\begin{aligned} &\beta^{(k)}_\lambda(u, k) = \zeta_\lambda(u, k), \\ &\beta^{(j)}_\lambda(u, k) = \beta^{(j)}_\lambda(u, k-1) + \zeta_\lambda(u, k)\beta^{(k-j)}_\lambda(u, k-1), \qquad j = 1, \ldots, k-1, \end{aligned} \tag{13.12}$$

for $k = 1, \ldots, K$, setting $\beta_\lambda(u, k) = \beta^{(K)}_\lambda(u, k)$. This parameterization guarantees that the roots of the characteristic equation associated with the polynomial $\beta_\lambda(u, z)$ are all in modulus greater than one for every u. Moreover, it is one to one and smooth (see [3, Theorem 2]).

In order to restrict the generalized partial correlation coefficients in the range $(-1, 1)$, we reparameterize them as the Fisher inverse transforms of the unbounded set of parameters $\vartheta_\lambda(u, k)$, by setting

$$\zeta_\lambda(u, k) = \frac{\exp(2\vartheta_\lambda(u, k)) - 1}{\exp(2\vartheta_\lambda(u, k)) + 1}. \tag{13.13}$$

Finally, we enforce the positivity of $\sigma^2_\lambda(u)$ by writing

$$\sigma_\lambda^2(u) = \exp\{2\vartheta_\lambda(u, 0)\}.$$

Eventually, the function G_λ that maps $\vartheta_\lambda(u, k)$ into $c_\lambda(u, k)$ is the composition of the following transformations:

$$G_\lambda : \vartheta_\lambda(u, k) \overset{(13.13)}{\mapsto} \zeta_\lambda(u, k) \overset{(13.12)}{\mapsto} \beta_\lambda(u, k) \overset{(13.8)}{\mapsto} c_\lambda(u, k). \quad (13.14)$$

As a result, the sdf is guaranteed to be positive, bounded away from zero and bounded from above.

The specification of the model for the time-varying sdf is then completed by setting

$$\vartheta_\lambda(u, k) = \theta_{k0} + \sum_{i=1}^{q} \theta_{ki} \Psi(u; \gamma_i, \tau_i),$$

for $k = 0, 1, 2, \dots, K$.

When $\lambda = 0$, the time-varying generalized cepstral coefficients can be directly parameterized as

$$c_0(u, k) = \theta_{k0} + \sum_{i=1}^{q} \theta_{ki} \Psi(u; \gamma_i, \tau_i), \quad (13.15)$$

i.e., the function G_0 is the identity.

13.3 Statistical Inference

Statistical inference deals with i) the choice of the Box–Cox transformation parameter λ; ii) the selection of the order K of the trigonometric polynomial $c_\lambda(u, \omega)$, i.e., the number of cepstral coefficients needed to represent the spectral density function; iii) the selection of the number of change points q and their locations; and iv) the estimation of the vector parameters

$$\theta = [\theta_{00}, \theta_{01}, \dots, \theta_{0q}; \theta_{10}, \theta_{11}, \dots, \theta_{1q}; \dots; \theta_{K0}, \theta_{K1}, \dots, \theta_{Kq}; \gamma_1, \dots, \gamma_q]'.$$

We set off by considering the last problem, conditioning on the triple (λ, K, q) and on the knowledge of the functions $\Psi(u; \gamma_i, \tau_i)$, whose estimation is postponed to Sect. 13.3.1.

Let us denote the time-varying sdf as $f_\theta(u, \omega)$, in order to emphasize its dependence on the structural parameters in θ. Estimation is carried out in the frequency domain, using a generalization of the Whittle likelihood for locally stationary processes due to [11].

$$-2\tilde{\ell}_n(\theta) = 2n\log(2\pi) + \frac{1}{2\pi}\frac{1}{n}\sum_{t=1}^{n}\int_{-\pi}^{\pi}\left\{\log f_\theta\left(\frac{t}{n},\omega\right) + \frac{J_n\left(\frac{t}{n},\omega\right)}{f_\theta\left(\frac{t}{n},\omega\right)}\right\}\mathrm{d}\omega, \tag{13.16}$$

where $J_n(\omega)$ is pre-periodogram (see [33]), which is localized version of the periodogram

$$J_n\left(\frac{t}{n},\omega\right) = \frac{1}{2\pi}\sum_{h:1\leq\lfloor t+\frac{1}{2}+\frac{h}{2}\rfloor,\lfloor t+\frac{1}{2}-\frac{h}{2}\rfloor\leq n} X_{\lfloor t+\frac{1}{2}+\frac{h}{2}\rfloor,n}X_{\lfloor t+\frac{1}{2}-\frac{h}{2}\rfloor,n}\exp(-\imath\omega h), \tag{13.17}$$

where $\lfloor x\rfloor$ denotes the largest integer not greater than x, and the summation runs for all the h from 1 to n, such that $\lfloor t+\frac{1\pm h}{2}\rfloor$ lies between one and n. The pre-periodogram has the following relation with the ordinary periodogram:

$$I_n(\omega) = \frac{1}{2\pi}\frac{1}{n}\sum_{t=1}^{n}J_n\left(\frac{t}{n},\omega\right),$$

so that the classical periodogram is a time average of the pre-periodogram, which makes the quasi-likelihood in (13.16) a true generalization of the Whittle likelihood, in the sense that if $f_\theta(\frac{t}{n},\omega) = f_\theta(\omega)$, i.e., the underlying process is stationary, then (13.16) is the traditional Whittle likelihood. The pre-periodogram can be regarded as a preliminary estimate of $f\left(\frac{t}{n},\omega\right)$ which has to be smoothed in the time and frequency directions in order to be a consistent estimator of $f_\theta(\frac{t}{n},\omega)$.

The Whittle estimator is given by

$$\hat{\theta}_n = \arg\min_{\theta\in\Theta}\ell_n(\theta), \tag{13.18}$$

where

$$\ell_n(\theta) = -\frac{1}{2\pi}\frac{1}{n}\sum_{t=1}^{n}\int_{-\pi}^{\pi}\left\{\log f_\theta\left(\frac{t}{n},\omega\right) + \frac{J_n\left(\frac{t}{n},\omega\right)}{f_\theta\left(\frac{t}{n},\omega\right)}\right\}\mathrm{d}\omega. \tag{13.19}$$

In order to derive the asymptotic properties of the estimator $\hat{\theta}_n$, let

$$\ell(\theta) = -\frac{1}{2\pi}\int_0^1\int_{-\pi}^{\pi}\left\{\log f_\theta(u,\omega) + \frac{f(u,\omega)}{f_\theta(u,\omega)}\right\}\mathrm{d}\omega, \tag{13.20}$$

be such that $\mathrm{E}(\ell_n(\theta)) = \ell(\theta)$ and let

$$\theta_0 = \arg\min_{\theta\in\Theta}\ell(\theta).$$

Further conditions are imposed to the class of sdfs implied by our modeling procedure,

$$\mathcal{F}_\lambda = \left\{ f_\theta(u,\omega) = \frac{1}{2\pi}\left[\sigma_\lambda^2(u)|\beta_\lambda(u,\omega)|^2\right]^{\frac{1}{\lambda}}, \lambda \in \mathbb{R} \right\}.$$

In essence, it is required that the implied spectral density $f_\theta(u,\omega) \in \mathcal{F}_\lambda$ is bounded away from zero, as in Assumption 13.1. Note that, by construction, the reciprocal $1/(f_\theta(u,\omega))$ is also an element of $\mathcal{F}_\lambda$ (case when $\lambda = -1$) and the functions in $\mathcal{F}_\lambda$ are not indexed by n.

Before stating the assumptions required for the asymptotic theory to be valid, let us define the total variation of a function $\phi(u,\omega)$ that in our case will be a transformation of the sdf:

$$V(\phi(u,\cdot)) = \sup\left\{ \sum_{k=1}^{\ell} |\phi(u_{k+1},\cdot) - \phi(u_k,\cdot)| : 0 \le u_0 < \cdots \le u_\ell = 1, \ell \in \mathbb{N} \right\},$$

and similarly,

$$V^2(\phi) = \sup\Bigg\{ \sum_{j,k=1}^{\ell,m} |\phi(u_j,\omega_k) - \phi(u_{j-1},\omega_k) - \phi(u_j,\omega_{k-1}) + \phi(u_{j-1},\omega_{k-1})| : \\ 0 \le u_0 < \cdots < u_\ell \le 1;\ \ -\pi \le \omega_0 < \cdots < \omega_m \le \pi;\ \ \ell, m \in \mathbb{N} \Bigg\}.$$

Also, let

$$\rho_2(\phi) = \left(\int_0^1 \int_{-\pi}^{\pi} |\phi(u,\omega)^2| \mathrm{d}\omega \mathrm{d}u \right)^{1/2},$$

and define

$$\hat{\phi}(u,k) = \int_{-\pi}^{\pi} \phi(u,\omega)\exp\{i\omega k\}\mathrm{d}\omega,$$

so that

$$v_\Sigma(\phi) = \sum_{k=-\infty}^{\infty} V(\hat{\phi}(\cdot,k)).$$

Moreover, we set

$$\|\phi\|_{\infty,V} = \sup_u V(\phi(u,\cdot)), \qquad \|\phi\|_{V,\infty} = \sup_\omega V(\phi(\cdot,\omega)),$$
$$\|\phi\|_{V,V} = V^2(\phi), \qquad \|\phi\|_{\infty,\infty} = \sup_{u,\omega} |\phi(u,\omega)|,$$

and let

$$\tau_{\infty,V} = \sup_{\phi\in\Phi} \|\phi\|_{\infty,V}, \qquad \tau_{V,\infty} = \sup_{\phi\in\Phi} \|\phi\|_{V,\infty},$$
$$\tau_{V,V} = \sup_{\phi\in\Phi} \|\phi\|_{V,V}, \qquad \tau_{\infty,\infty} = \sup_{\phi\in\Phi} \|\phi\|_{\infty,\infty}.$$

The following assumptions are required for consistency of the Whittle estimator $\hat{\theta}_n$.

Assumption 13.2 $\mathrm{E}(|\varepsilon|^k) \leq C_\varepsilon^k, \forall k \in \mathbb{N}$, if $\tau_{\infty,V}, \tau_{V,\infty}, \tau_{V,V}, \tau_{\infty,\infty}$ are finite.

Assumption 13.3 The Whittle likelihood $\ell(\theta)$ in (13.20) is continuous and has a unique minimum at $\theta_0 \in \Theta$, where Θ is a compact subset of $\mathbb{R}^{(K+1)\times(q+1)}$.

Assumption 13.4 The change points τ_i are distinct and the smoothness parameters $\gamma_i, i = 1, \ldots q$ are strictly positive.

The last assumption implies that the parameters θ_{ki} are identified.

Theorem 13.1 *Under Assumptions 13.1–13.4,*

$$\hat{\theta}_n \to_p \theta_0.$$

Proof See Sect. 13.6. □

The proof is largely based on the theory of empirical processes for locally stationary processes derived by [13, 14]. Note that the strong moment condition for the noise term, $\mathrm{E}(|\varepsilon_t|^k) \leq C_\varepsilon^k, \forall k \in \mathbb{N}$, compensates the weak Assumption 13.1 on the parameter curves. No Gaussian assumption is made here.

We now move to asymptotic normality. Let $\nabla = (\partial/(\partial\theta_{00}), \ldots, \partial/(\partial\theta_{Kq}), \partial/(\partial\gamma_1), \ldots, \partial/(\partial\gamma_q))'$, and $\nabla^2 = (\partial^2/(\partial^2\theta_{00}), \ldots, \partial^2/(\partial^2\theta_{Kq}), \ldots, \partial^2/(\partial\theta_{ki}\partial\gamma_i), \ldots, \partial^2/(\partial^2\gamma_1), \ldots, \partial^2/(\partial^2\gamma_q))'$. Then, we have

$$\nabla\ell_n(\theta) = \frac{1}{2\pi}\frac{1}{n}\sum_{t=1}^{n}\int_{-\pi}^{\pi}\left\{\left(\frac{J_n(\frac{t}{n},\omega)}{f_\theta(\frac{t}{n},\omega)} - 1\right)\frac{\nabla f_\theta(\frac{t}{n},\omega)}{f_\theta(\frac{t}{n},\omega)}\right\}\mathrm{d}\omega, \tag{13.21}$$

and

$$\begin{aligned}\nabla^2\ell_n(\theta) =& \frac{1}{2\pi}\frac{1}{n}\sum_{t=1}^{n}\int_{-\pi}^{\pi}\Bigg\{\left(\frac{J_n(\frac{t}{n},\omega)}{f_\theta(\frac{t}{n},\omega)} - 1\right)\frac{\nabla^2 f_\theta(\frac{t}{n},\omega)}{f_\theta(\frac{t}{n},\omega)}\\ &- 2\frac{J_n(\frac{t}{n},\omega)}{f_\theta^3(\frac{t}{n},\omega)}\nabla f_\theta\left(\frac{t}{n},\omega\right)\nabla f_\theta\left(\frac{t}{n},\omega\right)'\\ &+ \frac{1}{f_\theta^2(\frac{t}{n},\omega)}\nabla f_\theta\left(\frac{t}{n},\omega\right)\nabla f_\theta\left(\frac{t}{n},\omega\right)'\Bigg\}\mathrm{d}\omega.\end{aligned} \tag{13.22}$$

Note that the derivatives $\nabla f_\theta(\frac{t}{n}, \omega)$ and $\nabla^2 f_\theta(\frac{t}{n}, \omega)$, or equivalently $\nabla f_\theta(u, \omega)$ and $\nabla^2 f_\theta(u, \omega)$, depend on the derivatives $\nabla\sigma^2_\lambda(u)$, $\nabla^2\sigma^2_\lambda(u)$, $\nabla\beta_\lambda(u, \omega)$, and $\nabla^2\beta_\lambda(u, \omega)$. These derivatives are found in Sect. 13.6.

Theorem 13.2 *Under Assumptions 13.1–13.4,*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \to_d N(0, V),$$

where $V = I(\theta_0)^{-1}$ *and* $I(\theta_0) = -\mathbb{E}[\nabla^2\ell_n(\theta_0)]$ *is the Fisher information matrix given by*

$$I(\theta_0) = \int_0^1 \int_{-\pi}^{\pi} \frac{1}{f_\theta^2(u, \omega)} \nabla f_\theta(u, \omega)\nabla f_\theta(u, \omega)' \mathrm{d}u\mathrm{d}\omega.$$

Proof See Sect. 13.6. □

13.3.1 Model Selection

Model selection consists of choosing the global parameters in the triple (λ, K, q), the location of the change points, and the smoothness of the transition across regimes, regulated by the parameters γ_i in (13.11).

The selection is challenging and, in order to make it feasible, we restrict our specification search to the most typical values of λ, which are in $\mathcal{L} = \{-1, 0, 1\}$. Moreover, we expect K to be small, given that we allow for parameter changes to occur. Third, we consider the smoothness parameter to be invariant to the timing of the change point, i.e., $\gamma_i = \gamma$ and we restrict attention to $\mathcal{G} = \{10, 50, 100\}$, the first value corresponding to a monotonic trend in (0,1) and the last to a sharp transition. Finally, and perhaps more importantly, we perform a sequential search methodology proposed by [21], named QuickShift, which is a special case of the more general QuickNet proposed by [47] for specifying the number of hidden units in an artificial neural network model. See also [2] for a more recent application of QuickShift.

The algorithm is given as follows:

(1) Choose $\lambda \in \mathcal{L}$, and $K \in \mathcal{K} = \{1, \dots, K_{\max}\}$.
(2) Estimate the parameters θ_{k0}, $k = 0, \dots, K$ of the globally stationary model (with $\theta_{ki} = 0$, $i = 1, \dots, K$).
(3) Choose γ in $\mathcal{G}$.
(4) Apply QuickShift to identify sequentially the individual change points, one at a time, in a set of candidate values (e.g., a grid of equally spaced points in (0,1)). Let $i = 1, \dots, q$.

  – The location τ_1 is identified by considering the candidate for which the score test for $\theta_{1k} = 0$, $k = 1, \dots K$ is a maximum. The test is discussed below.
  – Estimate the parameters θ_{k0} and θ_{k1}, $k = 1, \dots, K$ by maximizing (13.16).

– Perform a likelihood ratio test of $H_0 : \theta_{k1} = 0$. If the null is rejected, then search for an additional change point and iterate until no further point is identified (i.e., the likelihood ratio test is not significant).
– For the final specification compute the Akaike Information Criterion (AIC).

(5) Repeat (4) for different values of λ, K, and γ.
(6) Select the specification with the smallest AIC.

The score test of $H_0 : \theta_{ik} = 0, i > 0, k = 1, \ldots, K$ is conducted as follows. The partial derivatives of the sdf with respect to the parameters are

$$\begin{aligned}\frac{\partial}{\partial \theta_{ik}} f(u,\omega) &= \frac{1}{2\pi}\left(1+\lambda c_\lambda(u)' z(\omega)\right)^{1/\lambda-1}\Psi(u;\gamma_i,\tau_i) z_k(\omega)\\ &= \frac{f(u,\omega)}{(1+\lambda c_\lambda(u)'z(\omega))}\Psi(u;\gamma_i,\tau_i) z_k(\omega),\end{aligned}$$

where $c_\lambda(u) = [c_\lambda u, 0, \ldots, c_\lambda u, K]'$ and $z(\omega) = [1, 2\cos(\omega), \ldots, 2\cos(\omega K)]'$, so that the partial score is

$$\begin{aligned}\frac{\partial \ell_n}{\partial \theta_{ik}} &= \frac{1}{2\pi n}\sum_t\sum_j\left[\frac{J_n\left(\frac{t}{n},\omega_j\right)}{f_\theta\left(\frac{t}{n},\omega_j\right)}-1\right]\frac{1}{f_\theta\left(\frac{t}{n},\omega_j\right)}\frac{\partial f_\theta\left(\frac{t}{n},\omega_j\right)}{\partial\theta_{\lambda k}}\\ &= \frac{1}{2\pi n}\sum_t\sum_j\left[\frac{J_n\left(\frac{t}{n},\omega_j\right)}{f_\theta\left(\frac{t}{n},\omega_j\right)}-1\right]\frac{1}{\left(1+\lambda c_\lambda(t/n)'z(\omega_j)\right)}\Psi(t/n;\gamma_i,\tau_i) z_k(\omega_j).\end{aligned} \tag{13.23}$$

The expectation of the second derivatives, after a sign change, is

$$\begin{aligned}\mathrm{E}\left(-\frac{\partial^2\ell}{(\partial\theta_{ik})^2}\right) &= \frac{1}{2\pi n}\sum_t\sum_j\frac{1}{\left(1+\lambda c_\lambda(t/n)'z(\omega_j)\right)^2}\Psi^2(t/n;\gamma_i,\tau_i) z_k^2(\omega_j)\\ \mathrm{E}\left(-\frac{\partial^2\ell}{\partial\theta_{ik}\partial\theta_{ir}}\right) &= \frac{1}{2\pi n}\sum_t\sum_j\frac{1}{\left(1+\lambda c_\lambda(t/n)'z(\omega_j)\right)^2}\Psi^2(u;\gamma_i,\tau_i) z_k(\omega_j) z_r(\omega_j),\end{aligned} \tag{13.24}$$

then the test statistic is obtained as n times the coefficient of determination of the regression of Pearson's residuals $\frac{J_n\left(\frac{t}{n},\omega_j\right)}{f_\theta\left(\frac{t}{n},\omega_j\right)} - 1$ on the explanatory variables $\frac{1}{\left(1+\lambda c_\lambda(t/n)'z(\omega_j)\right)}\Psi(t/n;\gamma_i,\tau_i) z(\omega_j)$.

13.4 Illustration

We illustrate our approach with reference to the annualized quarterly growth rate of the U.S. Gross Domestic Product. The quarterly time series, plotted in Fig. 13.1, is considered from 1947.2 to 2019.4.

The generalized linear model for the sdf has been fitted for the values of the transformation parameter $\lambda = -1, 0, 1$ and for $K = 1$. Hence, we compare the fit of a time-varying first-order AR model with that of an exponential and a MA model. Table 13.1 displays the values of the maximized Whittle likelihood and of the AIC as the number of change points varies from zero (time-invariant model, featuring 2 parameters, θ_{00} and θ_{01}), up to three. For a given q the minimum AIC is obtained for $\lambda = -1$, which can be considered as the optimal transformation parameter for

Fig. 13.1 Annualized quarterly growth rate of the US gross domestic product for the period 1947.2–2019.4

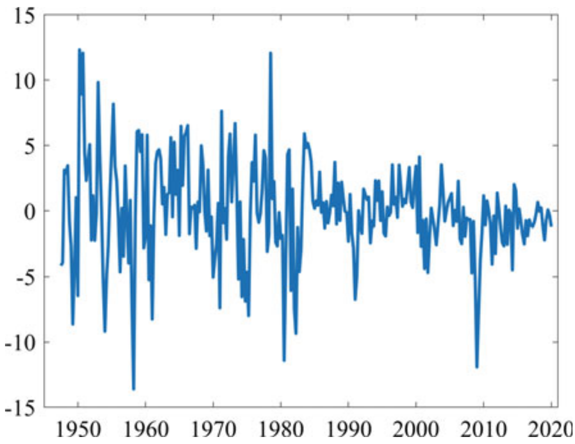


Table 13.1 Maximized Whittle likelihood and AIC for first-order ($K = 1$) generalized linear sdfs with q change points and transformation parameter $\lambda = -1, 0, 1$

		$\lambda = -1$		$\lambda = 0$		$\lambda = 1$	
q	N. parameters	Likelihood	AIC	Likelihood	AIC	Likelihood	AIC
0	2	−75.78	157.56	−76.62	159.25	−77.61	161.21
1	4	−64.19	136.38	−65.95	139.90	−67.13	142.26
2	6	−61.60	135.21	−63.27	138.53	−64.43	140.86
3	8	−60.71	137.41	−62.02	140.04	−63.04	142.09

the series. Increasing the order K does not lead to an improvement of the fit. The number of change points that is chosen as $q = 2$. The first change point is located in the fourth quarter of 1983 and corresponds to the inception of the period referred to as "the great moderation". Incorporating this into the model leads to a seizable increase in the likelihood.

The second change point occurs in the third quarter of 2016, where a further reduction of the volatility of the series seems to have taken place. The third change point is located in the fourth quarter of 2007 and marks the inception of a recessionary period triggered by the great financial crisis. Though its inclusion leads to an increase of the likelihood, AIC suggests considering $q = 2$, although we suspect that a loss of power results from allowing both prediction error variance and AR parameters to change.

Figure 13.2 displays the estimated time-varying sdf with three change points. The role of the great moderation can be appreciated, leading to a sizeable reduction in the variance of GDP growth. The great financial crisis led to an increase of the volatility and the persistence of GDP fluctuations. Both underwent a reduction after 2016. It is interesting to look at the implied patterns of the one-step-ahead prediction error variance, $\hat{\sigma}_{-1}^{-2}(u)$ and of the persistence parameter, here represented by the time-varying AR parameter, $-\hat{\beta}_{-1}(u, 1)$, see Fig. 13.3. The former has a sharp decline

Fig. 13.2 Estimated time-varying sdf of GDP growth

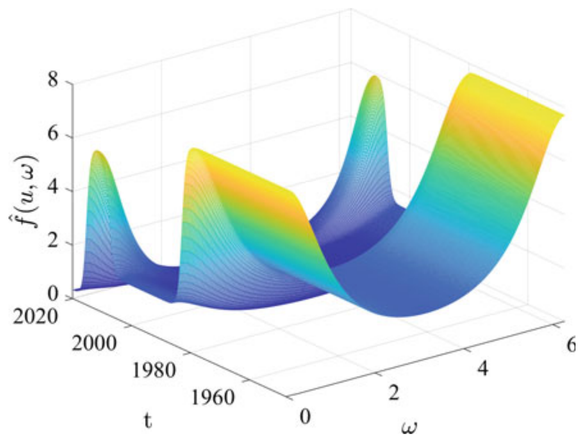
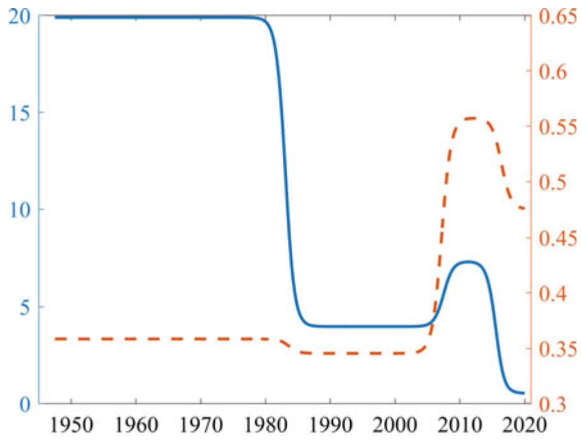


Fig. 13.3 Time-varying prediction error variance, $\hat{\sigma}_{-1}^{-2}(u)$ (solid line), and AR parameter, $-\hat{\beta}_{-1}(u, 1)$ (dashed line)



during the great moderation, and is very low after a recovery that followed the great recession. Interestingly, the persistence parameter also decreases during the great moderation, but only slightly, so that the unconditional variance decreases sharply.

13.5 Concluding Remarks

A flexible modeling framework for locally stationary processes has been developed in the frequency domain, which was illustrated to be effective in capturing the time-varying features of a real dataset, such as the time-varying sdf and prediction error variance, including change points. One limitation of the approach developed in this paper is related to the fact that the change affects simultaneously all the parameters of the model. We leave for future consideration the strategy of allowing for structural changes that are specific to an individual parameter.

13.6 Proofs

Proof of Theorem 13.1

The proofs of consistency and asymptotic normality of the estimator $\hat{\theta}_n$ are based on results by [14] on the empirical spectral process

$$E_n(\phi) = \sqrt{n}(F_n(\phi) - F(\phi))$$

evaluated at $\phi = f_\theta(\cdot, \omega)^{-1}$ and at $\phi = \frac{\partial}{\partial\theta} f_\theta(\cdot, \omega)^{-1}$, respectively, where

$$F(\phi) = \int_0^1 \int_{-\pi}^{\pi} \phi(u, \omega) f_\theta(u, \omega) \mathrm{d}\omega$$

is a spectral measure and

$$F_n(\phi) = \frac{1}{n} \sum_{t=1}^{n} \int_{-\pi}^{\pi} \phi\left(\frac{t}{n}, \omega\right) J_n\left(\frac{t}{n}, \omega\right) \mathrm{d}\omega$$

is an empirical spectral measure based on the pre-periodogram, defined in Eq. (13.17).

Under the assumptions of Theorem 13.1, almost sure consistency of $\hat{\theta}_n$ requires the uniform convergence of $\ell_n(\theta)$ to $\ell(\theta)$, defined in Eqs. (13.19) and (13.20), respectively,

$$\sup_{\theta\in\Theta} |\ell_n(\theta) - \ell(\theta)| \to 0, \quad n \to \infty.$$

Note that ([14, Example 3.1])

$$\begin{aligned}
\ell_n(\theta) - \ell(\theta) &= \int_0^1 \int_{-\pi}^{\pi} \left\{\log f_\theta(u, \omega) + \frac{f(u, \omega)}{f_\theta(u, \omega)}\right\} \mathrm{d}\omega \\
&\qquad - \frac{1}{n} \sum_{t=1}^{n} \int_{-\pi}^{\pi} \left\{\log f_\theta\left(\frac{t}{n}, \omega\right) + \frac{J_n\left(\frac{t}{n}, \omega\right)}{f_\theta\left(\frac{t}{n}, \omega\right)}\right\} \mathrm{d}\omega \\
&= \int_0^1 \int_{-\pi}^{\pi} \frac{1}{f_\theta(u, \omega)} f(u, \omega) \mathrm{d}\omega \\
&\qquad - \frac{1}{n} \sum_{t=1}^{n} \int_{-\pi}^{\pi} \frac{1}{f_\theta\left(\frac{t}{n}, \omega\right)} J_n\left(\frac{t}{n}, \omega\right) \mathrm{d}\omega + R_{\log}(f_\theta) \\
&= F(1/f_\theta) - F_n(1/f_\theta) + R_{\log}(f_\theta),
\end{aligned}$$

where F, F_n are defined above and

$$R_{\log}(f_\theta) = \int_{-\pi}^{\pi} \left[\int_0^1 \log f_\theta(u, \omega) \mathrm{d}u - \frac{1}{n} \sum_{t=1}^{n} \log f_\theta\left(\frac{t}{n}, \omega\right)\right] \mathrm{d}\omega. \tag{13.25}$$

Hence, if $\sup_{\theta\in\Theta} |R_{\log}(f_\theta)|$ is small, the convergence of $\ell_n(\theta) - \ell(\theta)$ may be controlled by $E_n(1/f_\theta)$, since, by Assumptions 13.1 and 13.2, we obtain

$$\sup_{\theta\in\Theta} |\ell_n(\theta) - \ell(\theta)| = \sup_{\theta\in\Theta} \left| \frac{1}{\sqrt{n}} E_n\left(\frac{1}{f_\theta}\right) + R_{\log}(f_\theta) \right| \leq \sup_{\theta\in\Theta} \left| \frac{1}{\sqrt{n}} E_n\left(\frac{1}{f_\theta}\right) \right| + \sup_{\theta\in\Theta} |R_{\log}(f_\theta)|.$$

Note that by Assumptions 13.1 and 13.2, we obtain that the empirical spectral process $E_n(1/f_\theta)$ has bounded variation in both the components of $f_\theta(u,\omega)$, see also the discussion in [14]. Then, the uniform convergence of $\ell_n(\theta) - \ell(\theta)$ could be easily obtained by Glivenko–Cantelli-type convergence results, see, for example, [14, Theorem 2.12], which ensures that we have

$$\sup_{\theta\in\Theta} \left| \frac{1}{\sqrt{n}} E_n\left(\frac{1}{f_\theta}\right) \right| \xrightarrow{P} 0.$$

To show that this result is valid also in our case, it suffices to verify [15, Assumption 2.4 (c)].

First, we note that by construction the reciprocal $1/(f_\theta(u,\omega))$ is an element of $\mathcal{F}_\lambda$ (when $\lambda = -1$) and the functions in $\mathcal{F}_\lambda$ are not indexed by n. Moreover, the implied spectral density $f_\theta(u,\omega) \in \mathcal{F}_\lambda$ is bounded away from zero.

Let us consider

$$\begin{aligned} f_\theta(u,\omega)^\lambda &= \sigma_\lambda^2(u)|\beta(u,\omega)|^2 \\ &= \sigma_\lambda^2(u)\left(\sum_{j=0}^{K} \beta_\lambda^2(u,j) + 2\sum_{k=1}^{K}\sum_{j=0}^{K-k} \beta_\lambda(u,j)\beta_\lambda(u,j+k)\cos(\omega k) \right). \end{aligned}$$

By Kolmogorov's formula ([7, Theorem 5.8.1]) we have

$$\int_{-\pi}^{\pi} \log f_\theta(u,\omega)\mathrm{d}\omega = 2\pi\lambda \log\left(\frac{\sigma_\lambda^2(u)}{2\pi}\right).$$

We define the class of functions

$$\mathcal{S}_\lambda = \Big\{ \sigma_\lambda^2(u) : [0,1] \mapsto \mathbb{R}_+, \text{ increasing with } 0 < L \leq \inf_u \sigma_\lambda^2(u) \leq \sup_u \sigma_\lambda^2(u) \leq B < \infty,\ \lambda \in \mathbb{R} \Big\},$$

and note that for uniformly bounded monotonic functions, the metric entropy is bounded

$$H(\eta, \mathcal{S}_\lambda, \rho_2) \leq C_2 B \epsilon^{-1},$$

for some $\epsilon > 0$, where C_2 is a constant, see [19]. Moreover, $\exists\ 0 < M_\star \leq 1 \leq M^\star < \infty$, $M_\star \leq f_\theta(u, \omega) \leq M^\star$, $\forall\ u, \omega$ and $f_\theta(u, \omega) \in \mathcal{F}_\lambda$. However, we do not need the lower bound $M_\star$ and then we may set $M_\star = 1$.

In addition, to ensure that

$$\left|1 + \sum_{k=1}^{K} \beta_\lambda(u, k) z^k\right| \neq 0,\ \forall 0 < |z| \leq 1 + \delta,\ \delta > 0,\ \lambda \in \mathbb{R},$$

we choose

$$\mathcal{B}_\lambda = \Bigg\{\boldsymbol{\beta}_\lambda(u) = (\beta_\lambda(u, 1), \ldots, \beta_\lambda(u, K))' : [0, 1]^K \mapsto \mathbb{R}^K, \\ \sup_u \left|1 + \sum_{k=1}^{K} \beta_\lambda(u, k) z^k\right| \neq 0,\ \forall 0 < |z| \leq 1 + \delta,\ \delta > 0,\ \lambda \in \mathbb{R}\Bigg\},$$

such that $\mathcal{B}_\lambda \in \mathcal{W}_2^\alpha$, that is, the Sobolev class of functions with $m \in \mathbb{N}$ smoothness parameter, such that the first $m \geq 1$ derivatives exist with finite L_2-norm, see [20]. For this class of functions, the metric entropy can be bounded by

$$H(\eta, \mathcal{B}_\lambda, \rho_2) \leq A \epsilon^{-1/m},$$

for some $A > 0$ and $\forall \epsilon > 0$. It follows that the metric entropy of the class of the time-varying sdf is bounded by

$$H(\eta, \mathcal{F}_\lambda, \rho_2) \leq A_2 B \epsilon^{-(1+1/m)}.$$

As a result, all the conditions of [13][Assumption 2.4 (c)] are easily satisfied, and hence we apply [13, Lemma A.2] to conclude that

$$\sup_{\theta \in \Theta} |R_{\log}(f_\theta)| = O\left(\frac{v_\Sigma}{n}\right),$$

which implies $\sup_{\theta \in \Theta} |R_{\log}(f_\theta)| \to 0$ as $n \to \infty$.

Moreover, by [13, Theorem 2.6], we obtain the convergence of $\rho_2(1/\hat{f}_\theta - 1/\hat{f}_\theta)$ by setting $\gamma = (1 + 1/m)$, so that

$$\rho_2\left(\frac{1}{\hat{f}_\theta}, \frac{1}{f_\theta}\right) = O_P\left(n^{-\frac{2-(1+1/m)}{4(1+1/m)}} (\log n)^{\frac{(1+1/m)-1}{2(1+1/m)}}\right), m > 1.$$

In conclusion, by compactness of Θ and the uniqueness of θ_0 implied by Assumptions 13.3 and 13.4, we obtain the claimed convergence $\hat{\theta}_n \to_p \theta_0$, which concludes the proof. □

13.7 Proof of Theorem 13.2

Before presenting the proof of Theorem 13.2, we need to check the smoothness conditions given in [9, Assumption 2.1].

First, we note that

$$\sigma_\lambda^2(u) = \exp\{2\Psi(u)'\theta_\sigma\},$$

where the $(q+1)$-vector $\Psi(u)$ is defined as

$$\Psi(u) = (1, \Psi(u, \gamma_1, \tau_1), \ldots, \Psi(u, \gamma_q, \tau_q))',$$

with $\Psi(u, \gamma_i, \tau_i) = [1 + \exp\{-\gamma_i(u - \tau_i)\}]^{-1}$ for $i = 1, \ldots, q$. Moreover $\theta_\sigma = (\theta_{00}, \ldots, \theta_{0q})'$ and define $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_q)'$.

Therefore, the first derivatives of $\sigma_\lambda^2(u)$ with respect to θ_σ and $\boldsymbol{\gamma}$ are

$$\nabla\sigma_\lambda^2(u) = \begin{bmatrix} \frac{\partial\sigma_\lambda^2(u)}{\partial\theta_\sigma} \\ \frac{\partial\sigma_\lambda^2(u)}{\partial\boldsymbol{\gamma}} \end{bmatrix},$$

where

$$\frac{\partial\sigma_\lambda^2(u)}{\partial\theta_\sigma} = 2\exp\{2\Psi(u)'\theta_\sigma\}\Psi(u),$$

and

$$\frac{\partial\sigma_\lambda^2(u)}{\partial\boldsymbol{\gamma}} = 2\exp\{2\Psi(u)'\theta_\sigma\}\Big[\theta_\sigma \odot \Psi(u) \odot (\mathbf{1}_{q+1} - \Psi(u)) \odot (\boldsymbol{u} - \boldsymbol{\tau})\Big],$$

with $\mathbf{1}_{q+1}$ being a $(q+1)$-vector of ones, $\boldsymbol{u} = (0, u, \ldots, u)' \in [0, 1]^{q+1}$ and $\boldsymbol{\tau} = (0, \tau_1, \ldots, \tau_q)'$.

The second derivatives of $\sigma_\lambda^2(u)$ with respect to θ_σ and $\boldsymbol{\gamma}$ are

$$\nabla^2\sigma_\lambda^2(u) = \begin{bmatrix} \frac{\partial^2\sigma_\lambda^2(u)}{\partial\theta_\sigma\partial\theta_\sigma'} & \frac{\partial^2\sigma_\lambda^2(u)}{\partial\theta_\sigma\partial\boldsymbol{\gamma}'} \\ \frac{\partial^2\sigma_\lambda^2(u)}{\partial\boldsymbol{\gamma}\partial\theta_\sigma'} & \frac{\partial\sigma_\lambda^2(u)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'} \end{bmatrix},$$

where

$$\frac{\partial^2 \sigma_\lambda^2(u)}{\partial \theta_\sigma \partial \theta_\sigma'} = 4\exp\{2\Psi(u)'\theta_\sigma\}\Psi(u)\Psi(u)',$$

$$\begin{aligned}\frac{\partial \sigma_\lambda^2(u)}{\partial \boldsymbol{\gamma} \partial \boldsymbol{\gamma}'} =& 4\exp\{2\Psi(u)'\theta_\sigma\} \\ & \times \left[\theta_\sigma \theta_\sigma' \odot \mathrm{diag}\Big(\Psi(u) \odot (\mathbf{1}_{q+1} - \Psi(u)) \odot (\boldsymbol{u} - \boldsymbol{\tau})\Big)^2\right] \\ & + 2\exp\{2\Psi(u)'\theta_\sigma\} \\ & \times \left[\mathrm{diag}\Big(\theta_\sigma \odot \Psi(u) \odot (\mathbf{1}_{q+1} - \Psi(u)) \odot (\mathbf{1}_{q+1} - 2\Psi(u)) \odot (\boldsymbol{u} - \boldsymbol{\tau})^2\Big)\right],\end{aligned}$$

and

$$\begin{aligned}\frac{\partial^2 \sigma_\lambda^2(u)}{\partial \theta_\sigma \partial \boldsymbol{\gamma}'} =& 4\exp\{2\Psi(u)'\theta_\sigma\} \\ & \times \left[\Psi(u)\theta_\sigma' \odot \mathrm{diag}\Big(\Psi(u) \odot (\mathbf{1}_{q+1} - \Psi(u)) \odot (\boldsymbol{u} - \boldsymbol{\tau})\Big)^2\right] \\ & + 2\exp\{2\Psi(u)'\theta_\sigma\} \\ & \times \left[\mathrm{diag}\Big(\Psi(u) \odot (\mathbf{1}_{q+1} - \Psi(u)) \odot (\boldsymbol{u} - \boldsymbol{\tau})\Big)\right].\end{aligned}$$

Furthermore, we require the first and second derivatives of $\beta_\lambda(u, \omega)$ with respect to θ_σ and $\boldsymbol{\gamma}$. However, we recall that for estimation purposes we adopt the parametrization $\beta_\lambda(u, k)$ for $k = 1, \ldots, K$ and each $\beta_\lambda(u, k)$ from the last iteration of the Durbin–Levinson recursion. As a result, we have $\beta_\lambda(u, k) = \zeta_\lambda(u, k)$, where $\zeta_\lambda(u, k)$ denotes the generalized partial autocorrelation coefficients, and we set, for $k = 1, \ldots, K$,

$$\zeta_\lambda(u, k) = \frac{\exp\{2\Psi(u)'\theta_\sigma\} - 1}{\exp\{2\Psi(u)'\theta_\sigma\} + 1}.$$

Thus, it suffices to derive $\zeta_\lambda(u, k)$ with respect to θ_σ and $\boldsymbol{\gamma}$. Then we have

$$\nabla \zeta_\lambda(u, k) = \begin{bmatrix} \frac{\partial \zeta_\lambda(u,k)}{\partial \theta_\sigma} \\ \frac{\partial \zeta_\lambda(u,k)}{\partial \boldsymbol{\gamma}} \end{bmatrix},$$

where

$$\frac{\partial \zeta_\lambda(u, k)}{\partial \theta_\sigma} = \frac{4\exp\{2\Psi(u)'\theta_\sigma\}}{(\exp\{2\Psi(u)'\theta_\sigma\} + 1)^2}\Psi(u)$$

and

$$\frac{\partial \zeta_\lambda(u, k)}{\partial \boldsymbol{\gamma}} = \frac{4\exp\{2\Psi(u)'\theta_\sigma\}}{(\exp\{2\Psi(u)'\theta_\sigma\} + 1)^2}\Big[\theta_\sigma \odot \Psi(u) \odot (\mathbf{1}_{q+1} - \Psi(u)) \odot (\boldsymbol{u} - \boldsymbol{\tau})\Big].$$

The second derivatives of $\zeta_\lambda(u,k)$ with respect to θ_σ and $\boldsymbol{\gamma}$ are

$$\nabla^2\zeta_\lambda(u,k) = \begin{bmatrix} \frac{\partial^2\zeta_\lambda(u,k)}{\partial\theta_\sigma\partial\theta_\sigma'} & \frac{\partial^2\zeta_\lambda(u,k)}{\partial\theta_\sigma\partial\boldsymbol{\gamma}'} \\ \frac{\partial^2\zeta_\lambda(u,k)}{\partial\boldsymbol{\gamma}\partial\theta_\sigma'} & \frac{\partial\zeta_\lambda(u,k)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'} \end{bmatrix},$$

where

$$\frac{\partial^2\zeta_\lambda(u,k)}{\partial\theta_\sigma\partial\theta_\sigma'} = \frac{8\exp\{2\Psi(u)'\theta_\sigma\}(1-\exp\{2\Psi(u)'\theta_\sigma\})}{(\exp\{2\Psi(u)'\theta_\sigma\}+1)^3}\Psi(u)\Psi(u)',$$

$$\begin{aligned}\frac{\partial\zeta_\lambda(u,k)}{\partial\boldsymbol{\gamma}\partial\boldsymbol{\gamma}'} =& \frac{8\exp\{2\Psi(u)'\theta_\sigma\}(1-\exp\{2\Psi(u)'\theta_\sigma\})}{(\exp\{2\Psi(u)'\theta_\sigma\}+1)^3} \\ &\times\left[\Psi(u)\Psi(u)'\odot\operatorname{diag}\Big(\Psi(u)\odot(\mathbf{1}_{q+1}-\Psi(u))\odot(\boldsymbol{u}-\boldsymbol{\tau})\Big)^2\right] \\ &+\frac{4\exp\{2\Psi(u)'\theta_\sigma\}}{(\exp\{2\Psi(u)'\theta_\sigma\}+1)^2} \\ &\times\left[\operatorname{diag}\Big(\theta_\sigma\odot\Psi(u)\odot(\mathbf{1}_{q+1}-\Psi(u))\odot(\mathbf{1}_{q+1}-2\Psi(u))\odot(\boldsymbol{u}-\boldsymbol{\tau})^2\Big)\right],\end{aligned}$$

and finally

$$\begin{aligned}\frac{\partial^2\zeta_\lambda(u,k)}{\partial\theta_\sigma\partial\boldsymbol{\gamma}'} =& \frac{16\exp\{4\Psi(u)'\theta_\sigma\}}{(\exp\{2\Psi(u)'\theta_\sigma\}+1)^3} \\ &\times\left[\Psi(u)\theta_\sigma'\odot\operatorname{diag}\Big(\Psi(u)\odot(\mathbf{1}_{q+1}-\Psi(u))\odot(\boldsymbol{u}-\boldsymbol{\tau})\Big)\right] \\ &+\frac{4\exp\{2\Psi(u)'\theta_\sigma\}}{(\exp\{2\Psi(u)'\theta_\sigma\}+1)^2} \\ &\times\left[\operatorname{diag}\Big(\Psi(u)\odot(\mathbf{1}_{q+1}-\Psi(u))\odot(\boldsymbol{u}-\boldsymbol{\tau})\Big)\right].\end{aligned}$$

Then, we conclude that all the derivatives derived above reveal that the first and second derivatives $\nabla\sigma_\lambda^2(u)$, $\nabla^2\sigma_\lambda^2(u)$, $\nabla\beta_\lambda(u,\omega)$, and $\nabla^2\beta_\lambda(u,\omega)$ satisfy all the smoothness conditions given in [9, Assumption 2.1]. Therefore, we can prove the asymptotic normality of our estimator in (13.18) by checking the remaining relevant conditions.

Our estimator solves the likelihood equations $\nabla\ell_n(\hat{\theta}_n)=0$. Hence the mean value theorem gives

$$\nabla\ell_n(\hat{\theta}_n)-\nabla\ell_n(\theta_0)=\nabla^2\ell_n(\theta_n^\star)(\hat{\theta}_n-\theta_0),$$

where $|\theta_n^\star-\theta_0|\le|\hat{\theta}_n-\theta_0|$ and, since $\hat{\theta}_n\in\Theta$, $\nabla\ell_n(\hat{\theta}_n)=0$. As we have proved in Theorem 1, Assumptions 13.1–13.4 ensure the consistency of $\hat{\theta}_n$, and thus a central limit theorem for $\sqrt{n}(\hat{\theta}_n - \theta_0)$ could be obtained by following analogous arguments discussed in the proof of [9, Theorem 2.4]. Then, the result follows by proving that

1. $\sqrt{n}\nabla\ell_n(\theta_0) \to_d N(\mathbf{0}, \Gamma)$,
2. $\nabla^2\ell_n(\theta_n^\star) - \nabla^2\ell_n(\theta_0) \to_p 0$, and
3. $\nabla^2\ell_n(\theta_0) \to_p \mathbb{E}[\nabla^2\ell_n(\theta_0)]$ where $\mathbb{E}[\nabla^2\ell_n(\theta_0)] = V$.

However, as in [9, Remark 2.6], if the model is correctly specified we have $\Gamma = V$, where

$$V = \int_0^1 \int_{-\pi}^{\pi} \nabla \log f_\theta(u, \omega)\nabla \log f_\theta(u, \omega)' \mathrm{d}u\mathrm{d}\omega.$$

As far as convergence in distribution of the score vector in the item Assumption 13.7, we first prove the equicontinuity of $\nabla\ell_n(\theta)$. We use similar arguments as those used in the proof of Theorem 13.1. From Eq. (13.21), one has

$$\begin{aligned}\nabla\ell(\theta) - \nabla\ell_n(\theta) = &\int_{-\pi}^{\pi}\Big[\int_0^1 \frac{f(u,\omega)}{f_\theta^2(u,\omega)}\nabla f_\theta(u,\omega)\mathrm{d}u \\ &- \frac{1}{n}\sum_{t=1}^{n}\frac{J_n(\frac{t}{n},\omega)}{f_\theta^2(\frac{t}{n},\omega)}\nabla f_\theta\Big(\frac{t}{n},\omega\Big)\Big]\mathrm{d}\omega \\ &+ \int_{-\pi}^{\pi}\Big[\int_0^1 \frac{\nabla f_\theta(u,\omega)}{f_\theta(u,\omega)}\mathrm{d}u - \frac{1}{n}\sum_{t=1}^{n}\frac{\nabla f_\theta(\frac{t}{n},\omega)}{f_\theta(\frac{t}{n},\omega)}\Big]\mathrm{d}\omega.\end{aligned}$$

By Leibniz's rule, it is easy to see that the preceding equation is equivalent to

$$\begin{aligned}\nabla\ell(\theta) - \nabla\ell_n(\theta) &= \frac{1}{\sqrt{n}}E_n\Big(\nabla\frac{1}{f_\theta}\Big) \\ &\quad + \int_{-\pi}^{\pi}\Big[\int_0^1 \nabla\log f_\theta(u,\omega)\mathrm{d}u - \frac{1}{n}\sum_{t=1}^{n}\nabla\log f_\theta\Big(\frac{t}{n},\omega\Big)\Big]\mathrm{d}\omega \\ &= E_n\Big(\nabla\frac{1}{f_\theta}\Big) \\ &\quad + \nabla\int_{-\pi}^{\pi}\Big[\int_0^1 \log f_\theta(u,\omega)\mathrm{d}u - \frac{1}{n}\sum_{t=1}^{n}\log f_\theta\Big(\frac{t}{n},\omega\Big)\Big]\mathrm{d}\omega \\ &= E_n\Big(\nabla\frac{1}{f_\theta}\Big) + \nabla R_{\log}(f_\theta),\end{aligned}$$

where $R_{\log}(f_\theta)$, defined in (13.25), was proved to be asymptotically negligible. The same bounds obtained in the proof of Theorem 13.1 can be applied to the class of functions

$$\mathcal{F}_\lambda^\nabla = \Big\{\nabla f_\theta(u, \cdot), u \in [0,1], \theta \in \Theta, \lambda \in \mathbb{R}\Big\},$$

such that, by using Assumptions 13.1 and 13.2, we get Glivenko–Cantelli-type convergence result

$$\sup_{\theta\in\theta}\left\|\frac{1}{\sqrt{n}}E_n\Big(\nabla\frac{1}{f_\theta}\Big)\right\| \to_p 0,$$

which implies the equicontinuity in probability of $\nabla\ell_n(\theta)$. Thus,

$$\sup_{\theta\in\Theta}|\nabla\ell(\theta)-\nabla\ell_n(\theta)| \to_p 0.$$

Clearly, since by Assumptions 13.3 and 13.4 the parameter space Θ is compact with a unique minimum at $\theta_0\in\Theta$, we have $\nabla\ell(\theta_0)=0$, and so, evaluating Eq. (13.21) at the true value θ_0 is equivalent to write $\nabla\ell_n(\theta_0)$ as an empirical process of the form

$$\sqrt{n}\nabla\ell_n(\theta_0)=E_n\Big(\nabla\frac{1}{f_{\theta_0}}\Big).$$

Note also that, under the assumption that the model is correctly specified, it holds that $\mathbb{E}[\nabla\ell_n(\theta_0)]=0$, and therefore, we obtain

$$n\mathbb{V}[\nabla\ell_n(\theta)]=n\mathbb{E}\Big[\int_{-\pi}^{\pi}\frac{1}{n^2}\sum_{t=1}^{n}\frac{J_n^2(\frac{t}{n},\omega)}{f_\theta^2(\frac{t}{n},\omega)}\nabla\log f_\theta\Big(\frac{t}{n},\omega\Big)\nabla\log f_\theta\Big(\frac{t}{n},\omega\Big)'\mathrm{d}\omega\Big]\to\Gamma,$$

since all the required Assumptions in [9, Lemma A.5] are fulfilled.

Now we focus on the item Assumption 13.7, which is satisfied if $\nabla^2\ell_n(\theta)$ is equicontinuous. From Eq. (13.22) and the results obtained in the proof of condition Assumption 13.7, one can express $\nabla^2\ell_n(\theta)$ as

$$\begin{aligned}\nabla^2\ell_n(\theta)&=\frac{1}{\sqrt{n}}E_n\Big(\nabla^2\frac{1}{f_\theta}\Big)+\frac{1}{n}\sum_{t=1}^{n}\int_{-\pi}^{\pi}\frac{1}{f_\theta^2(\frac{t}{n},\omega)}\nabla f_\theta\Big(\frac{t}{n},\omega\Big)\nabla f_\theta\Big(\frac{t}{n},\omega\Big)'\mathrm{d}\omega\\&=\frac{1}{\sqrt{n}}E_n\Big(\nabla^2\frac{1}{f_\theta}\Big)+\frac{1}{n}\sum_{t=1}^{n}\int_{-\pi}^{\pi}\nabla\log f_\theta\Big(\frac{t}{n},\omega\Big)\nabla\log f_\theta\Big(\frac{t}{n},\omega\Big)'\mathrm{d}\omega,\end{aligned}\tag{13.26}$$

where

$$\nabla^2 f_\theta(u,\omega)^{-1}=\frac{2}{f_\theta^3(u,\omega)}\nabla f_\theta(u,\omega)\nabla f_\theta(u,\omega)'-\frac{1}{f_\theta^2(u,\omega)}\nabla^2 f_\theta(u,\omega).$$

Therefore, by similar arguments as above, under Assumptions 13.1 and 13.2, we can show the uniform convergence

$$\sup_{\theta\in\theta}\left\|\frac{1}{\sqrt{n}}E_n\Big(\nabla^2\frac{1}{f_\theta}\Big)\right\|\to_p 0,$$

which further implies the equicontinuity in probability of $\nabla^2\ell_n(\theta)$. Indeed, the second term in (13.26) converges to its expectation as $n\to\infty$, and, hence,

$$\sup_{\theta\in\Theta}|\nabla^2\ell(\theta)-\nabla^2\ell_n(\theta)|\to_p 0.$$

The proof of the item Assumption 13.7 follows directly, since, under the maintained Assumptions, Theorem 13.1 implies that the second term of Eq. (13.26) tends to V in probability for $\theta\to\theta_0$. To conclude, all the relevant conditions stated in [9, Theorem 2.4] are fulfilled and then the asymptotic normality follows, see also [11, Theorem 3.1]. □

References

1. ADAK, S. (1998). Time-dependent spectral analysis of nonstationary time series. *Journal of the American Statistical Association* **93** 1488–1501.
2. AMADO, C. AND TERÄSVIRTA, T. (2013). Modelling volatility by variance decomposition. *Journal of Econometrics* **175** 142–153.
3. BARNDORFF-NIELSEN, O. AND SCHOU, G. (1973). On the parametrization of autoregressive models by partial autocorrelations. *Journal of Multivariate Analysis* **3** 408–419.
4. BLOOMFIELD, P. (1973). An exponential model for the spectrum of a scalar time series. *Biometrika* **60** 217–226.
5. BOGERT, B. P., HEALY, M. J. R. AND TUKEY, J. W. (1963). *The frequency analysis of time series for echoes: cepstrum, pseudo-autocovariance, cross-cepstrum, and saphe cracking*. Proceedings of the Symposium on Time Series Analysis, Wiley, New York. Chap 15, 209–243.
6. BOX, G. E. AND COX, D. R. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B* **26** 211–243.
7. BROCKWELL, P. AND DAVIS, R. (1991). *Time Series: Theory and Methods*. Springer Series in Statistics, Springer.
8. CRESSIE, N. (1993). *Statistics for Spatial Data*. Wiley Series in Probability and Statistics, Wiley.
9. DAHLHAUS, R. (1996). Maximum likelihood estimation and model selection for locally stationary processes. *Journal of Nonparametric Statistics* **6** 171–191.
10. DAHLHAUS, R. (1997). Fitting Time Series Models to Nonstationary Processes. *The Annals of Statistics* **25** 1–37.
11. DAHLHAUS, R. (2000). A likelihood approximation for locally stationary processes. *The Annals of Statistics* **28** 1762–1794.
12. DAHLHAUS, R. (2012). Locally stationary processes. 351–408.
13. DAHLHAUS, R. AND POLONIK, W. (2006). Nonparametric quasi-maximum likelihood estimation for gaussian locally stationary processes. *Ann Statist* **34** 2790–2824.
14. DAHLHAUS, R. AND POLONIK, W. (2009). Empirical spectral processes for locally stationary time series. *Bernoulli* **15** 1–39.
15. DAHLHAUS, R. AND SUBBA-RAO, S. (2006). Statistical inference for time-varying ARCH processes. *The Annals of Statistics* **34** 1075–1114.
16. DAVIS, R. A., LEE, T. C. M. AND RODRIGUEZ-YAM, G. A. (2006). Structural break estimation for nonstationary time series models. *Journal of the American Statistical Association* **101** 223–239.

17. Dette, H., Preuss, P. and Vetter, M. (2011). A measure of stationarity in locally stationary processes with applications to testing. *Journal of the American Statistical Association* **106** 1113–1124.
18. Durbin, J. (1960). The fitting of time-series models. *Revue de l'Institut International de Statistique* **28** 233–244.
19. van de Geer, S. (1993). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *The Annals of Statistics* **21** 14–44.
20. van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press.
21. González, A. and Teräsvirta, T. (2008). Modelling autoregressive processes with a shifting mean. *Studies in Nonlinear Dynamics & Econometrics* **12**.
22. Gould, H. W. (1974). Coefficient identities for powers of taylor and dirichlet series. *The American Mathematical Monthly* **81** 3–14.
23. Grenander, U. and Szegö, G. (1958). *Toeplitz forms and their applications*. Univ of California Press.
24. Guo, D. M. O. H.W. and von Sachs, R. (2003). Smoothing spline anova for time-dependent spectral analysis. *Journal of the American Statistical Association* **98** 643–652.
25. Krampe, J., Kreiss, J. P. and Paparoditis, E. (2018). Estimated Wold Representation and Spectral Density Driven Bootstrap for Time Series. *Journal of the Royal Statistical Society, Series B* **80** 703–726.
26. Kreiss, J.- P. and Paparoditis, E. (2015). Bootstrapping locally stationary processes. *Journal of the Royal Statistical Society Series B* **77** 267–290.
27. Levinson, N. (1946). The Wiener (root mean square) error criterion in filter design and prediction. *Studies in Applied Mathematics* **25** 261–278.
28. Liu, Y., Xue, Y. and Taniguchi, M. (2020). Robust linear interpolation and extrapolation of stationary time series in L^p. *Journal of Time Series Analysis* **41** 229–248.
29. Liu, Y., Taniguchi, M. and Ombao, H. (2021). Statistical inference for local Granger causality. *arXiv*: arxiv.org/abs/2103.00209.
30. Monahan, J. (1984). A note enforcing stationarity in autoregressive-moving average models. *Biometrika* **71** 403–404.
31. Nason, G. P. (2013). A test for second-order stationarity and approximate confidence intervals for localized autocovariances for locally stationary time series. *Journal of the Royal Statistical Society Series B* **75** 879–904.
32. Nason, G. P., von Sachs, R. and Kroisandt, G. (2000). Wavelet processes and adaptive estimation of the evolutionary wavelet spectrum. *Journal of the Royal Statistical Society Series B* **62** 271–292.
33. Neumann, M. H. and von Sachs, R. (1997). Wavelet thresholding in anisotropic function classes and applications to adaptive estimation of evolutionary spectra. *The Annals of Statistics* **25** 38–76.
34. Ombao, H. C., Raz, J. A., von Sachs, R. and Malow, B. A. (2001). Automatic statistical analysis of bivariate nonstationary time series. *Journal of the American Statistical Association* **96** 543–560.
35. Ombao, H. C., Von Sachs, R. and Guo, W. (2005). Slex analysis of multivariate nonstationary time series. *Journal of the American Statistical Association* **100** 519–531.
36. Pourahmadi, M. (1984). Taylor expansion of and some applications. *The American Mathematical Monthly* **91** 303–307.
37. Priestley, M. B. (1965). Evolutionary spectra and non-stationary processes. 204–237. .
38. Proietti, T. and Luati, A. (2015). The generalised autocovariance function. *Journal of Econometrics* **186** 245–257.
39. Proietti, T. and Luati, A. (2019). Generalized linear cepstral models for the spectrum of a time series. *Statistica Sinica* **29** 1561–1583.
40. Qin, L. and Wang, Y. (2008). Nonparametric spectral analysis with applications to seizure characterization using eeg time series. 1432–1451. .

41. Rosen, O., Stoffer, D. and Wood, S. (2009). Local spectral analysis via a Bayesian mixture of smoothing splines. *Journal of the American Statistical Association* **104** 249–262.
42. Rosen, O., Wood, S. and Stoffer, D. (2012). Adaptspec: Adaptive spectral estimation for nonstationary time series. *Journal of the American Statistical Association* **107** 1575–1589.
43. Sakiyama, K. and Taniguchi, M. (2003). Testing composite hypotheses for locally stationary processes. *Journal of Time Series Analysis* **24** 483–504.
44. Silverman, R. (1957). Locally stationary random processes. *IRE Transactions on Information Theory* **3** 182–187.
45. Taniguchi, M. (1980). On estimation of the integrals of certain functions of spectral density. *Journal of Applied Probability* **17** 73–83.
46. West, M., Prado, R. and Krystal, A. D. (1999). Evaluation and comparison of eeg traces: Latent structure in nonstationary time series. *Journal of the American Statistical Association* **94** 375–387.
47. White, H. (2006). Approximate nonlinear forecasting methods. *Handbook of economic forecasting* **1** 459–512.

Chapter 14
Tango: Music, Dance and Statistical Thinking

Anna Clara Monti and Pietro Scalera

Abstract After briefly reviewing the history of Argentine tango, the paper considers a model describing the elements which drive the dance and allow the attunement and the synchronization within the couple of dancers. Subsequently a statistical approach is proposed to analyze the activity of the dancers when the number of leaders and the number of followers differ.

14.1 Introduction

Tango appeared in Buenos Aires before the end of the nineteenth century. Between 1821 and 1932 Argentina was the second largest recipient of immigrants (after the United States)—mainly Italians, Spaniards and French [2]. The immigrants got together with the native born Argentines and the former slaves from African countries. Tango is the outcome of a melting pot of cultures, musics and dances from different peoples and shows the influence of different musics such the European instrumentation and harmony as well as the Cuban Habanera and the African rhythm [2, 17]. Initially it was a working-class dance, identified as the music of immigrants either for its roots and for its lyrics which reflects the melancholy of the immigrants [6]. At the beginnings of the twentieth century tango was exported from Argentina to Europe and, as a result, found acceptance in upper-class Argentinian society.

The history of tango is strictly linked to the history of Argentina. Its golden age occurs around 1940 ('*anos cuarenta*'). This is a period of outstanding creativity with great composers, arrangers, lyricists, singers and orchestra directors such as Juan D'Arienzo, Rodolfo Biagi, Carlos Di Sarli, Osvaldo Pugliese, Alfredo De Angelis,

A. C. Monti (✉)
University of Sannio, Piazza Arechi II, 82100 Benevento, Italy
e-mail: acmonti@unisannio.it

P. Scalera
Conservatorio di Musica di San Pietro a Majella, Via San Pietro a Majella 35, 80138 Naples, Italy

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_14

Miguel Caló, Annibal Troilo, Julio De Caro, Osvaldo Fresedo, Francisco Canaro, among others. Musical compositions, performers and dance styles of this period continue to be considered the yardstick against which all subsequent innovations and developments are measured [18]. Tango experienced a gloomy period during the military dictatorship (1976–1983), when its dance was discouraged. The renaissance of tango dates back to the 1980s. Ever since tango has become more and more popular around the world. See [5, 7, 12, 17] for the history of tango, from the origins until the most recent developments.

Tango is currently a global dance which attracts dancers and musicians from all the continents [18]. It is extremely difficult to estimate how many tango dancers there are around the world. Nevertheless because of its popularity an entire industry has emerged around tango dancing and tourism, involving tango academies, tango ballrooms, tango holidays, tango shoes and clothes boutiques, tango orchestras, tango CDs, and so forth [3, 6].

Tango dancing requires rapid movements and decision making. It is a beneficial physical activity, keeps the mind alert and gives a better balance. It is used as a therapy for physical and mental neurologic disorders, especially for Parkinson's disease [8, 10, 11, 15, and reference therein], multiple sclerosis [4], cancer [13], depression [14]; see [1] and reference therein.

In 2009 tango was included in the UNESCO Intangible Cultural Heritage Lists.

Tango dance includes an endless variety of movements (steps, turns, embellishments, etc.). On the other side, tango is one of the first dances in history which does not rely on choreographed steps. One of its key features is *improvisation*. Tango couples respond to the same music with diverse rhythms, movement qualities, and figures. The dance is made of relatively free sequences where each improvisational element is made up of a complex combination of jointly made forward steps, backward steps, side steps, and pivots; as well as *mixes*, in which one dancer executes a movement and the other another [9, 16]. In tango nothing is known in advance [6]. Before the music starts, neither of the dancers knows how the dance will unfold, nevertheless the couple manages to be continuously synchronized throughout the dance. Section 14.2 describes the flow of information which contribute to the dancers' performance.

A couple dancing tango is made by a *leader* ('líder' in spanish) and a *follower* ('seguidora'). In tango events dancers with one role may appreciably outnumber dancers with the other role. Section 14.3 outlines statistical modeling for the activity and the waiting time of the outnumbering role.

Few notes on tango music are in the Appendix.

14.2 Dancers' Interaction

In statistical models there is a response variable which needs to be predicted. To this end information available before the experiment and experimental information are combined into a model, which is used for prediction.

In our context, the aim is predicting the movements of the dancers by using either information available in advance or information which become available during the dance. The model is initially developed for the follower and eventually an analogous model is shown to apply also to the leader.

Usually—but not necessarily—the leader is the male dancer and the follower is the female dancer. Experienced dancers can play both roles, and some dancers may enjoy switching from one role to the other.

Tango assigns very strict roles to the leader and the follower. The leader is the dancer mainly responsible for interpreting the music, who makes the decision on the steps or figures and on the timing. Coordination between the dancers is achieved through nonverbal communication. During the dance the leader proposes the movements, by communicating his '*intention*' through a measured, but directionally very precise weight projection of his body [9]. In turn the follower picks up this piece of information and executes the response interpreting the suggested movement, its width and orientation and the speed of performance, returning a clear feedback.

Denote by $Step_t^F$ the step of the follower at time t, by $Intention_t^L$ the intention communicated by the leader at time t and by $Music_t$ the music on air at the same time. A very simple model is

$$Step_t^F = f\left(Intention_t^L, Music_t\right) + \epsilon_t. \tag{14.1}$$

The music is a further explanatory variable as it represents a constraint for both dancers. The dance needs to be harmoniously respondent to the music. The error term ϵ_t derives by various factors: the quality of the leader's guide, the follower's fitness at the moment, incidental interactions with other couples, room conditions, and so forth.

Model (14.1) is very simple and assumes that the follower's performance depends only on the current pieces of information: the leader's intention and the music on air. Indeed tango dance requires great receptivity skills, and the ability to be attuned to the other body's subtle changes of balance, timing, speed and direction [6]. Nevertheless there are objective and subjective information which can enhance the follower's interpretation.

One important kind of information is the piece of music. There are three kinds of music: '*tango*', '*vals*' and '*milonga*'.

Milonga is the most archaic kind of tango music. As other kinds of music born in the Americas during the twentieth century, it is a combination of African music with an 'ostinato' (i.e. a polyrhythmic figure called 'habanera'), which is usually played in the background throughout all the song length, and European music with respect to melody and harmony.

Tango is the natural development of milonga. It is characterized by a more steady an less syncopated rhythm. Lyrics in tango are generally sad, while melodies are usually divided in joyful and dramatic parts.

Tango vals is a form of tango on a ternary tempo. The rhythm recalls the Viennese waltz, but it relies on the melody, the lyrics and the musical instruments of tango.

Tango rhythms invite the dancers to walk. Linear and circular movements (ochos, giros, half-giros, calesita, ...) are mixed, with preference to the first ones when the music is upbeat and the latter ones when the music is downbeat. Tango admits the largest freedom in terms of steps or figures, and pauses allow dancers to perform embellishments, little movements by one of the dancer which enrich the dance.

Vals is danced in a rather flowing way. Circular movements are typical of vals, as they adapt well to a ternary tempo and comply with the fluidity of the music. The dance is faster and it is characterized by the absence of pauses.

Dancing a milonga implies a strong emphasis on rhythm. Milongas are more cheerful pieces of music which are interpreted in a more joyful way. The steps recall the basic elements of tango, but the performance is faster and the steps are usually smaller. Furthermore milonga can be danced in the 'lisa' version or with 'traspié', i.e. double steps on the 'ostinato' which make the dance even faster.

In summary the different kinds of music encourage different steps or figures, impose a different rhythm and hence a different speed in the movements.

Another piece of information is given by the 'embrace', i.e. the way dancers hold each other. There is a common agreement that there are three main kinds of embrace: the '*milonguero embrace*', the '*salon embrace*' and the '*open embrace*' [17].

In the milonguero embrace the dancers are very close to each other, their chests are connected and the follower's left arm and the leader's right arm are tightly wrapped around the partner. Little space is allowed for fancy figures. Furthermore, to keep the chest-to-chest contact, all movements must be below the hips, in the knees and in the feet. From the milonguero embrace follows the milonguero style of dancing tango. It originates from Argentine milongas where dance floors were crowded, and it is characterized by a selection of steps and techniques proper to dance in crowded places. In a milonguero style the movements are typically smaller with respect to other styles.

The salon embrace is characterized by a closed connection on the right side of the leader but there is no chest connection. This embrace allows more space for large steps and figures.

Finally in the *open embrace* only the arms of the dancers are in contact. It provides the dancers with plenty of space for their performances.

The milonguero, salon and open embrace are shown in Fig. 14.1.[1]

The three embraces affect the freedom of movements of the dancers. For instance, a closed embrace is suitable for walking, while an open embrace is appropriate for figures where the legs need room to move.

Some dancers switch from one kind of embrace to another during a single song, being in a close embrace when appropriate and opening up for moves that require extra space. Nevertheless each dancer has his/her favorite embrace which affects

[1] Noelia Hurtado and Carlitos Espinoza in the milonguero embrace (screenshot from the Youtube video https://www.youtube.com/watch?v=-xMsReD4x8U), Michelle Marsidi and Joachim Dietiker in the salon embrace (https://www.michelleyjoachim.com/photos#photos-tango-argentino-basel), Daiana Guspero and Miguel Angel Zotto in the open embrace (https://www.tangofilia.net/miguel-angel-zotto-il-volto-del-tango/#gallery-3).

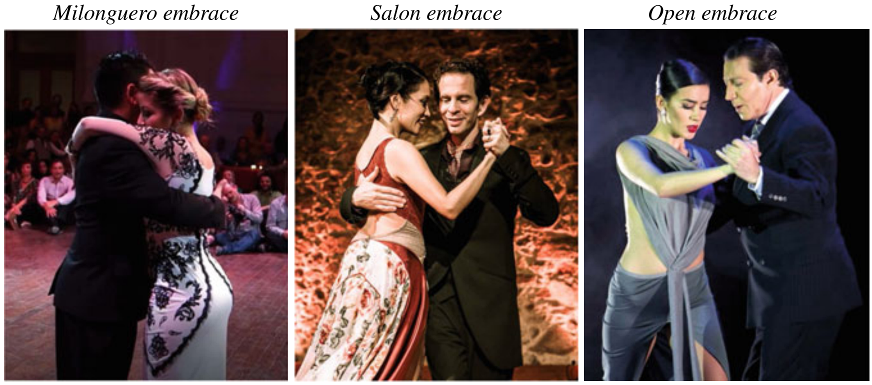


Fig. 14.1 Milonguero, salon and open embrace in tango dance

his/her variety of steps and figures. Hence the embrace is an important element to consider when trying to predict the dancers' movements.

The music on air, tango, vals or milonga, and the embrace are objective information since they are available to anyone in the room. In addition each dancer has his/her personal information.

The first piece of personal information derives from the dancer's expertise and knowledge of steps and figures. The more extensive experience the dancer has, the more extensive his/her technical skills are, the easier it is for the follower to interpret the leader's guide and respond appropriately.

The second piece of subjective information derives from the knowledge of the leader. Consciously or unconsciously, each dancer has his own dancing style characterized by a personal repertory of favorite steps and figures. The knowledge of the leader can derive from previous joint dances, but also by observing him in the dance hall. Many followers have in mind a three way array as the one displayed in Fig. 14.2 with the distribution of the steps and figures in tango, vals and milonga for various leaders.[2]

The technical skills along with the knowledge of the leader are typical of each single follower and will be indicated as the follower's expertise E_t^F at time t

$$E_t^F = \left\{ \textit{Technical skills}_t, \textit{Knowledge of the leader}_t \right\}.$$

The follower's expertise is updated any time the dancer experiments with tango, by participating to a tango event, by dancing a piece of music, by meeting a new leader, by watching other couples dancing.

Let

$$\Omega_{t-1}^F = \left\{ \textit{Kind of music, Embrace, } E_{t-1}^F \right\}$$

[2] The dancers in the pictures are (from left to right) Pablo Garcia, Pablo Veron, Miguel Angel Zotto, Carlitos Espinoza and Joachim Dietiker.

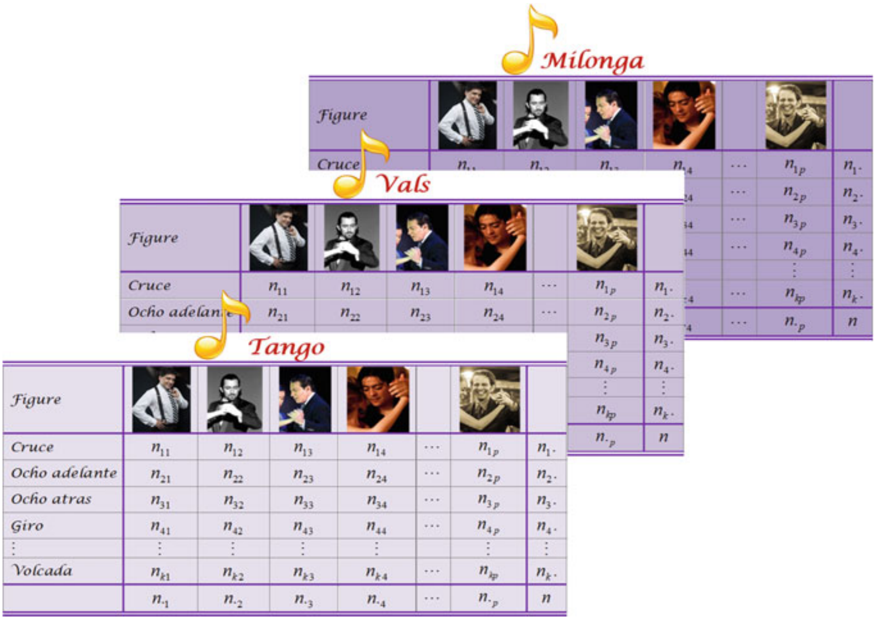


Fig. 14.2 Three-way array with the distribution of steps and figures for each kind of music and various leaders

denote the set of all the information available to the follower at time $t-1$. They give rise to a prior distribution of the steps and figures the follower may be expected to perform.

At time t the follower receives the current pieces of information, given by the leader's intention and the music on air, which can be regarded as experimental information at time t. The model, which describes the follower's step at time t, exploiting all the information available, is

$$Step_t^F = f\left(Intention_t^L, Music_t; \Omega_{t-1}^F\right) + \epsilon_t.$$

A similar model can be defined also for the leader's intention. The leader's expertise E_t^L at time t, analogously to the follower's expertise, includes the technical skills. Also for the leader the knowledge of the follower is an important piece of information. Novices may limit the variety of movements he can propose, whereas skilled followers, reactive to the guide, allow the leader more freedom in the choice of the movements.

There is however another element in the expertise of the leader, which is the knowledge of the song. While the follower is driven by the leader, leaders are responsible for providing feedforward information, planning ahead, interpreting the music [9]. In this context the knowledge of the song, which provides the awareness of how the rhythm and the melody are progressing, gives the leader more confidence in proposing the next moves. Hence the personal expertise of the leader is made of the technical skills, the knowledge of the follower and the familiarity with the song. The leader's expertise at time t is

$$E_t^L = \left\{ \textit{Technical skills}_t, \textit{Knowledge of the follower}_t, \textit{Familiarity with the song}_t \right\}.$$

The set of information available to the leader at time $t-1$ is

$$\Omega_{t-1}^L = \left\{ \textit{Kind of music}, \textit{Embrace}, E_{t-1}^L \right\}.$$

In addition to interpreting the music, the leader is also responsible to navigate the couple around the dance floor. Availability of space or crowd around the couple can augment or limit, respectively, the variety of movements the leader can propose. Let $DanceFloor_t$ denote the conditions of the dance floor around the couple at time t, the model for the leader's intention is

$$\textit{Intention}_t = f\left(\textit{Music}_t, \textit{Dancefloor}_t; \Omega_{t-1}^L\right) + \epsilon_t.$$

14.3 Activity of the Dancers and Waiting Time

An important issue when tango events take place is whether there are the same number of leaders and followers. Very often in *milongas*—the places where people dance tango – there is an excess in number of dancers with one role with respect to the other. In this context statistics can provide an approach to study the effect of this discrepancy.

In what follows, for simplicity, it will be assumed that there is an excess of followers, but of course in many cases the situation is the other way round with an excess of leaders. In such a case the analysis is the same with the role of followers and leaders swapped.

When there are more followers than leaders, not all the followers can dance at the same time. Some of them are forced to be idling, so that two questions arise:

Q.1 *how long is the idling time*?

Q.2 *how much can a follower dance*?

These two questions are clearly relevant for the dancer, the longer he or she is idling, the less the dancer gets to dance, the less enjoyable is the event. However these questions are also important for the organizers of the events, since unhappy dancers may not attend future events with negative economic relapses.

In milongas, dancers are usually expected to rotate their partners. A couple of dancers get together for a group of three or four songs, called '*tanda*', and then the dancers move to the next partners. Typically there are two tandas of tango, alternated by a tanda of three vals, then there are two tandas of tango again, followed by a tanda of three milongas. Between two tandas there is a *cortina*, a short piece of non-tango music, during which the dancers swap their partners. In what follows the unit of time is the tanda.

Three assumptions are made:

A.1 *leaders choose their partners randomly*;

A.2 *at any tanda the coupling of dancers occurs independently by previous history*;

A.3 *leaders are always active*.

Under Assumption A.1 the probability π of a follower to be invited to dance is constant. Let N_L and N_F be the number of leaders and the number of followers in the milonga, respectively. Denote by M_i the event 'a follower is invited by the leader i', for $i = 1, \ldots, N_L$. We have

$$\begin{aligned}\pi &= P(\textit{A follower is invited to dance}) = P\left(M_1 \cup M_2 \cup \cdots \cup M_{N_L}\right)\\ &= P\left(M_1\right) + P\left(M_2\right) + \cdots + P\left(M_{N_L}\right) = \frac{1}{N_F} + \frac{1}{N_F} + \cdots + \frac{1}{N_F} = \frac{N_L}{N_F}.\end{aligned} \tag{14.2}$$

Under Assumption A.2 the coupling of dancers at the current tanda is independent from previous ones. Hence any tanda can be regarded as an independent trial, whose outcome is a success if a given follower gets to dance and a failure otherwise. In this circumstances, the waiting time T for a follower to be invited has a geometric distribution, i.e.

$$T \sim G(\pi).$$

Hence the expected time for the first tanda is

$$E\left[T\right] = 1/\pi = N_F/N_L.$$

The larger the excess of followers with respect to leaders is, the longer the follower has to wait.

Furthermore if n_j^T is the time (in terms of tandas) the follower j spends in milonga, the number of tandas X_j^T he/she dances has a binomial distribution

$$X_j^T \sim B(n_j^T, \pi).$$

Consequently the expected number of tandas the follower dances is

$$E\left[X_j^T\right] = n_j^T \pi = n_j^T \frac{N_L}{N_F},$$

and the expected idling time is $n_j^T(1 - \pi)$.

Under Assumptions A.1–A.3 modeling the waiting time and the activity of the dancers is pretty straightforward.

Suppose now that Assumption A.1 does not hold. In fact more skilled followers are more likely to be invited. The choice of the partner may be affected by the kind of music, often some dancers are chosen for tango and others for vals or milongas. Friendship between the dancers may also encourage invitation.

Furthermore location of the dancers can have a significant impact on the choice of the partner. Traditionally the invitation to dance is proposed and accepted through '*mirada*' and '*cabeceo*'. The dancer, who takes the initiative, stares or smiles at his/her target partner, performing a 'mirada'. If the other dancer reciprocates with a smile or by nodding his/her head then the invitation is accepted and the two dancers get together for a tanda. Mirada and cabeceo are conceived to ensure the woman's right to choose her partner. They allow both parties to avoid unpleasant publicly evident rejections. Clearly it is more difficult to address a mirada to someone sitting far away, hence the invitation is more easily directed to dancers close by.

When Assumption A.1 is not fulfilled, there is a probability π_j of being invited for the follower j, which differs from the probability for the other followers. Equation (14.2) is replaced by

$$\begin{aligned}\pi_j &= P\,(Follower\ j\ is\ invited\ to\ dance) = P_j\left(M_1 \cup M_2 \cup \cdots \cup M_{N_L}\right)\\ &= P_j\,(M_1) + P_j\,(M_2) + \cdots + P_j\left(M_{N_L}\right)\end{aligned}$$

where $P_j\,(M_i)$ is the probability of follower j being invited by leader i. Without Assumption A.1 the probability of being invited depends on the personal features of the follower and there is no any easy formula to compute it. However, if Assumption A.2 still holds, any tanda is an independent trial for follower j whose outcome is either getting to dance or being idling.

The waiting time T_j of the follower j is still geometrically distributed with parameter π_j, $T_j \sim G(\pi_j)$, though π_j varies across dancers. Furthermore the number of tandas the follower j dances still has a binomial distribution with parameters n_j^T (i.e. the time (in tandas) the follower j spends in milonga) and π_j, $X_j^T \sim B\left(n_j^T, \pi_j\right)$.

Complications arise when Assumption A.2 is removed. On one side dancers are expected to rotate their partners, so that a couple is unlikely to dance more than a tanda in a row. On the the other side couples, who have experienced a good connection in previous tandas, may be willing to replicate. At this point there is a probability $\pi_{ji}^t\,(\Omega_{t-1})$ for the follower j to be invited by the leader i at time t which is a function of the history of all the previous tandas Ω_{t-1}.

Also Assumption A.3 may not be fulfilled. Apart from physical resistance which can affect the activity of the leaders, many dancers may enjoy dancing one kind of music among tango, vals and milonga more than the others, or they may have specific preference for songs from different musicians. Hence dancers with both roles may decide to skip one or more tandas if they don't enjoy the kind of music or they don't like the song.

In these circumstances the probability $\pi_{ij}^{t}\,(\Omega_{t-1})$ becomes a complicated function of various factors. On both the follower and the leader side, there are their technical skills as well as their personal preferences on the kind of music and on the songs. On the couple side there are the time elapsed from previous joint tandas (if any), the location of the two dancers, their connection in dancing and their friendship.

Appendix: Notes on Tango Music

Tango has a timbre that is extremely recognizable even by those who are casual listeners. The bands in tango are called *Tango Orquestra* and may include from four to few dozens of musicians. The key instrument is the bandoneon, a sort of small accordion. The bandoneon is the core instrument which characterizes the timbre of a Tango Orquestra. Other essential instruments are the fiddles, the upright bass and the piano. The sound of Tango Orquestras is characterized by many harmonic instruments and by the absence of drums, reeds and brasses, which are typical in jazz bands and latin bands born in the same historical period in other parts of America.

Since tango has popular roots the shape of the songs is very similar to the popular songs. The tango repertoire is extremely wide and it is impossible to outline a scheme which embeds every song, hence we focus on the most recurrent form. The typical form in tango is a two alternated sections song (A–B–A–B) where every section is formed by 16 or 32 bars. Sometimes these sections are repeated twice. The two sections are clearly distinguished since one is joyful (major harmonies) and the other one dramatic (minor harmonies). The choice of the position is left to the composer.

As mentioned in Sect. 14.2, there are three kinds of music in tango which differ for the time signature:

– milonga has a $\frac{2}{4}$ time signature;
– vals has a $\frac{3}{4}$ time signature;
– tango has a $\frac{4}{4}$ time signature.

The milonga has a rhythmic pattern that is called habanera. In Fig. 14.3 there is the transcription of the first four bars of the bassline from the cinematic version of ‘*Se dice de mi*’ by Francisco Canaro (music) and Ivo Pelay (lyrics), sung by Tita Merello, from the movie ‘Mercado De Abasto’ (Lucas Demare, 1955).

Figure 14.4 shows the first four bars of the vals ‘*Desde el alma*’ by Rosita Melo (1911) for a piano solo arrangement. The left hand (which interpret the notes on the second line) plays a rhythmic pattern very close to the classic European waltz. The

Fig. 14.3 First four bars of ‘*Se dice de mi*’

Fig. 14.4 First four bars of '*Desde el alma*'

Fig. 14.5 Generic tango four quarters rhythmic comping in G major

first quarter has an accent while the other two are played piano and this gives rise to a rotatory feeling.

The (properly said) tango is slightly different from the other two genres. The Habanera is no longer played, and the rhythmic section plays mainly the four quarters as shown in Fig. 14.5. This simple groove is the main feature of the musical genre which makes tango very recognizable.

Since this comping is very basic a lot of composers, arrangers and orchestra directors manipulate this pattern to make it more suitable to the melody of the song. These notes describe just the basic structure, on the basis of which any musician and arranger develops his personal style.

References

1. ALSALEH M. A. (2019). Benefits of Argentine tango in diseases and disorders. *Psychology and Behavioral Science International Journal* **11** 555807.
2. AZZI M. S. (1996). Multicultural tango: the impact and the contribution of the Italian immigration to the Tango in Argentina. *International Journal of Musicology* **5** 437–453.
3. CARA A. C. (2009). Entangled tangos: passionate displays, intimate dialogues. *The Journal of American Folklore* **122** 438–465.
4. COLLEBRUSCO L., TOCCO G. D., BELLANTI A. (2018). Tango-Therapy: an alternative approach in early rehabilitation of multiple sclerosis. *Open Journal of Therapy and Rehabilitation* **6** 85–93.
5. COLLIER S. (1992). The popular roots of the Argentine tango. *History Workshop Journal* **34** 92–100.
6. GRITZNER K. (2017). Between commodification and emancipation: the tango encounter. *Dance Research: The Journal of the Society for Dance Research* **35** 49–60.
7. GOERTZEN C., AZZI M. S. (1999). Globalization and the tango. *Yearbook for Traditional Music* **31** 67–76.
8. HACKNEY M. E., KANTOROVICH S., LEVIN R., EARHART G.M. (2007). Effects of tango on functional mobility in Parkinson's disease: a preliminary study. *J Neurological Physical Therapy* **31** 173–179.
9. KIMMEL M. (2012). Intersubjectivity at close quarters: how dancers of tango Argentino use imagery for interaction and improvisation. *Cognitive Semiotics* **4** 76–124.
10. KOH Y., NOH G. (2020). Tango therapy for Parkinson's disease: effects of rush elemental tango therapy. *Clinical Case Report* **8** 970–977.

11. Lötzke D., Ostermann T., Büssing A. (2015). *Argentine tango in Parkinson disease–a systematic review and meta-analysis*. *BMC Neurology* **15** 226.
12. Megenney W. W. (2003). The river plate “tango”: etymology and origins. *Afro-Hispanic Review* **22** 39–45.
13. Oei S.L., Rieser T., Becker S., Groß J., Matthes H., Schad F., Thronicke A. (2021). TANGO: effect of tango Argentino on cancer-associated fatigue in breast cancer patients-study protocol for a randomized controlled trial. *Trials* **22** 866.
14. Pinniger R., Brown R. F., Thorsteinsson E. B., McKinley P. (2012). Argentine tango dance compared to mindfulness meditation and a waiting-list control: a randomised trial for treating depression. *Complementary Therapies in Medicine* **20** 377–384.
15. Rabinovich, D. B., Garretto, N. S., Arakaki, T., DeSouza J. FX (2021). A high dose tango intervention for people with Parkinson’s disease (PwPD), *Advances in Integrative Medicine* **8** 272–277.
16. Rahmatian S. (2018). The structure of Argentine tango. *Journal of Mathematics and the Arts* **12** 225–243.
17. Stepputat, K. (2020). Tango musicality and tango danceability: reconnecting strategies in current cosmopolitan tango Argentino practice. *The World of Music* **9** 51–68.
18. Stepputat K., Kienreich W., Dick C. S. (2019). Digital methods in intangible cultural heritage research: a case study in tango Argentino. *Journal on Computing and Cultural Heritage* **12** 1–22.

Chapter 15
Z-Process Method for Change Point Problems in Time Series

Ilia Negri

Abstract Z-process method was introduced as a general unified approach based on partial estimation functions to construct a statistical test in change point problems not only for ergodic models but also for some non-ergodic models where the Fisher information matrix is random. In this paper, we consider the problem of testing for parameter changes in time series models based on this Z-process method. As an example, we consider the parameter change problem in some linear time series models. Some possibilities for nonlinear models are also discussed.

15.1 Introduction

This paper overviews some results for change points, or structural breaks, in time series analysis. We focus on models described in time domain frameworks such as parametric linear and nonlinear models, where the dependence structure is explicitly described according to potential breaks in the observation. For the particular choice of the models, one can consider score-based test statistics, using the likelihood principle. Here, our aim is to show that "Z-method" introduced in [18] for diffusion processes, can be seen as a possible alternative method to test for parameter change for linear, and possible nonlinear time series models. The problem for an i.i.d. sample was first considered in [19, 20]. Let $X_1, \dots X_n$ be a sample of size n of independent observations admitting density $f_i = f(\cdot, \theta_i)$, $i = 1, \dots, n$, respectively. The change point problems is to test: $H_0 : \theta_1 = \theta_2 = \cdots = \theta_n$ against $H_1 : \theta_1 = \cdots = \theta_k \neq \theta_{k+1} = \cdots = \theta_n$, where $1 \leq k < n$ is a change point. The paper [9] gave a complete review on change point detection and estimation for i.i.d. model and linear (multiple) regression models. See also [8] for parametric methods and analysis. For regression models, see for example, [6, 7, 12, 13, 21]. We refer

I. Negri (✉)
University of Calabria, Via Pietro Bucci, 87036 Rende, CS, Italy
e-mail: ilia.negri@unical.it

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_15

the readers to [5] for a complete review of methods to identify change points for sequential random sequences. The most frequent methods used are likelihood-ratio procedure, information criterion methods (AIC and SIC), Bayes solution, cumulative sum (CUSUM) method, and wavelet transformation method. The paper [14] was the first to introduce a test statistic based on the Fisher-score process to test a parameter change for independent data from which the idea of the "Z-methods" arise. Let us give an illustration of the method from an example of independent data. Let $(\mathcal{X}, \mathcal{A}, \mu)$ be a measure space and $f(\cdot;\theta)$ a parametric family of probability densities with respect to μ, where $\theta \in \Theta \subset \mathbb{R}^d$. Let $X_1, X_2, \ldots$ be an independent sequence of $\mathcal{X}$-valued random variables from this parametric model. Given $\mathbb{M}_n(\theta) = \sum_{k=1}^n \log f(X_k;\theta)$, there are at least two ways to define the maximum likelihood estimator (MLE) in statistics. The first, that is, a special case of M-estimators, is $\hat{\theta}_n = \underset{\theta\in\Theta}{\operatorname{argmax}}\, \mathbb{M}_n(\theta)$. The second is to define it as the solution to the estimating equation $\mathbb{Z}_n(\hat{\theta}_n) = 0$, where $\mathbb{Z}_n(\theta) = \dot{\mathbb{M}}_n(\theta) = \left(\frac{\partial}{\partial\theta_1}\mathbb{M}_n(\theta), \ldots, \frac{\partial}{\partial\theta_d}\mathbb{M}_n(\theta)\right)^T$. The latter is a special case of Z-estimators. See [24, 25] for these details. It is well known that the MLE $\hat{\theta}_n$ is asymptotically normal, and that, for any bounded continuous function $\psi : \mathbb{R}^d \to \mathbb{R}$,

$$\lim_{n\to\infty} E[\psi(\sqrt{n}(\widehat{\theta}_n - \theta_0))] = E[\psi(I(\theta_0)^{-1/2}\zeta)],$$

where $I(\theta_0)$ is the Fisher information matrix and ζ is a standard normal random vector.

The natural space in change point problem is $D[0, 1]$, the space of functions defined on $[0, 1]$ which are right-continuous and have left-hand limits, taking values in a finite-dimensional Euclidean space. We equip this space with the Skorohod metric, see [4, Chapter 2] for the properties of this space. Throughout this paper, all random processes are assumed to take values in $D[0, 1]$. Moreover, to consider the change point problem, we set $u \in [0, 1]$ and the partial sum process is defined as $\mathbb{M}_n(u,\theta) = \sum_{k=1}^{\lfloor un\rfloor} \log f(X_k;\theta)$, $\forall u \in [0, 1]$. Its gradient vector is denoted by $\mathbb{Z}_n(u,\theta) = \dot{\mathbb{M}}_n(u,\theta)$, $\forall u \in [0, 1]$.

The test problem can be formulated as follows:

H_0: the true value $\theta_0 \in \Theta$ does not change during $u \in [0, 1]$,
H_1: there is a change in some $u \in (0, 1)$.

Under the alternative, there exists a certain $u = u^*$, where the value of the parameter changes. It means that, letting $k = \lfloor u^* n\rfloor$, the observations $X_1, \ldots, X_k$ have density $f(\cdot;\theta_0)$, while $X_{k+1}, \ldots, X_k$ have density $f(\cdot;\theta_1)$, $\theta_0 \neq \theta_1$. Let $\widehat{\theta}_n$ be the MLE for the full data $X_1, \ldots, X_n$, that is, $\widehat{\theta}_n$ is the solution to the estimating equation $\mathbb{Z}_n(1,\theta) = \dot{\mathbb{M}}_n(1,\theta) = 0$.

It is well known from Donsker's theorem that under H_0, $n^{-1/2} I(\theta_0)^{-1/2}\mathbb{Z}_n(u,\theta_0)$ converges weakly to $B(u)$ in the Skorohod space $D[0, 1]$, where $B(u)$ is a vector of independent standard Brownian motions, and $I(\theta_0)$ denotes the Fisher Information matrix. Moreover, the score random process $n^{-1/2}\hat{I}_n^{-1/2}\mathbb{Z}_n(u,\widehat{\theta}_n)$ converges weakly to $B^\circ(u)$ in $D[0, 1]$, where $B^\circ(u)$ is a vector of independent standard Brownian bridges and $\hat{I}_n$ is a consistent estimator for $I(\theta_0)$. The test statistic considered in [14] is $\mathcal{T}_n = n^{-1} \sup_{u\in[0,1]} \mathbb{Z}_n(u, \widehat{\theta}_n)^\top \widehat{I}_n^{-1} \mathbb{Z}_n(u, \widehat{\theta}_n)$, where $\widehat{I}_n$ is a consistent estimator for the Fisher Information matrix $I(\theta_0)$. It was shown that $\mathcal{T}_n$ converges in distribution to $\sup_{u\in[0,1]} ||B^\circ(u)||^2$. See [14] for this asymptotic result. The asymptotic behavior of the test under the alternative was not discussed in [14]. The paper [17] applied the same approach based on the Fisher-score process to the change point problem for an ergodic diffusion process model based on continuous observations. They studied the asymptotic behavior of the test statistics under the alternative, and the consistency of the test under an alternative which has sufficient generality was also proved. In [18], a generalized version of Horváth and Parzen's theory was proposed. The proposed method is not just a simple generalization of the Fisher-score process method proposed by [14] for the case of independent random sequences, at least for three reasons. It makes possible to treat new applications in a broad spectrum of statistical change point problems including not only models for ergodic-dependent data but also non-ergodic cases. The proofs of the main results are based on some asymptotic representations of Z-estimators that are new from the viewpoint of mathematical statistics. Moreover, an argument to prove the consistency of the test based on the proposed method under some specified alternatives is developed.

The rest of the paper is organized as follows. In Sect. 15.2, we summarize the general Z-process method for change point problems. In Sect. 15.3, we review the change point problem for time series in light of the Z-process method. Finally, in Sect. 15.4, we give some ideas on how to apply it to nonlinear time series models.

15.2 *Z*-process Method for Change Point Problems

In this section, the main results of Z-process method for change point problems are presented. See [18] for the details. The notations $\to^p$ and $\to^d$ mean the convergence in probability and the convergence in distribution, as $n \to \infty$, respectively.

Let Θ be a bounded, open, convex subset of $\mathbb{R}^d$. For every $n \in \mathbb{N}$, let $\mathbb{Z}_n(u, \theta)$ be an $\mathbb{R}^d$-valued random process indexed by $\theta \in \Theta$, defined on a probability space $(\Omega, \mathcal{F}, P)$ that is common for all $n \in \mathbb{N}$. Let denote by $\dot{\mathbb{Z}}_n(u, \theta)$ the $d \times d$ random matrix whose (i, j)-component is $\dot{\mathbb{Z}}_n^{i,j}(u, \theta) = \frac{\partial}{\partial\theta_j}\mathbb{Z}_n^i(u, \theta)$. The case where $\mathbb{Z}_n(u, \theta)$ is given as the gradient vector $\dot{\mathbb{M}}_n(u, \theta)$ of an $\mathbb{R}$-valued random field $\mathbb{M}_n(u, \theta)$, which is assumed to be two times continuously differentiable with the Hessian matrix $\ddot{\mathbb{M}}_n(u, \theta)$, is included in this framework. The following testing problem is considered: H_0: the true value $\theta_0 \in \Theta$ does not change during $u \in [0, 1]$, against the alternative H_1: "there is a change of the parameters for some u". Under H_0, let us consider the following conditions.

Assumption 15.1

(i) There exists a sequence of diagonal matrices Q_n and a limit (could be also random) $Z_{\theta_0}(1, \theta)$ such that

$$\sup_{\theta\in\Theta} ||Q_n^{-2}\mathbb{Z}_n(1,\theta) - Z_{\theta_0}(1,\theta)|| \to^p 0.$$

(ii) The limit $Z_{\theta_0}(1,\theta)$ satisfies $\inf_{\theta:||\theta-\theta_0||>\varepsilon} ||Z_{\theta_0}(1,\theta)|| > 0, \quad \forall\varepsilon > 0$, and $||Z_{\theta_0}(1,\theta_0)|| = 0$, almost surely.

(iii) There exist a sequence of diagonal matrices R_n and a sequence of matrix-valued random processes $V_n(u,\theta_0)$ such that $V_n(1,\theta_0)$ is non-singular almost surely and that for any sequence of Θ-valued random vectors $\widetilde{\theta}_n(u)$ indexed by $u \in [0, 1]$ satisfying $\sup_{u\in[0,1]} ||\widetilde{\theta}_n(u) - \theta_0|| \to^p 0$,

$$\sup_{u\in[0,1]} ||Q_n^{-1}\dot{\mathbb{Z}}_n(u,\widetilde{\theta}_n(u))R_n^{-1} - (-V_n(u,\theta_0))|| \to^p 0.$$

(iv) In $D[0, 1]$,

$$(Q_n^{-1}\mathbb{Z}_n(u,\theta_0), V_n(u,\theta_0)) \to^d ((u^{-1}V(u,\theta_0))^{1/2}B(u), V(u,\theta_0)),$$

where $B(u)$ is a d-dimensional standard Brownian motion, and the value of $u^{-1}V(u,\theta_0))^{1/2}B(u)$, at $u = 0$ should be read as zero.

(v) There exists a sequence of matrix-valued random processes, $\widehat{V}_n(u)$, such that $\sup_{u\in[0,1]} ||\widehat{V}_n(u) - V(u,\theta_0)|| \to^p 0$.

Remark 15.1 In the examples presented in [18], the rate matrices Q_n and R_n are diagonal matrices. In some cases, these matrices have a form like $\sqrt{n}I_d$, where I_d denotes the $d \times d$ identity matrix, but in general, they need not be diagonal.

Remark 15.2 In one of the examples presented in [18], the limit $Z_{\theta_0}(1,\theta)$ is random.

Let $\widehat{\theta}_n$ be any sequence of Θ-valued random vectors such that $||Q_n^{-1}\mathbb{Z}_n(1,\widehat{\theta}_n)|| = o_P(1)$ under H_0. Let $V(u,\theta_0)$ be a non-negative definite matrix-valued random process such that $V(1,\theta_0)$ is positive definite almost surely, and let $B(u)$ be a vector of independent standard Brownian motions; we assume that $V(u,\theta_0)$ and $B(u)$ are independent. Let $\widehat{V}_n(u)$ be any sequence of matrix-valued random processes, which are non-singular except for $u = 0$ almost surely, and it should be a uniformly consistent sequence of estimators for the non-negative definite matrix-valued random process $V(u,\theta_0)$. Introduce the test statistic

$$\mathcal{T}_n = \sup_{u\in(0,1]} (Q_n^{-1}\mathbb{Z}_n(u,\widehat{\theta}_n))^\top (u\widehat{V}_n(u)^{-1})Q_n^{-1}\mathbb{Z}_n(u,\widehat{\theta}_n). \qquad (15.1)$$

We state the result about the asymptotic behavior of the test statistic under the null hypothesis; for the proof, refer to [18].

Theorem 15.1 *Suppose that Assumption 15.1 holds. Then,*

$$\mathcal{T}_n \to^d \sup_{u\in[0,1]} ||B(u) - u^{1/2}V(u,\theta_0)^{1/2}V(1,\theta_0)^{-1/2}B(1)||^2. \qquad (15.2)$$

Some remarks are necessary about the limit appearing in (15.2). If $V(u, \theta_0) = uV(1, \theta_0)$ for every $u \in [0, 1]$, then the test is asymptotically distribution free. In this case, the limit reduces to $\sup_{u\in[0,1]} ||B^\circ(u)||^2$, where $B^\circ(u) = B(u) - uB(1)$ is a vector of independent standard Brownian bridges. The paper [15] gave a table of the critical values of $\sup_{u\in[0,1]} ||B^\circ(u)||^2$ for different significance levels and for different values of the dimension d, see also [22]. Moreover, due to the independence of $V(u, \theta_0)$ and $B(u)$, the limit in (15.2) is approximated by

$$\sup_{u\in[0,1]} ||B(u) - u^{1/2}\widehat{V}_n(u)^{1/2}\widehat{V}_n(1)^{-1/2}B(1)||^2.$$

The approximate distribution of the limit can be computed by some computer simulations for the standard Brownian motions $B(u)$. In [18], this method was successfully applied to a change point problem for ergodic diffusion process and to a change point problem for volatility of a diffusion process. In particular, an interesting point of this last example is that the limit $V(u, \theta_0)$ of the normalized $\dot{\mathbb{Z}}_n(u, \widetilde{\theta}_n(u))$ is random and depend on $u \in [0, 1]$ in a complex way. In [18], some results on consistency were also shown.

15.3 Change Point Problems for Time Series

Although the change point analysis arose for independent observations, it has become an important area of research for dependent observations in particular in time series models. The paper [10] gave a general survey of parametric and non-parametric methods for dependent observations. The paper [16], using some limiting theorems for the Wald test statistics, developed a general asymptotic theory for change point problems in a general class of time series models including ARIMA models, under the null hypothesis. The review paper [1] gave an account of some work on structural breaks in time series models. A particular attention is devoted to the applications of the CUSUM test and likelihood-ratio test to time series models including AR, GARCH, and fractionally integrated ARMA models. They also briefly discussed several approaches for estimating and locating multiple break points in the observations. The paper [2] introduced a procedure based on quasi-likelihood scores to detect changes in the parameters of a GARCH model. See also [3] where a test for a change in the parameters of a GARCH model based on approximate likelihood scores that does not require the observations to have finite variance was proposed.

As the general Z-process method for change point problems presented in the previous section is based on a generalization of the score statistics, we review in details, using the involved Z process method, some works where the score-based test is applied to change point problems for time series models. The paper [11] considered change point problems for AR models. The test statistics are based on the efficient score vector and the large sample properties of the change point estimator are also explored. Let us consider the AR model $X_k = a(X_{k-1}, \ldots, X_{k-p}) + \varepsilon_k$,

$k = 1, 2 \ldots$, where $a(x_1, \ldots x_p) = \sum_{i=1}^{p} \phi_i x_i$ is an $\mathbb{R}^p$ to $\mathbb{R}$ linear function, satisfy the stationary conditions for the AR process $\{X_k\}$, and $\{\varepsilon_k\}$ is an i.i.d. sequence of $N(0, \sigma^2)$. Under the Gaussian assumption, the log-likelihood function based on observations $X_1, \ldots, X_n$ may be easily derived. Let us denote it by $\ell_n(\theta)$, where θ is the vector of unknown parameters ϕ_i, $i = 1, \ldots, p$ and σ^2. Let us denote by $I(\theta)$ the information matrix for this AR model. So, for $0 \le u \le 1$, $\mathbb{M}_n(u, \theta) = \ell_{\lfloor un \rfloor}(\theta)$ and its gradient vectors is $\mathbb{Z}_n(u, \theta) = \dot{\ell}_{\lfloor un \rfloor}(\theta)$. Let us denote by $\hat{\theta}_n$ the MLE of the parameter θ, that is the solution of the equations $\mathbb{Z}_n(1, \theta) = 0$. The paper [11] considered the efficient score statistic $\mathcal{S}(u) = n^{-1/2} I^{-1/2}(\hat{\theta}_n) \mathbb{Z}_n(u, \hat{\theta}_n)$, and proved that it converges weakly in $D[0, 1]$ to a vector of independent standard Brownian bridges $B^\circ(u)$. In a similar way, the test statistics given by (15.1) can be defined in this framework and the result of Theorem 15.1 holds under the same assumption as [11].

A limited number of studies have been presented for parameter change in ARMA-GARCH models, except that [23] compared the score test and residual-based CUSUM test and derived their limiting null distributions. Our interest is in the score test. The quasi-maximum likelihood estimator for the parameters of an ARMA(P,Q)-GARCH(p,q) model are derived. Their score test statistic is defined using the quasi-likelihood $\tilde{L}_{\lfloor un \rfloor}(\theta)$ for the ARMA-GARCH process. It can be rewritten as a Z-process. Indeed, we can set $\mathbb{M}_n(u, \theta) = \lfloor un \rfloor \tilde{L}_{\lfloor un \rfloor}(\theta)$. The paper [23] proved essentially that $n^{-1/2} I^{-1/2}(\hat{\theta}_n) \mathbb{Z}_n(u, \hat{\theta}_n)$, where $\mathbb{Z}_n(u, \theta) = \dot{\mathbb{M}}_n(u, \theta)$ and $I(\theta)$ is the Fisher information matrix for the model considered, converges weakly in $D[0, 1]$ to a vector of independent standard Brownian bridges $B^\circ(u)$. The test statistics given by (15.1) can be also defined for ARMA-GARCH process and the result of Theorem 15.1 holds under the same assumption as [23].

15.4 Nonlinear Time Series Models

The general Z process method may be applied to the time series models of the form

$$X_k = a(X_{k-1}, X_{k-2}, \ldots X_{k-p}; \theta) + b(X_{k-1}, X_{k-2}, \ldots X_{k-q}; \theta)\varepsilon_k, \quad k = 1, 2, \ldots. \tag{15.3}$$

Here, $\{\varepsilon_k\}$ is an i.i.d. sequence with $E[\varepsilon_1] = 0$ or, more generally, a martingale difference sequence with respect to the filtration $(\mathcal{F}_k)_{k \ge 0}$, where $\mathcal{F}_k = \sigma(X_k, X_{k-1}, \ldots)$. The functions $a : \mathbb{R}^p \times \Theta \to \mathbb{R}$ and $b : \mathbb{R}^q \times \Theta \to \mathbb{R}$ are known functions of $(x, \theta) \in \mathbb{R}^p \times \Theta$, and $(x, \theta) \in \mathbb{R}^q \times \Theta$, respectively, where Θ is a subset of $\mathbb{R}^d$. Let us assume that a and b are twice continuously differentiable with respect to θ. A possible way to define the estimating functions is that

$$\mathbb{Z}_n(1, \theta) = \dot{\mathbb{M}}_n(1, \theta) \quad \text{and} \quad \mathbb{Z}_n(u, \theta) = \dot{\mathbb{M}}_n(u, \theta),$$

where

$$\mathbb{M}_n(u,\theta) = -\sum_{k=1}^{[un]} \left\{ \log b(X_{k-1}, X_{k-2}, \ldots; \theta) + \frac{|X_k - a(X_{k-1}, X_{k-2}, \ldots; \theta)|^2}{2b(X_{k-1}, X_{k-2}, \ldots; \theta)^2} \right\}.$$

The rate matrix is typically given by $Q_n = R_n = \sqrt{n} I_d$. In [26], different nonlinear time series models were considered including the exponential AR model (ExpAR). See also [27]. The theory can be applied to the following ExpAR model.

ExpAR(1)-model with GARCH error.
Let us consider the function

$$a(x_1, x_2; \theta) = \left\{ \theta_1 + \theta_2 \exp(-\theta_3 x_2^2) \right\} x_1,$$

where $\theta_1 \in \mathbb{R}$, $\theta_2 \neq 0$, $\theta_3 > 0$, and $b(x; \theta) = 1$. Then the model in (15.3) is

$$X_k = \left\{ \theta_1 + \theta_2 \exp(-\theta_3 X_{k-2}^2) \right\} X_{k-1} + \varepsilon_k, \quad k = 1, 2, \ldots.$$

Here, $\theta = (\theta_1, \theta_2, \theta_3)^T$ and $\Theta \in \mathbb{R}^3$. The Z process method can be applied to test if there is a change point in the parameter θ. The limit process can be derived starting from the computation of the Fisher information matrix. Under some reasonable settings, Assumption 15.1 can be verified and the test statistic (15.1) can be computed to test for a chance in one or more components of the parameter θ.

References

1. Aue, A., Horváth, L. (2013). Structural breaks in time series. *J*ournal of Time Series Analysis **34** 1–16
2. Berkes, I., Gombay, E., Horváth, L., Kokoszaka, P. (2004). Sequential change-point detection in GARCH(p, q) models. *E*conometric Theory **20** 1140–1167.
3. Berkes, I., Horváth, L., Kokoszaka, P. (2004). Testing for parameter constancy in GARCH(p, q) models. *S*tatist. Probab. Lett. **4** 263–273.
4. Billingsley, P. (1968). *C*onvergence of Probability Measures. Wiley, NewYork.
5. Brodsky, B. E., Darkhovsky, B. S. (2000). *N*on-parametric Statistical Diagnosis Problems and Methods. Kluwer, Dordrecht.
6. Brown, R. L., Durbin, J., Evans, J. M. (1975). Techniques for testing the constancy of regression relationships over time. With discussion. *J*ournal of the Royal Statistical Society. Series B. Methodological, **37** 149–192.
7. Chen, J. (1998). Testing for a change point in linear regression models. *C*ommunications in Statistics. Theory and Methods **27** 2481–2493.
8. Chen, J., Gupta, A. K. (2000). *P*arametric Statistical Change Point Analysis. Borkäuser, Boston.
9. Chen, J., Gupta, A. K. (2001). On change point detection and estimation. Communications in Statistics. *S*imulation and Computation **30** 665–697.
10. Csörgo, M., Horváth, L. (1997). *L*imit Theorems in Change-Point Analysis. Wiley, New York.
11. Gombay, E. (2008). Change detection in autoregressive time series. *J*ournal of Multivariate Analysis **99** 451–464.

12. Hinkley, D. V. (1969). Inference about the intersection in two-phase regression. *Biometrika* **56** 495–504.
13. Hinkley, D. V. (1970). Inference about the change-point in a sequence of random variables. *Biometrika* **57** 1–17.
14. Horváth, L., Parzen, E. (1994). Limit theorems for Fisher-score change processes. *In*: Change-point Problems, (Edited by Carlstein, E., Müller H.-G. and Siegmund, D.) IMS Lecture Notes – Monograph Series **23** 157–169.
15. Lee, S., Ha, J., Na, O., Na, S. (2003). The cusum test for parameter change in time series models. *Scand. J. Statist.* **30** 781–796.
16. Ling, S. (2007). Testing for change points in time series models and limiting theorems for NED sequences. *The Annals of Statistics* **35** 1213–1237.
17. Negri, I. and Nishiyama, Y. (2012). Asymptotically distribution free test for parameter change in a diffusion process model. *Ann. Inst. Statist. Math.* **64** 911–918
18. Negri, I. and Nishiyama, Y. (2017). Z-process method for change point problems with applications to discretely observed diffusion processes. *Stat Methods Appl* **26** 231–250.
19. Page, E. S. (1954). Continuous inspection schemes. *Biometrika* **41** 100–115.
20. Page, E. S. (1955). A test for a change in a parameter occurring at an unknown point. *Biometrika* **42** 523–527.
21. Quandt, R. E. (1960). Tests of the hypothesis that a linear regression system obeys two separate regimes. *Journal of the American Statistical Association* **55** 324–330.
22. Song, J., Lee, S. (2009). Test for parameter change in discretely observed diffusion processes. *Statist. Inference Stoch. Process.* **12** 165–183.
23. Song, J., Kang, J. (2018). Parameter change tests for ARMA-GARCH models. *Computational Statistics and Data Analysis* **121** 41–56
24. van der Vaart, A.W. (1998). *Asymptotic Statistics.* Cambridge University Press, Cambridge.
25. van der Vaart, A.W. and Wellner, J.A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics.* Springer-Verlag, New York.
26. Teräsvirta, T. (1994). Specification, estimation, and evaluation of smooth transition autoregressive models. *Journal of the American Statistical Association* **89** 208–218.
27. Tong, H. (1995). *Non-linear Time Series: A Dynamical System Approach.* Oxford statistical science series 6, Clarendon Press, Oxford.

Chapter 16
Copula Bounds for Circular Data

Hiroaki Ogata

Abstract We propose an extension of the Fréchet–Hoeffding copula bounds for circular data. The copula is a powerful tool for describing the dependency of random variables. In two dimensions, the Fréchet–Hoeffding upper (lower) bound indicates the perfect positive (negative) dependence between two random variables. However, for circular random variables, the usual concept of dependency may not be accepted because of their periodicity. In this paper, we redefine the Fréchet–Hoeffding bounds and consider modified Fréchet and Mardia families of copulas for modelling the dependency of two circular random variables. Simulation studies are also given to demonstrate the behaviour of the model.

16.1 Introduction

Circular statistics deals with observations that are represented as points on a unit circle. Directional data is a typical example, because a direction is represented as an angle from a certain zero direction. Common examples often cited are wind direction, river flow direction and the direction of migrating birds in flight. For modelling circular data, many distributions have been proposed such as von Mises distribution, cardioid distribution, wrapped normal distribution, wrapped Cauchy distribution and so on. Recently, [8] proposed the way of constructing circular densities based on the spectra of complex-valued stationary process. In this sense, the branches of circular statistics and time series analysis have a strong relationship. For a comprehensive explanation of circular statistics, we refer to [1, 3, 6], to name a few.

If the observation is a pair of circular data, usually denoted by $(\theta, \phi) \in \mathbb{T}^2 = [0, 2\pi)^2$, it is represented as a point on a torus. When we have bivariate data, we can investigate the relationship between them. Although the Pearson correlation coefficient is the representative measure of a relationship between two random variables, it does not work for circular data because of its periodic structure. Many measures

H. Ogata (✉)
Tokyo Metropolitan University, 1-1, Minami-Osawa, Hachioji, Tokyo, Japan
e-mail: hiroakiogata@tmu.ac.jp

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_16

of association for circular data have been proposed so far. The book [6, Sect. 11.2.2.] provides a summary of circular-circular correlations.

Another powerful tool for describing the dependency between two random variables is a copula. It is a function rather than a single value and provides much more information regarding the dependency. Mathematically speaking, the copula is a joint distribution function on $\mathbf{I}^2 = [0, 1]^2$ whose margins are uniform. For a circular version, [4] introduced a circular analogue of copula density, called 'circulas', assuming the existence of density function. In this paper, we consider a copula function—not a copula density—for circular data. Because we do not assume the existence of the density function, we can deal with singular copulas. We give special consideration to circular analogues of the Fréchet–Hoeffding copula bounds, which are examples of singular copulas. Due to the arbitrary nature of the origin in the circular variable, they are not uniquely specified.

The remainder of the paper is organized as follows: Sect. 16.2 provides the definition of the equivalence class of circular copula functions from the aspect of the arbitrariness of the zero direction. Section 16.3 introduces circular analogues of the Fréchet–Hoeffding copula bounds. We prove that the following (i) and (ii) are equivalent.

(i) The bivariate circular random variables have the circular Fréchet–Hoeffding upper (lower) copula bounds.
(ii) Support of the bivariate circular random variables is nondecreasing (nonincreasing) in the sense of mod 2π.

Section 16.4 gives a Monte Carlo simulation. We generate observations from the circular version of the Mardia copula family introduced in [5], and investigate their behaviour. Section 16.5 provides summary and conclusion of this paper.

16.2 Equivalence Class of Circular Copula Functions

A copula is the joint distribution function on $\mathbf{I}^2 = [0, 1]^2$ whose margins are uniform. Sklar's theorem insists any joint distribution function $H(x, y)$ is written by the copula function $C(u, v)$ and their marginal distribution function $F(x)$, $G(y)$, i.e.

$$H(x, y) = C(F(x), G(y)).$$

In the case of circular random variables, marginal and joint distribution functions depend on the choice of the zero direction. However, we should not consider the difference caused solely by the difference of the zero direction. This section gives the concept of the equivalence class of circular copula functions in the sense of the arbitrariness of the choice of the zero direction.

First, let us give the definitions of circular distribution functions, quasi-inverses of circular distribution functions and circular joint distribution functions, in the same manner as [7, Definitions 2.3.1, 2.3.6, and 2.3.2].

Definition 16.1 A *circular distribution function* is a function F with domain $\mathbb{T} = [0, 2\pi)$ such that

1. F is nondecreasing,
2. $F(0) = 0$ and $\lim_{\theta \nearrow 2\pi} F(\theta) = 1$.

Definition 16.2 Let F be a circular distribution function. Then a *quasi-inverse* of F is any function $F^{(-1)}$ with domain $\mathbf{I}$ such that

$$F^{(-1)}(u) = \inf\{\theta | F(\theta) \geq u\}.$$

Definition 16.3 A *circular joint distribution function* is a function H with domain $\mathbb{T}^2$ such that

1. $H(\theta_2, \phi_2) - H(\theta_2, \phi_1) - H(\theta_1, \phi_2) + H(\theta_1, \phi_1) \geq 0$ for any rectangle $[\theta_1, \theta_2] \times [\phi_1, \phi_2] \subset \mathbb{T}^2$,
2. $H(\theta, 0) = H(0, \phi) = 0$ and $\lim_{\theta \nearrow 2\pi, \phi \nearrow 2\pi} H(\theta, \phi) = 1$.

Now let us extend the domain of F to $\tilde{\mathbb{T}} = [0, 4\pi)$ in the following way:

$$\tilde{F}(\theta) = \begin{cases} F(\theta) & (0 \leq \theta < 2\pi) \\ F(\theta - 2\pi) + 1 & (2\pi \leq \theta < 4\pi). \end{cases}$$

If we change the zero direction to $\alpha \in \mathbb{T}$, its distribution function becomes

$$F_\alpha(\theta) = \tilde{F}(\theta + \alpha) - \tilde{F}(\alpha).$$

The choice of α is arbitrary, so we can define the equivalence class of circular distribution functions $\{F_\alpha | \alpha \in \mathbb{T}\}$.

Similarly, we extend the domain of H to $\tilde{\mathbb{T}}^2$ in the following way:

$$\begin{aligned} &\tilde{H}(\theta, \phi) \\ &= \begin{cases} H(\theta, \phi) & (0 \leq \theta < 2\pi, 0 \leq \phi < 2\pi) \\ F(\theta) + H(\theta, \phi - 2\pi) & (0 \leq \theta < 2\pi, 2\pi \leq \phi < 4\pi) \\ G(\phi) + H(\theta - 2\pi, \phi) & (2\pi \leq \theta < 4\pi, 0 \leq \phi < 2\pi) \\ 1 + F(\theta - 2\pi) + G(\phi - 2\pi) + H(\theta - 2\pi, \phi - 2\pi) & (2\pi \leq \theta < 4\pi, 2\pi \leq \phi < 4\pi). \end{cases} \end{aligned}$$

If we change the zero directions to $(\alpha, \beta) \in \mathbb{T}^2$, its joint distribution function becomes

$$H_{\alpha,\beta}(\theta, \phi) = \tilde{H}(\theta + \alpha, \phi + \beta) - \tilde{H}(\alpha, \phi + \beta) - \tilde{H}(\theta + \alpha, \beta) + \tilde{H}(\alpha, \beta).$$

The choice of (α, β) is arbitrary, so we can define the equivalence class of circular joint distribution functions $\{H_{\alpha,\beta} | (\alpha, \beta) \in \mathbb{T}^2\}$.

From Sklar's theorem, the circular copula function is defined by

$$C_{\alpha,\beta}(u, v) = H_{\alpha,\beta}(F_\alpha^{(-1)}(u), G_\beta^{(-1)}(v)). \tag{16.1}$$

The choice of zero directions (α, β) is arbitrary, so we can define the equivalence class of circular copula functions $\{C_{\alpha,\beta} | (\alpha, \beta) \in \mathbb{T}^2\}$.

16.3 Circular Fréchet–Hoeffding Copula Bounds

A copula is known to have both upper and lower bounds. That is, for every $(u, v) \in \mathbf{I}^2$,

$$W(u, v) := \max(u + v - 1, 0) \leq C(u, v) \leq \min(u, v) =: M(u, v).$$

The bounds $M(u, v)$ and $W(u, v)$ are themselves copulas and are called *Fréchet–Hoeffding upper bounds* and *Fréchet–Hoeffding lower bounds*, respectively.

Here, we introduce the concept of a nondecreasing (nonincreasing) set in $\bar{\mathbf{R}}^2 = (-\infty, \infty)^2$.

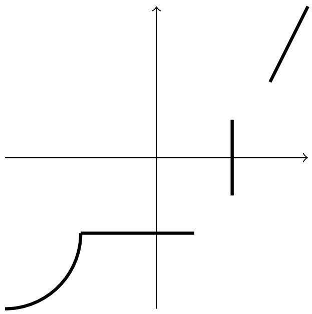
Definition 16.4 ([7, Definition 2.5.1]) A subset S of $\bar{\mathbf{R}}^2$ is *nondecreasing* if for any (x, y) and (u, v) in S, $x < u$ implies $y \leq v$. Similarly, a subset S of $\bar{\mathbf{R}}^2$ is *nonincreasing* if for any (x, y) and (u, v) is S, $x < u$ implies $y \geq v$.

The example of a nondecreasing set is given in Fig. 16.1.

If the pair of random variables (X, Y) have a Fréchet–Hoeffding upper (lower) bound, then the support of (X, Y) is nondecreasing (nonincreasing). See [7, Theorems 2.5.4 and 2.5.5]. In this sense, a Fréchet–Hoeffding upper (lower) bound indicates the perfect positive (negative) dependence between X and Y.

Now let us consider the pair of circular (angular) random variables $(\Theta, \Phi) \in \mathbb{T}^2$. When (Θ, Φ) has perfect positive dependence, its support should be like the examples shown in the top row of Fig. 16.2. The figures in the left column are for a continuous circular random variable and those in the right column are for a discrete variable. The figures in the middle and bottom rows are redrawings of those in the top row when we regard $(5\pi/4, \pi/8)$ and $(7\pi/4, 7\pi/8)$ as zero directions (indicated by cross marks in the top row figures), respectively. Due to the arbitrary nature of the zero directions, all support plots in Fig. 16.2 should indicate perfect positive dependence.

Fig. 16.1 Nondecreasing set



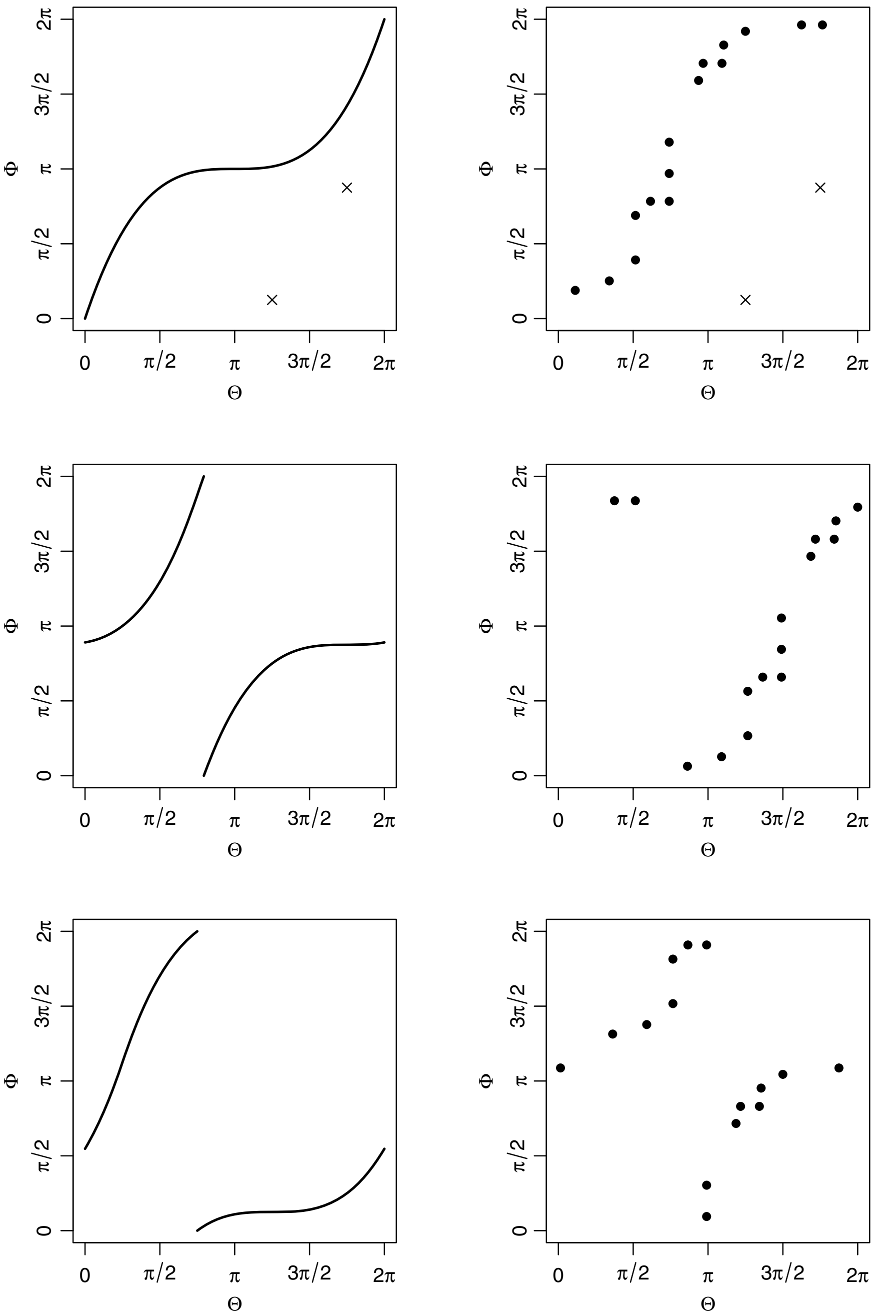


Fig. 16.2 Supports of perfect positive dependent circular random variables. The figures in the left column are for a continuous random variable and those in the right column are for a discrete random variable. The middle and bottom rows are redrawings of those in the top row when we regard $(5\pi/4, \pi/8)$ and $(7\pi/4, 7\pi/8)$ as zero directions (indicated by cross marks in the figures in the top row), respectively

Now, we give the equivalence class of the circular Fréchet–Hoeffding copula upper bounds in the following theorem; its proof is given in Sect. 16.6.

Theorem 16.1 (Circular Fréchet–Hoeffding copula upper bound) *Let* $0 \leq a \leq 1$. *The equivalence class of the circular Fréchet–Hoeffding copula upper bounds is given by*

$$M_a(u, v) = \begin{cases} \min(u, v - a) & (u, v) \in [0, 1 - a] \times [a, 1] \\ \min(u + a - 1, v) & (u, v) \in [1 - a, 1] \times [0, a] \\ \max(u + v - 1, 0) & \textit{otherwise.} \end{cases}$$

The copula $M_a(u, v)$ was introduced in [7, Exercise 3.9]. It is the joint distribution function when the probability mass is uniformly distributed on two line segments, one joining $(0, a)$ to $(1 - a, 1)$ with mass $1 - a$, and the other joining $(1 - a, 0)$ to $(1, a)$ with mass a. Theorem 16.1 clarifies that this M_a corresponds to the circular Fréchet–Hoeffding copula upper bound.

We can also find the equivalence class of the circular Fréchet–Hoeffding copula lower bounds in the following theorem.

Theorem 16.2 (Circular Fréchet–Hoeffding copula lower bound) *Let* $0 \leq a \leq 1$. *The equivalence class of the circular Fréchet–Hoeffding copula lower bounds is given by*

$$W_a(u, v) = \begin{cases} \max(u + v - a, 0) & (u, v) \in [0, a]^2 \\ \max(u + v - 1, a) & (u, v) \in [a, 1]^2 \\ \min(u, v) & \textit{otherwise.} \end{cases}$$

Proof is omitted because it is similar to that of Theorem 16.1.

The copula $W_a(u, v)$ was introduced in [7, Example 3.4]. It is the joint distribution function when the probability mass is uniformly distributed on two line segments, one joining $(0, a)$ to $(a, 0)$ with mass a, and the other joining $(a, 1)$ to $(1, a)$ with mass $1 - a$.

When describing the dependency between Θ and Φ, their periodic structure must be considered. The paper [2] considered the complete dependence between Θ and Φ as

$$\Phi \equiv \Theta + \alpha_0 \pmod{2\pi}, \text{ positive association,} \tag{16.2}$$

$$\Phi \equiv -\Theta + \alpha_0 \pmod{2\pi}, \text{ negative association,} \tag{16.3}$$

where α_0 is an arbitrary fixed direction. If we use $\mathbb{T}^2$ display, the supports of (16.2) are expressed as in Fig. 16.3. Because of arbitrary nature of α_0, not only the top but also the middle and bottom rows indicate complete positive dependence. Figure 16.3 is the special case of Fig. 16.2. Therefore, the circular perfect dependence with the concept of nondecreasing (nonincreasing) set can be considered as a generalization of complete dependence in the sense of [2].

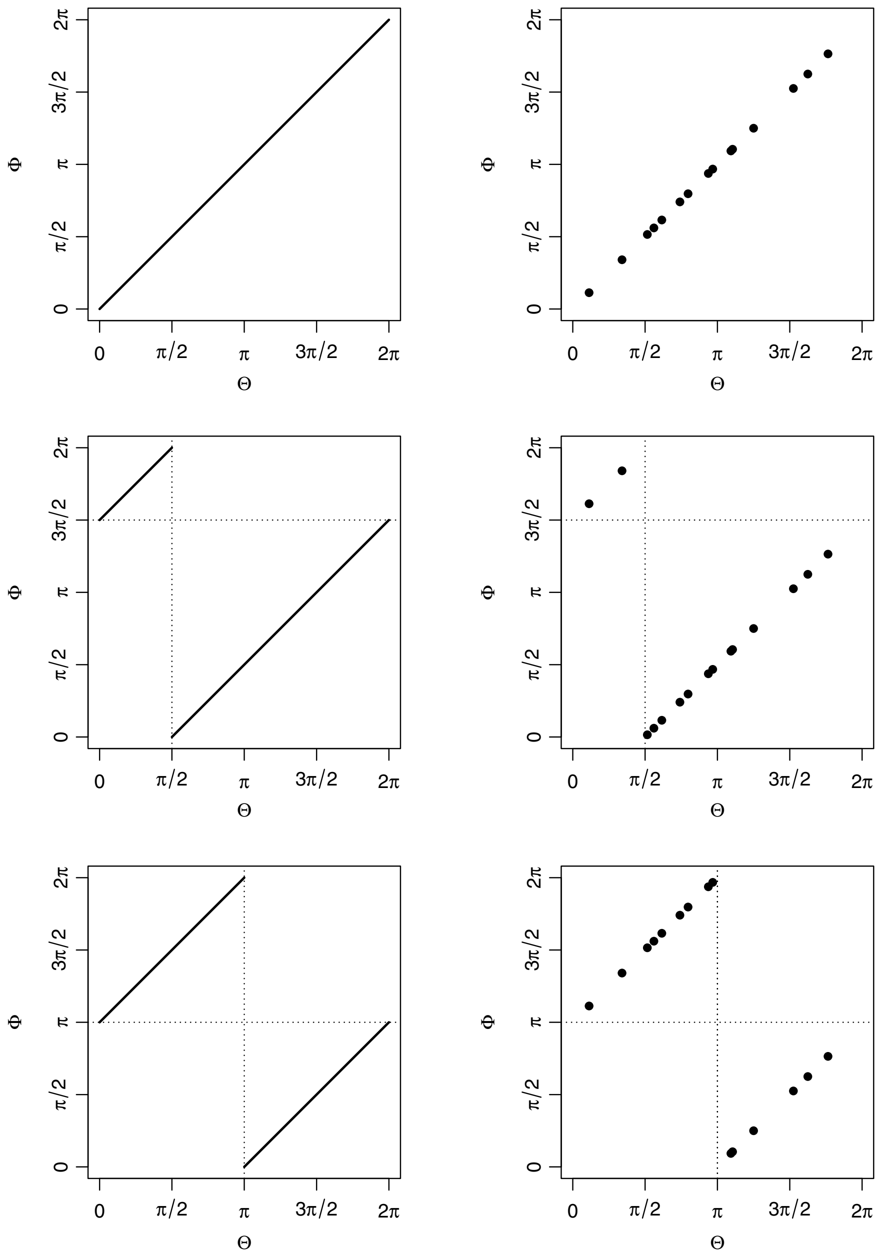


Fig. 16.3 Perfect positive dependence in the sense of [2] with $\alpha_0 = 0$ (top), $\alpha_0 = 3\pi/2$ (middle), $\alpha_0 = \pi$ (bottom). Left are for continuous and right are for discrete

16.4 Monte Carlo Simulations

In this section, we consider the copula

$$C_{\gamma,a,b}(u,v) = \frac{\gamma^2(1+\gamma)}{2} M_a(u,v) + (1-\gamma^2)\Pi(u,v) + \frac{\gamma^2(1-\gamma)}{2} W_b(u,v),$$
$$(\gamma, a, b) \in [-1,1] \times \mathbf{I} \times \mathbf{I}. \quad (16.4)$$

Here, $M_a(u,v)$, $W_b(u,v)$ are the copulas introduced in Theorems 16.1, 16.2, and $\Pi(u,v) := uv$ is an independent copula. This is a linear combination of these three copulas and an analogue to the model in [5]. The parameter γ controls the weights of these three copulas, and $\gamma = 1, 0, -1$ correspond M_a (perfect positive dependence), Π (independent), W_b (perfect negative dependence), respectively.

We generate a random circular bivariate sample $(\theta_i, \phi_i)_{i=1,\ldots,500}$ from (16.4). Both marginals are set to a cardioid distribution, whose distribution function is given by

$$F_{\mathrm{Ca}}(\theta) = \frac{\rho}{\pi}\sin(\theta-\mu) + \frac{\theta}{2\pi} + \frac{\rho}{\pi}\sin\mu.$$

The parameters for marginals are set to $\rho_F = 0.1, \mu_F = \pi$ for θ and $\rho_G = 0.3, \mu_G = \pi/3$ for ϕ. Figure 16.4 shows the simulated circular bivariate plots with $\gamma = 0.7$ (top left), -0.7 (top right), 0.5 (middle left), -0.5 (middle right), 0.3 (bottom left), -0.3 (bottom right). When γ is close to 1 and -1, (16.4) is close to M_a and W_b, which are singular copulas. When $\gamma = \pm 0.7$, we can see the singular components.

16.5 Summary and Conclusions

We have considered the equivalence class of univariate and bivariate circular distribution functions ascribed to the arbitrary nature of the zero direction. Then, we have introduced the equivalence class of circular copula functions with Sklar's theorem. Using the concept of the equivalence class of circular copula functions, we have introduced circular analogues of the Fréchet–Hoeffding copula upper and lower bounds. We have explained they are the generalizations of the complete positive and negative dependence in the sense of [2]. We also have introduced the circular analogue of Mardia's copula model and simulated a dataset from this model. When the model is close to its extremes, we can find the singular components in the plots.

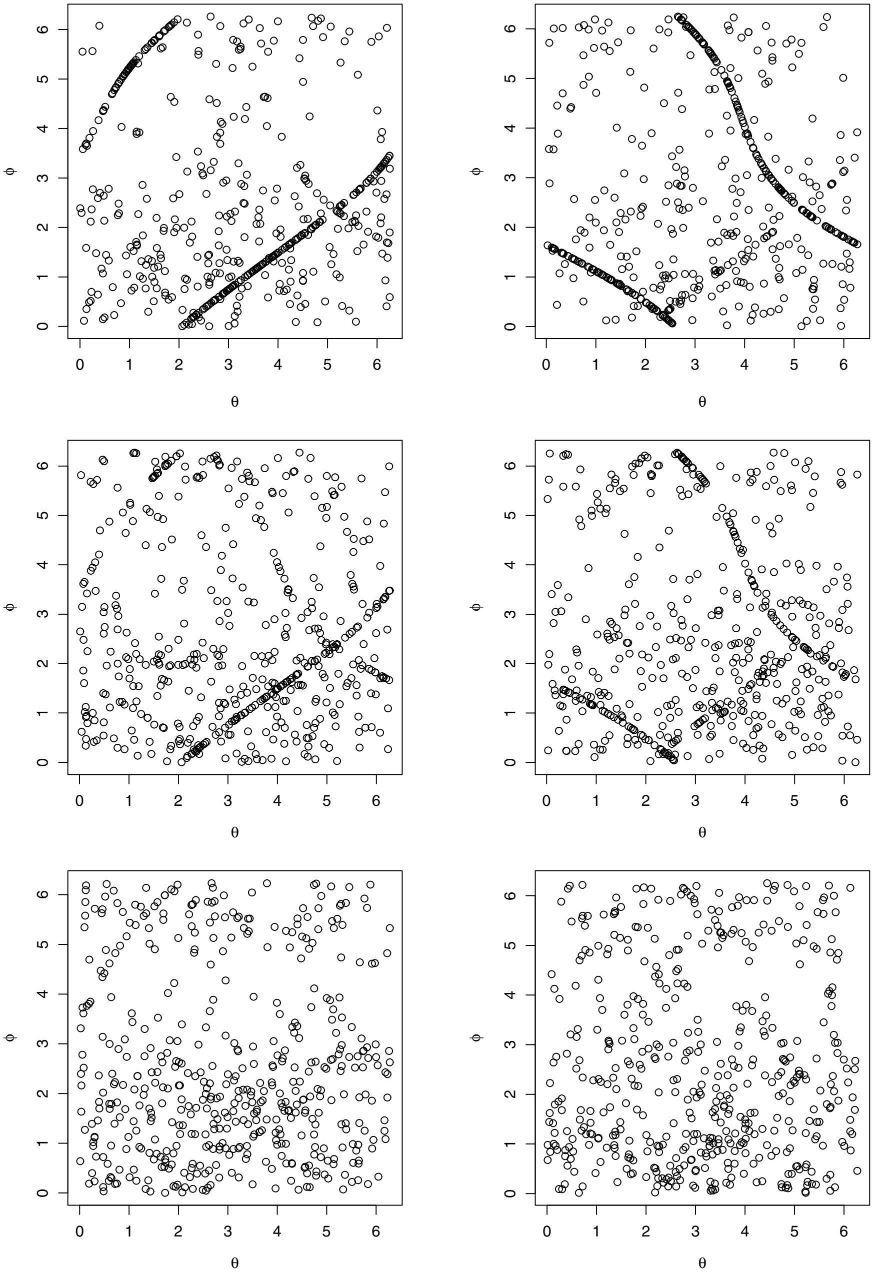


Fig. 16.4 Simulated circular bivariate plots from the model (16.4). The sample size is 500. The marginal for θ is a Cardioid distribution with $\rho_F = 0.1$, $\mu_F = \pi$, and the marginal for ϕ is a Cardioid distribution with $\rho_G = 0.3$, $\mu_G = \pi/3$. The parameters a and b are fixed to 0.7 and 0.4, respectively. The parameter γ is set to 0.7 (upper left), -0.7 (upper right), 0.5 (middle left), -0.5 (middle right), 0.3 (lower left), -0.3 (lower right)

16.6 Proof of Theorem 16.1

Let

$$H(\theta, \phi) = M(F(\theta), G(\phi)) = \min(F(\theta), G(\phi)).$$

Fix arbitrary zero directions $(\alpha, \beta) \in \mathbb{T}$. From the discussion in Sect. 16.2, the circular Fréchet–Hoeffding copula upper bound is

$$\begin{aligned} C_{\alpha,\beta}(u, v) &= H_{\alpha,\beta}(F_\alpha^{(-1)}(u), G_\beta^{(-1)}(v)) \\ &= \tilde{H}(F_\alpha^{(-1)}(u) + \alpha, G_\beta^{(-1)}(v) + \beta) \\ &\quad - \tilde{H}(\alpha, G_\beta^{(-1)}(v) + \beta) - \tilde{H}(F_\alpha^{(-1)}(u) + \alpha, \beta) + \tilde{H}(\alpha, \beta). \end{aligned}$$

When (Θ, Φ) is discrete, (u, v) is restricted on {range of F_α} $\times$ {range of G_β}. Now, let us divide $\mathbf{I}^2$ into the following four regions:

$$\begin{aligned} \mathbf{I}_1 &= \{(u, v) \mid F_\alpha^{(-1)}(u) + \alpha < 2\pi,\ G_\beta^{(-1)}(v) + \beta < 2\pi\} \\ &= \{(u, v) \mid u < 1 - F(\alpha),\ v < 1 - G(\beta)\}, \\ \mathbf{I}_2 &= \{(u, v) \mid F_\alpha^{(-1)}(u) + \alpha < 2\pi,\ G_\beta^{(-1)}(v) + \beta \geq 2\pi\} \\ &= \{(u, v) \mid u < 1 - F(\alpha),\ v \geq 1 - G(\beta)\}, \\ \mathbf{I}_3 &= \{(u, v) \mid F_\alpha^{(-1)}(u) + \alpha \geq 2\pi,\ G_\beta^{(-1)}(v) + \beta < 2\pi\} \\ &= \{(u, v) \mid u \geq 1 - F(\alpha),\ v < 1 - G(\beta)\}, \\ \mathbf{I}_4 &= \{(u, v) \mid F_\alpha^{(-1)}(u) + \alpha \geq 2\pi,\ G_\beta^{(-1)}(v) + \beta \geq 2\pi\} \\ &= \{(u, v) \mid u \geq 1 - F(\alpha),\ v \geq 1 - G(\beta)\}. \end{aligned}$$

For $(u, v) \in \mathbf{I}_1$,

$$\begin{aligned} &C_{\alpha,\beta}(u, v) \\ &= H(F_\alpha^{(-1)}(u) + \alpha, G_\beta^{(-1)}(v) + \beta) \\ &\quad - H(\alpha, G_\beta^{(-1)}(v) + \beta) - H(F_\alpha^{(-1)}(u) + \alpha, \beta) + H(\alpha, \beta) \\ &= \min\{F(F_\alpha^{(-1)}(u) + \alpha), G(G_\beta^{(-1)}(v) + \beta)\} - \min\{F(\alpha), G(G_\beta^{(-1)}(v) + \beta)\} \\ &\quad - \min\{F(F_\alpha^{(-1)}(u) + \alpha), G(\beta)\} + \min\{F(\alpha), G(\beta)\}. \end{aligned}$$

Here,

$$\begin{aligned} F(F_\alpha^{(-1)}(u) + \alpha) &= F_\alpha(F_\alpha^{(-1)}(u)) + F(\alpha) = u + F(\alpha), \\ G(G_\beta^{(-1)}(v) + \beta) &= G_\beta(G_\beta^{(-1)}(v)) + G(\beta) = v + G(\beta). \end{aligned}$$

Therefore,

$$C_{\alpha,\beta}(u,v) = \min\{u+F(\alpha), v+G(\beta)\} - \min\{F(\alpha), v+G(\beta)\} \\ - \min\{u+F(\alpha), G(\beta)\} + \min\{F(\alpha), G(\beta)\}.$$

When $G(\beta) > F(\alpha)$, we divide $\mathbf{I}_1$ into

$$\begin{aligned} \mathbf{I}^G_{1-1} &= \{(u,v) \in \mathbf{I}_1 \mid u \le G(\beta) - F(\alpha)\}, \\ \mathbf{I}^G_{1-2} &= \{(u,v) \in \mathbf{I}_1 \mid u > G(\beta) - F(\alpha),\ v > u - (G(\beta) - F(\alpha))\}, \\ \mathbf{I}^G_{1-3} &= \{(u,v) \in \mathbf{I}_1 \mid v \le u - (G(\beta) - F(\alpha))\}. \end{aligned}$$

When $G(\beta) \le F(\alpha)$, we divide $\mathbf{I}_1$ into

$$\begin{aligned} \mathbf{I}^F_{1-1} &= \{(u,v) \in \mathbf{I}_1 \mid v \le F(\alpha) - G(\beta)\}, \\ \mathbf{I}^F_{1-2} &= \{(u,v) \in \mathbf{I}_1 \mid v > F(\alpha) - G(\beta),\ v \le u + (F(\alpha) - G(\beta))\}, \\ \mathbf{I}^F_{1-3} &= \{(u,v) \in \mathbf{I}_1 \mid v > u + (F(\alpha) - G(\beta))\}. \end{aligned}$$

See also Fig. 16.5. Then,

$$C_{\alpha,\beta}(u,v) = \begin{cases} 0 & (u,v) \in \mathbf{I}^G_{1-1} \\ u - (G(\beta) - F(\alpha)) & (u,v) \in \mathbf{I}^G_{1-2} \\ v & (u,v) \in \mathbf{I}^G_{1-3} \\ 0 & (u,v) \in \mathbf{I}^F_{1-1} \\ v - (F(\alpha) - G(\beta)) & (u,v) \in \mathbf{I}^F_{1-2} \\ u & (u,v) \in \mathbf{I}^F_{1-3}. \end{cases} \tag{16.5}$$

Similarly, for $(u,v) \in \mathbf{I}_2$, we have

$$C_{\alpha,\beta}(u,v) = u + \min\{u+F(\alpha), v-(1-G(\beta))\} - \min\{F(\alpha), v-(1-G(\beta))\} \\ - \min\{u+F(\alpha), G(\beta)\} + \min\{F(\alpha), G(\beta)\}.$$

When $G(\beta) > F(\alpha)$, we divide $\mathbf{I}_2$ into

$$\begin{aligned} \mathbf{I}^G_{2-1} &= \{(u,v) \in \mathbf{I}_2 \mid u \le G(\beta) - F(\alpha),\ v \le 1 - (G(\beta) - F(\alpha))\}, \\ \mathbf{I}^G_{2-2} &= \{(u,v) \in \mathbf{I}_2 \mid u > G(\beta) - F(\alpha),\ v \le 1 - (G(\beta) - F(\alpha))\}, \\ \mathbf{I}^G_{2-3} &= \{(u,v) \in \mathbf{I}_2 \mid v > u + \{1 - (G(\beta) - F(\alpha))\}\}, \\ \mathbf{I}^G_{2-4} &= \{(u,v) \in \mathbf{I}_2 \mid u \le G(\beta) - F(\alpha), \\ &\qquad 1 - (G(\beta) - F(\alpha)) < v \le u + \{1 - (G(\beta) - F(\alpha))\}\}, \\ \mathbf{I}^G_{2-5} &= \{(u,v) \in \mathbf{I}_2 \mid u > G(\beta) - F(\alpha),\ v > 1 - (G(\beta) - F(\alpha))\}. \end{aligned}$$

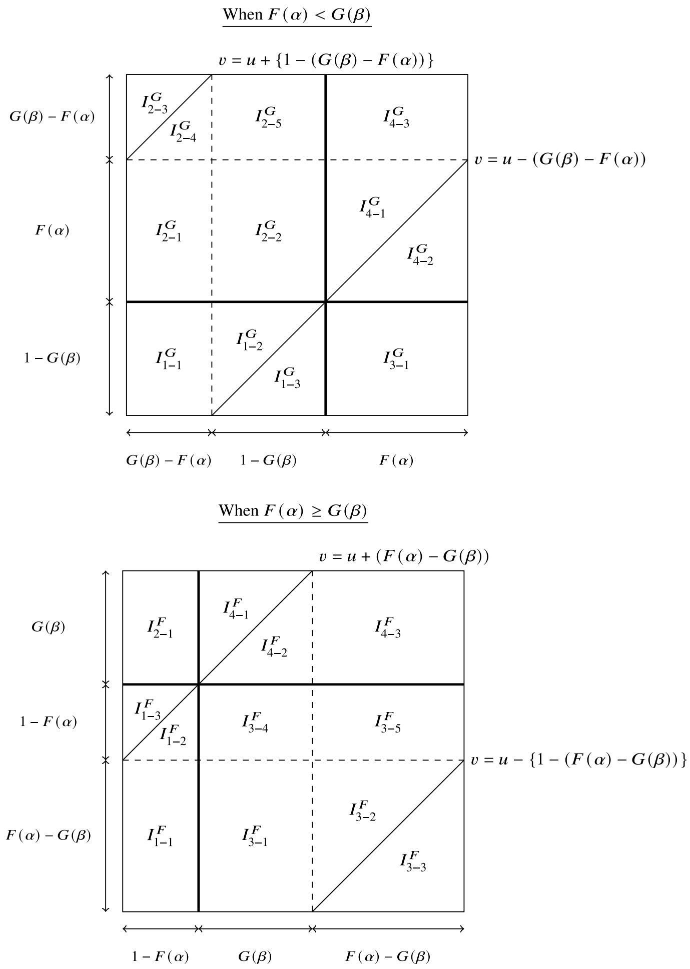


Fig. 16.5 Division of $\mathbf{I}^2 = [0, 1]^2$

When $G(\beta) \leq F(\alpha)$, we do not divide $\mathbf{I}_2$ further and rename it as $\mathbf{I}^F_{2-1}$. Then,

$$C_{\alpha,\beta}(u, v) = \begin{cases} 0 & (u, v) \in \mathbf{I}^G_{2-1} \\ u - (G(\beta) - F(\alpha)) & (u, v) \in \mathbf{I}^G_{2-2} \\ u & (u, v) \in \mathbf{I}^G_{2-3} \\ v - \{1 - (G(\beta) - F(\alpha))\} & (u, v) \in \mathbf{I}^G_{2-4} \\ u + v - 1 & (u, v) \in \mathbf{I}^G_{2-5} \\ u & (u, v) \in \mathbf{I}^F_{2-1}. \end{cases} \tag{16.6}$$

For $(u, v) \in \mathbf{I}_3$, we have

$$C_{\alpha,\beta}(u,v) = v + \min\{u-(1-F(\alpha)), v+G(\beta)\} - \min\{u-(1-F(\alpha)), G(\beta)\} - \min\{F(\alpha), v+G(\beta)\} + \min\{F(\alpha), G(\beta)\}.$$

When $G(\beta) > F(\alpha)$, we do not divide $\mathbf{I}_3$ further and rename it $\mathbf{I}^G_{3-1}$. When $G(\beta) \leq F(\alpha)$, we divide $\mathbf{I}_3$ into

$$\begin{aligned}
\mathbf{I}^F_{3-1} &= \{(u,v) \in \mathbf{I}_3 \mid u \leq 1-(F(\alpha)-G(\beta)),\ v \leq F(\alpha)-G(\beta)\},\\
\mathbf{I}^F_{3-2} &= \{(u,v) \in \mathbf{I}_3 \mid u > 1-(F(\alpha)-G(\beta)),\\
&\qquad v \leq F(\alpha)-G(\beta),\ v > u-\{1-(F(\alpha)-G(\beta))\}\},\\
\mathbf{I}^F_{3-3} &= \{(u,v) \in \mathbf{I}_3 \mid v \leq u-\{1-(F(\alpha)-G(\beta))\}\},\\
\mathbf{I}^F_{3-4} &= \{(u,v) \in \mathbf{I}_3 \mid u \leq 1-(F(\alpha)-G(\beta)),\ v > F(\alpha)-G(\beta)\},\\
\mathbf{I}^F_{3-5} &= \{(u,v) \in \mathbf{I}_3 \mid u > 1-(F(\alpha)-G(\beta)),\ v > F(\alpha)-G(\beta)\}.
\end{aligned}$$

Then,

$$C_{\alpha,\beta}(u,v) = \begin{cases} v & (u,v) \in \mathbf{I}^G_{3-1}\\ 0 & (u,v) \in \mathbf{I}^F_{3-1}\\ u-\{1-(F(\alpha)-G(\beta))\} & (u,v) \in \mathbf{I}^F_{3-2}\\ v & (u,v) \in \mathbf{I}^F_{3-3}\\ v-(F(\alpha)-G(\beta)) & (u,v) \in \mathbf{I}^F_{3-4}\\ u+v-1 & (u,v) \in \mathbf{I}^F_{3-5}. \end{cases} \tag{16.7}$$

For $(u,v) \in \mathbf{I}_4$, we have

$$\begin{aligned}
&C_{\alpha,\beta}(u,v)\\
&= u+v-1+\min\{u-(1-F(\alpha)), v-(1-G(\beta))\}\\
&\quad - \min\{F(\alpha), v-(1-G(\beta))\} - \min\{u-(1-F(\alpha)), G(\beta)\} + \min\{F(\alpha), G(\beta)\}.
\end{aligned}$$

When $G(\beta) > F(\alpha)$, we divide $\mathbf{I}_4$ into

$$\begin{aligned}
\mathbf{I}^G_{4-1} &= \{(u,v) \in \mathbf{I}_4 \mid v \leq u-(G(\beta)-F(\alpha))\},\\
\mathbf{I}^G_{4-2} &= \{(u,v) \in \mathbf{I}_4 \mid v \leq 1-(G(\beta)-F(\alpha)),\ v > u-(G(\beta)-F(\alpha))\},\\
\mathbf{I}^G_{4-3} &= \{(u,v) \in \mathbf{I}_4 \mid v > 1-(G(\beta)-F(\alpha))\}.
\end{aligned}$$

When $G(\beta) \leq F(\alpha)$, we divide $\mathbf{I}_4$ into

$$\begin{aligned}
\mathbf{I}^F_{4-1} &= \{(u,v) \in \mathbf{I}_4 \mid v > u+(F(\alpha)-G(\beta))\},\\
\mathbf{I}^F_{4-2} &= \{(u,v) \in \mathbf{I}_4 \mid u \leq 1-(F(\alpha)-G(\beta)),\ v \leq u+(F(\alpha)-G(\beta))\},\\
\mathbf{I}^F_{4-3} &= \{(u,v) \in \mathbf{I}_4 \mid u > 1-(F(\alpha)-G(\beta))\}.
\end{aligned}$$

Then,

$$C_{\alpha,\beta}(u,v) = \begin{cases} v & (u,v) \in \mathbf{I}^G_{4-1} \\ u - (G(\beta) - F(\alpha)) & (u,v) \in \mathbf{I}^G_{4-2} \\ u + v - 1 & (u,v) \in \mathbf{I}^G_{4-3} \\ u & (u,v) \in \mathbf{I}^F_{4-1} \\ v - (F(\alpha) - G(\beta)) & (u,v) \in \mathbf{I}^F_{4-2} \\ u + v - 1 & (u,v) \in \mathbf{I}^F_{4-3}. \end{cases} \tag{16.8}$$

From (16.5)–(16.8), we find $C_{\alpha,\beta}(u,v)$ is equivalent to $M_a(u,v)$ in Theorem 16.1 by taking

$$a = \begin{cases} 1 - (G(\beta) - F(\alpha)) & (F(\alpha) \leq G(\beta)) \\ F(\alpha) - G(\beta) & (F(\alpha) > G(\beta)). \end{cases}$$

Acknowledgements Financial support for this research was received from JSPS KAKENHI in the form of grant 18K11193. We would like to thank Editage (www.editage.jp) for English language editing. Thanks are extended to an anonymous referee for valuable comments improving the paper.

References

1. Fisher, N. I. (1993). *Statistical Analysis of Circular Data.* Cambridge: Cambridge University Press.
2. Fisher, N. I. and Lee, A. J. (1983). A correlation coefficient for circular data. *Biometrika.* **70** 327–332.
3. Jammalamadaka, S. A. and SenGupta, A. S. (2001). *Topics in Circular Statistics.* New York: World Scientific Publishing Co.
4. Jones, M. C., Pewsey, A. and Kato, S. (2015). On a class of circulas: copulas for circular distributions. *Ann Inst Stat Math.* **67** 843–862.
5. Mardia, K. V. (1970). *Families of Bivariate Distributions.* Darien, Connecticut: Hafner Publishing Company.
6. Mardia, K. V. and Jupp, P. E. (1999). *Directional Statistics.* Chichester: Wiley.
7. Nelsen, R. B. (2006). *An Introduction to Copulas.* New York: Springer.
8. Taniguchi, M., Kato, S., Ogata, H. and Pewsey, A. (2020), Models for circular data from time series spectra. *J. Time Series Anal.* **41** 808–829.

Chapter 17
Topological Data Analysis for Directed Dependence Networks of Multivariate Time Series Data

Anass El Yaagoubi and Hernando Ombao

Abstract Topological data analysis (TDA) approaches are becoming increasingly popular for studying the dependence patterns in multivariate time series data. In particular, various dependence patterns in brain networks may be linked to specific tasks and cognitive processes, which can be altered by various neurological impairments such as epileptic seizures. Existing TDA approaches rely on the notion of distance between data points that is symmetric by definition for building graph filtrations. For brain dependence networks, this is a major limitation that constrains practitioners from using only symmetric dependence measures, such as correlations or coherence. However, it is known that the brain dependence network may be very complex and can contain a directed flow of information from one brain region to another. Such dependence networks are usually captured by more advanced measures of dependence such as partial directed coherence, which is a Granger causality-based dependence measure. These dependence measures will result in a non-symmetric distance function, especially during epileptic seizures. In this paper, we propose to solve this limitation by decomposing the weighted connectivity network into its symmetric and anti-symmetric components using matrix decomposition and comparing the anti-symmetric component prior to and post seizure. Our analysis of epileptic seizure EEG data shows promising results.

17.1 Introduction

Over the past twenty years, topological data analysis (TDA) has witnessed significant advances that aim to study and understand various patterns in the data, thanks to the pioneering work of [11, 13, 14, 19]. Many TDA techniques for the analysis of various types of data have emerged, like barcodes, persistence diagrams, and persistence

A. El Yaagoubi · H. Ombao (✉)
King Abdullah University of Science and Technology, Thuwal 23955-6900 Thuwal, Kingdom of Saudi Arabia
e-mail: hernando.ombao@kaust.edu.sa

A. El Yaagoubi
e-mail: anass.bourakna@kaust.edu.sa

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_17

landscapes. These TDA features aim to summarize the intrinsic shape of the data (cloud of points or weighted network usually), by keeping track of the specific scales at which any topological feature (e.g., holes and cavities, etc.) is either born or dead. The goal of such approaches is to provide insight on the geometrical patterns included in high dimensional data that are not visible using conventional approaches.

There has been an increasing trend in the literature that applies the TDA techniques to data sets with a temporal structure (e.g., financial time series, brain signals, etc.). For such data sets, it has been proposed over the years that dependence networks of multivariate time series data, particularly for multivariate brain signals such as electroencephalograms (EEG) and local field potentials (LFP), can provide promising results, as shown in [15].

The topology of the structural and functional brain networks is organized according to principles that maximize the (directed-)flow of information and minimize the energy cost for maintaining the entire network, such as small world networks, see [35, 39, 43]. More formally, define $X(t) \in \mathbb{R}^d$ to be a vector-valued stochastic brain process at time t. Due to physiological reasons, the influence of $X_q(t)$ on $X_p(t)$ may differ from the influence of $X_p(t)$ on $X_q(t)$, where $X_p(t)$ and $X_q(t)$ are two univariate time series components of $X(t)$.

Neurological disorders such as Alzheimer's disease, Parkinson's disease, and Epilepsy may alter the topological structure of the brain network. For instance, it is known that the onset of epileptic seizures can alter brain networks and, in particular, brain connectivity, leading to multiple studies that compare the pre-ictal, inter-ictal, and post-ictal network characteristics. However, studying this change in the topology of the brain network based on a symmetric measure of dependence such as correlation or coherence is going to result in the loss of information regarding the spatio-temporal evolution of the seizure process (how abnormal electrical activity in one brain region may spread to other regions). Indeed, it is known that brain seizures are usually initiated from a localized region (or multiple sub-regions), which then propagate to the rest of the network, for more details refer to [5, Chap. 1]. Such propagation ought to be non-symmetric, as there is a directed flow of information going from the source region to the rest of the brain.

Therefore, in order to capture this asymmetry in the brain connectivity we ought to use a non-symmetric measure of dependence, that allows for the influence of a channel $X_q(t)$ on another channel $X_p(t)$ to be different from the reverse influence, i.e., of channel $X_p(t)$ on channel $X_q(t)$. In particular, partial directed coherence (PDC), which estimates the intensity of information flow from one brain region to another based on the notion of the Granger causality, i.e., the ability of time series components to predict each other at various lag values. Refer to Sect. 17.2 for more details regarding PDC. Using the distance measure based on PDC in [15] introduces the following problem. For a function d to be a valid distance function it has to respect the four axioms of a distance:

Axiom 17.1 $d(x, y) \geq 0$;

Axiom 17.2 $d(x, y) = 0 \iff x = y$;

Axiom 17.3 $d(x, y) = d(y, x)$;

Axiom 17.4 $d(x, y) \leq d(x, z) + d(z, y)$.

After defining the distance using PDC (e.g., $d = 1 - \text{PDC}$), we notice immediately that it fails to satisfy Axiom 17.3, i.e., $d(x, y) \neq d(y, x)$ which is a significant restriction. Some of the aforementioned axioms (such as the second and the fourth) can be relaxed in some situations. However, Axiom 17.3 cannot be ignored, and doing so will lead to major conflicts and contradictions when creating the Rips–Vietoris filtration.

Motivated by this challenge, we propose a novel approach that will allow existing TDA techniques to assess the topology of the asymmetry in oriented brain networks. We propose an approach based on the decomposition of a weighted network into its symmetric and anti-symmetric components. In Sect. 17.2, we provide a brief review of vector autoregressive (VAR) models and PDC. In Sect. 17.3, we present the network decomposition approach. In Sect. 17.4, we provide a brief review of persistent homology and Vietoris–Rips filtration. In Sect. 17.5, we analyze the impact of an epileptic seizure on the flow of information within the brain using an EEG data set.

17.2 VAR Models and PDC

One standard approach to modeling brain signals is to use parametric models such as VAR models, see [22, 23, 29, 30]. This class of models is very flexible, and has the ability to capture all linear dependencies in the multivariate time series, see [31]. Given a stationary multivariate time series $X(t) \in \mathbb{R}^d$, the VAR model of order K is expressed in the following way:

$$X(t) = \sum_{k=1}^{K} \Phi_k X(t-k) + E(t), \tag{17.1}$$

$$E(t) \sim \mathcal{N}(0, \Sigma_E), \tag{17.2}$$

where $X(t)$ is d-dimensional vector of observations, and Φ_k is $d \times d$ mixing weight matrix for lag k, which are chosen such that $X(t)$ is causal, $E(t)$ is the innovations vector. The model order K is usually selected based on various information criteria, such as Akaike information criterion (AIC) and Bayesian information criterion (BIC), see [2, 41, 42].

The concept of PDC, as presented in [3], is supposed to represent the concept of Granger causality in the frequency domain, as it measures the intensity of information flow between a pair of EEG channels. Assuming we fit a VAR model to the observed multivariate brain signals, we get the following expression for the Fourier transform of the VAR model parameters:

$$\overline{A}(\omega) = I - \sum_{k=1}^{K} \Phi_k \exp(-i2\pi k\omega), \tag{17.3}$$

which in turn is used as follows to compute the PDC:

$$\text{PDC}_{p,q}(\omega) = \frac{\left|\overline{A}_{p,q}(\omega)\right|}{\sqrt{\overline{A}^H_{.,q}(\omega)\overline{A}_{.,q}(\omega)}}, \tag{17.4}$$

where $\text{PDC}_{p,q}(\omega)$ denotes the direction and intensity of the information flow from channel q to channel p at the frequency ω and the symbol ^{H} denotes the Hermitian transpose, when the dimension d is equal to one then the Hermitian transpose becomes the complex conjugate. From this definition, it is obvious that PDC is not symmetric measure of information as $\text{PDC}_{p,q}(\omega) \neq \text{PDC}_{q,p}(\omega)$. Such a PDC cannot be used directly with TDA to analyze the shape of the network. Therefore, another approach is necessary, which we will develop in the next section.

Assume we observe two multivariate processes $Y^{(1)}(t)$ and $Y^{(2)}(t)$ that are mixtures of AR(2) processes $Z_j^{(k)}(t)$ as follows:

$$\begin{aligned}
Y_1^{(1)}(t) &= Y_5^{(1)}(t-1) + Z_1^{(1)}(t),\\
Y_2^{(1)}(t) &= Y_1^{(1)}(t-1) + Y_4^{(1)}(t-1) + Z_2^{(1)}(t),\\
Y_3^{(1)}(t) &= Y_2^{(1)}(t-1) + Y_4^{(1)}(t-1) + Z_3^{(1)}(t),\\
Y_4^{(1)}(t) &= Y_2^{(1)}(t-1) + Y_3^{(1)}(t-1) + Z_4^{(1)}(t),\\
Y_5^{(1)}(t) &= Y_4^{(1)}(t-1) + Z_5^{(1)}(t),
\end{aligned}$$

and

$$\begin{aligned}
Y_1^{(2)}(t) &= Y_5^{(2)}(t-1) + Z_1^{(2)}(t),\\
Y_2^{(2)}(t) &= Y_1^{(2)}(t-1) + Y_4^{(1)}(t-1) + Z_2^{(2)}(t),\\
Y_3^{(2)}(t) &= Y_2^{(2)}(t-1) + Z_3^{(2)}(t),\\
Y_4^{(2)}(t) &= Y_3^{(2)}(t-1) + Z_4^{(2)}(t),\\
Y_5^{(2)}(t) &= Y_4^{(2)}(t-1) + Z_5^{(2)}(t),
\end{aligned}$$

with distinct lead-lag directed dependence networks, as can be seen in Fig. 17.1. Choosing $Z_j^{(k)}(t)$ to be AR(2) processes allows this dependency pattern to be frequency specific. Both networks are oriented and have multiple edges in common; however, they present distinct oriented connectivity among the nodes 2–4.

Using a non-oriented measure of dependence such as coherence or correlation would result in misleading conclusions, as these measures will ignore the orientation of the dependence structure. Hence, the two distinct structures will be confused into

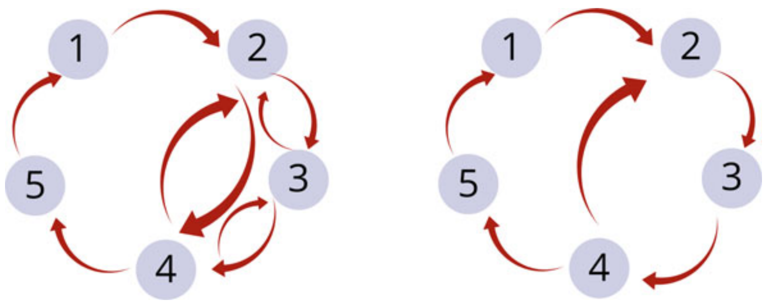


Fig. 17.1 Example of distinct directed dependence networks. Symmetric and non-symmetric cyclic structure (Left) and non-symmetric only cyclic structure (Right)

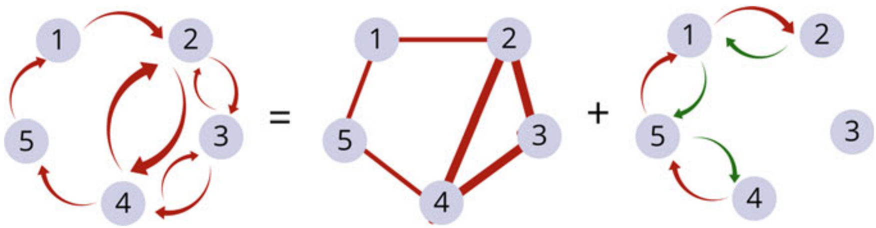


Fig. 17.2 First example of non-symmetric network decomposition. Original network (Left), symmetric component (Middle), and anti-symmetric component (Right)

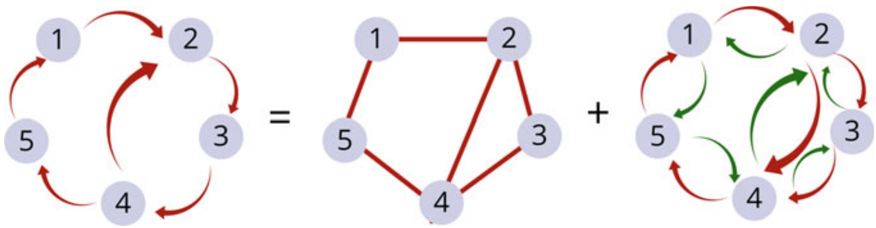


Fig. 17.3 Second example of non-symmetric network decomposition. Original network (Left), symmetric component (Middle), and anti-symmetric component (Right)

the same structure. However, using an oriented measure of dependence such as PDC will fit the structure properly. Therefore, the latter approach will enable us to take into account the orientation in the dependence structure by decomposing oriented networks into their symmetric and anti-symmetric components as can be seen in Figs. 17.2 and 17.3.

From Figs. 17.2 and 17.3, it is clear that applying TDA to the symmetric component (Middle) will results in disastrous conclusion, as it will not be able to distinguish between the two networks, since they share very similar topological structure. However, applying TDA to the anti-symmetric component (Right) will capture the topo-

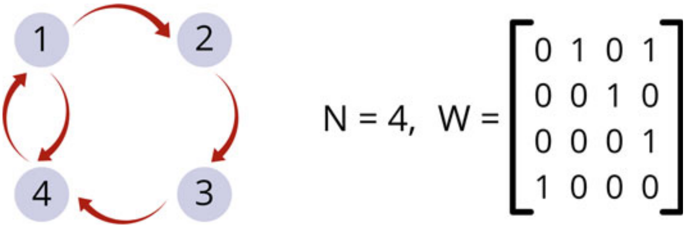


Fig. 17.4 Original weight matrix and graph representation

logical differences between the two networks, as the first example doesn't contain any asymmetric cycles, whereas the second example contains two main cycles.

17.3 Network Decomposition

There are multiple techniques for decomposing weighted network into multiple networks with various properties. In this section, we consider the transformation that decomposes a network into a unique pair of symmetric and anti-symmetric components. Given an oriented weighted network $G = (N, W)$, see Fig. 17.4, with $|N|$ nodes and weight matrix W, we propose the following graph decomposition that results in two graphs G_s and G_a, where $G_s = (N, W_s)$ is the symmetric component and $G_a = (N, W_a)$ is the anti-symmetric component as can be seen in Figs. 17.5 and 17.6.

$$W_s = \frac{1}{2}(W + W^{'}), \tag{17.5}$$

$$W_a = \frac{1}{2}(W - W^{'}), \tag{17.6}$$

where the signs $+$ and $-$ are the usual matrix addition and subtraction, and $W_s^{'} = \frac{1}{2}(W^{'} + W) = W_s$ is a symmetric weight matrix corresponding to the symmetric component G_s and $W_a^{'} = \frac{1}{2}(W^{'} - W) = -W_a$ the anti-symmetric weight matrix that corresponds to the anti-symmetric component. Note that $W = W_s \bigoplus W_a$. This decomposition is unique, since it is the result of the projection of the weight matrix W on the space of symmetric and anti-symmetric matrices, since W_s and W_a are respectively the closest symmetric and anti-symmetric matrices to W in terms of the Frobenius norm; furthermore, the spaces of symmetric and anti-symmetric matrices are orthogonal with respect to standard matrix inner product.

In order to analyze the alterations in the topology of the network following the onset of seizure and during the seizure episode, we can use either W_s or W_a. Since W_s is a symmetric matrix, it can be used through a monotonically decreasing function, as it is the case with correlation and coherence matrices. However, W_a being anti-

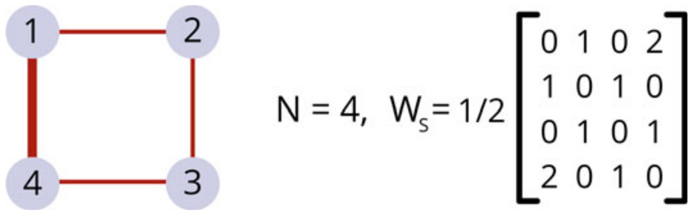


Fig. 17.5 Symmetric component representation

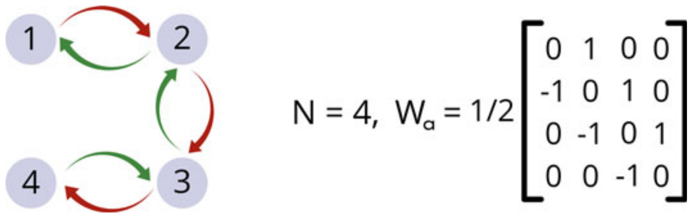


Fig. 17.6 Anti-symmetric component representation

symmetric, it needs to be transformed since a distance function cannot take negative values. Therefore, we propose to build the Rips–Vietoris filtration based on $|W_a|$, where the entries $|W_a|_{p,q} = \frac{1}{2}|W_{p,q} - W_{q,p}|$. Indeed, $|W_a|_{p,q}$ can be thought of as a distance, since it measures departure from symmetry. If W is perfectly symmetrical, then $\forall p, q$, $(W_a)_{p,q} = 0$, and as W departs from symmetry, i.e., $W_{p,q}$ changes and differs from $W_{q,p}$, $|W_a|_{p,q}$ becomes larger than zero. In the following section, we provide a brief review of persistent homology, before choosing to focus on $|W_a|$ to analyze the temporal evolution of the changes in directional asymmetry in brain connectivity during the seizure process.

17.4 Overview of Persistent Homology and Vietoris–Rips Filtration

The goal of persistent homology is to provide computational tools that can distinguish between topological objects. It analyzes their topological features, such as connected components, holes, cavities, etc. In this paper we consider weighted networks, with weights corresponding to some loose notion of distance as defined by $|W_a|$.

In order to analyze the shape of these networks, we need to build the homology of the data by looking at an increasing sequence of networks of neighboring data points at varying scales/distances, as seen in Fig. 17.7. The goal of this approach is to analyze the characteristics of the geometrical patterns when they appear (birth) then disappear (death) for a wide range of scales, see [11, 13, 14, 19]. The Vietoris–Rips (VR) filtration as presented above is constructed based on the notions of a simplex

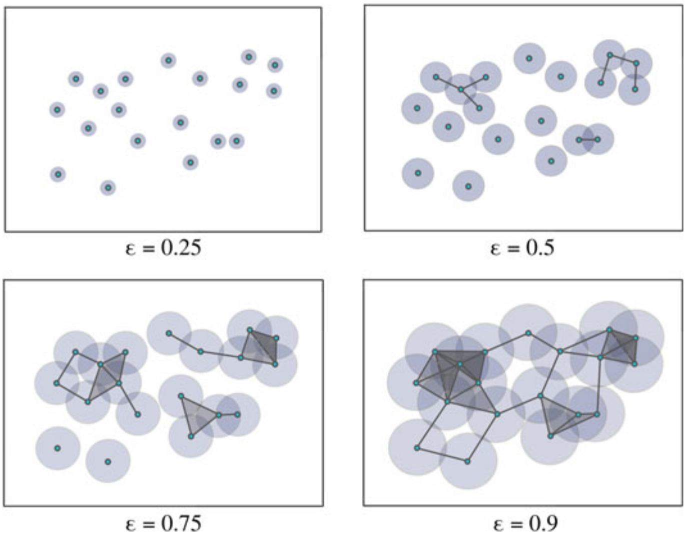


Fig. 17.7 Example of a Vietoris–Rips filtration on a cloud of points. As the radius of the balls grows, there's birth and death of various topological features

and simplicial complex, which can be thought of as a finite collection of sets that is closed under the subset relation, see Figs. 17.8 and 17.9. Simplicial complexes are generally understood as higher dimensional generalizations of graphs. Simplicial complexes are meant to represent in an abstract form the shape of the data at various scales, in order to simplify and allow abstract manipulations. Simplicial complexes as shown in Fig. 17.9 can be very simple like a group of disconnected nodes, or more complex, e.g., a combination of pairs of connected nodes, triplets of triangles, or any higher dimensional simplex. This notion of a simplicial complex can be understood as a generalization of the notion of networks, to include surfaces, volumes, and higher dimensional objects. Given this definition of simplicial complexes, we can formalize the notion of the VR filtration to be an increasing sequence of simplicial complexes. Let $X_p(t)$ be an observed time series of brain activity at location $p \in N$ and time $t \in \{1, \ldots, T\}$. Therefore, we can think of a distance between brain channels at locations p and q to be their distance from symmetry, i.e., imbalance in information flow $D(p, q) = |W_{a,(pq)}|$. Using this measure, we construct the Vietoris–Rips filtration by connecting nodes that have a distance less or equal to some given threshold ϵ, which results in the following filtration:

$$\mathcal{X}_{\epsilon_1} \subset \mathcal{X}_{\epsilon_2} \subset \cdots \subset \mathcal{X}_{\epsilon_n}, \tag{17.7}$$

where $0 < \epsilon_1 < \epsilon_2 < \cdots < \epsilon_{n-1} < \epsilon_n$ are the distance thresholds. Nodes within some given distance ϵ_i are connected to form different simplicial complexes, $\mathcal{X}_{\epsilon_1}$ is the first simplicial complex (single nodes) and $\mathcal{X}_{\epsilon_n}$ is the last simplicial complex

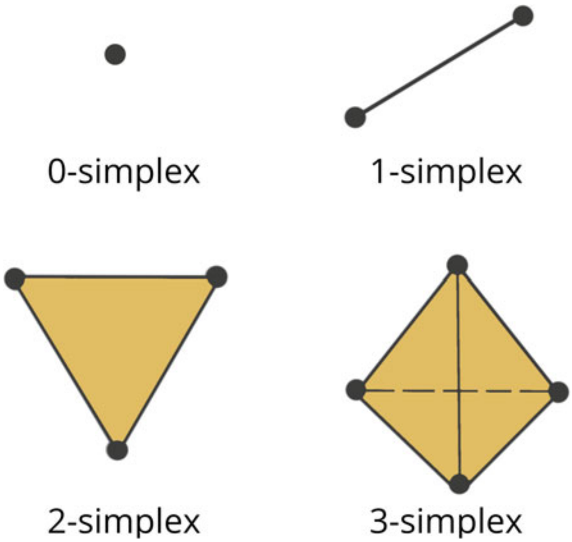


Fig. 17.8 Examples of the four lowest dimensional simplices

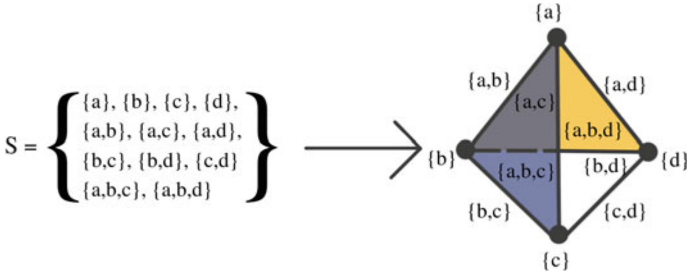


Fig. 17.9 Example of a simplicial complex with four nodes, six edges, and two faces. If S is a simplicial complex, then every face of a simplex in S must also be in S

(all nodes connected, i.e., a clique of size d, where d is the number of nodes or dimension of the multivariate time series). Refer to [26] for a review of how to build the VR filtrations.

The VR filtration is a complex object. Therefore, in practice practitioners consider a topological summary that is known as the persistence diagram (PD), which is a diagram that represents the times of birth and death of the topological features in the VR filtration as seen in Fig. 17.10. Every birth-death pair is represented by a point in the diagram, e.g., (ϵ_1, ϵ_2) and (ϵ_2, ϵ_3). The points in the PD are colored based on the dimension of the feature they correspond to (e.g., one color for the connected components, another color for the cycles, etc.).

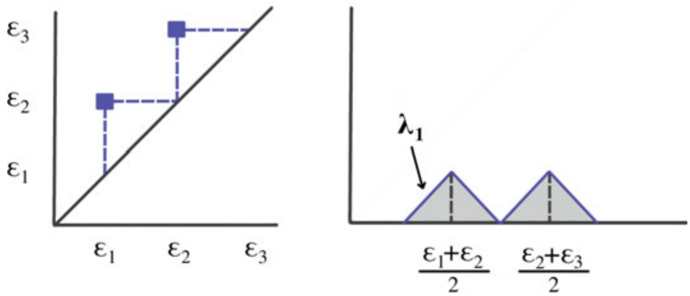


Fig. 17.10 Example of a PD (Top) and a PL (Bottom). This figure is inspired form the paper by [20]

Comparing PDs can be quite challenging and time consuming. Indeed, taking averages or computing distances, like the Bottleneck or Wasserstein distances between persistence diagrams is time consuming as it is necessary to find point correspondence, see [1]. For these reasons, we prefer to analyze a simpler representation of the PD called the persistence landscape (PL), which is a simpler object (one-dimensional function), as defined in [6]. The use of PLs lends to a more rigorous statistical analysis that produce more easily interpretable results. As the PLs are functions of a real variable, it is easy to compute group averages and to derive confidence regions.

17.5 Oriented TDA of Seizure EEG

We investigate the topological features of the asymmetric brain network component based on the decomposition presented previously by contrasting the network before and after the epileptic seizure. The data was collected from an epileptic patient at the Epilepsy Disorder Laboratory at the University of Michigan (PI: Beth Malow, M.D.). With a sampling rate 100 Hz and observation period of 8 min, the following results are based on 19 scalp differential electrodes (no reference), see Fig. 17.11.

To analyze this data set on the pre-seizure an seizure onset, we start by fitting a VAR model with dimension $d = 19$ (number of observed time series components) and order $K = 5$ to the pre-seizure part and on the seizure onset. Connectivity between all pairs of channels was assessed using PDC, for the delta (0–4 Hz), alpha (8–12 Hz), beta (12–30 Hz), and gamma (30–50 Hz) frequency bands. After we decompose each network into its symmetric and anti-symmetric components, then we build the VR filtration based on $|W_a^{\Omega}|$, where Ω stands for the frequency band of interest, delta, alpha, beta, gamma, etc. We report the PDs in Figs. 17.12, 17.13, 17.14, and 17.15.

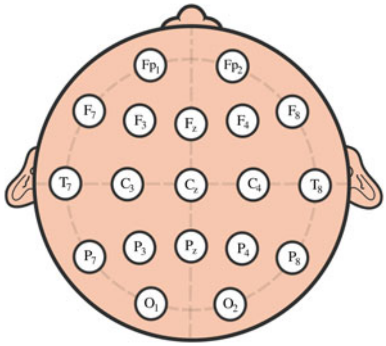


Fig. 17.11 Differential electrodes positioning map on the scalp

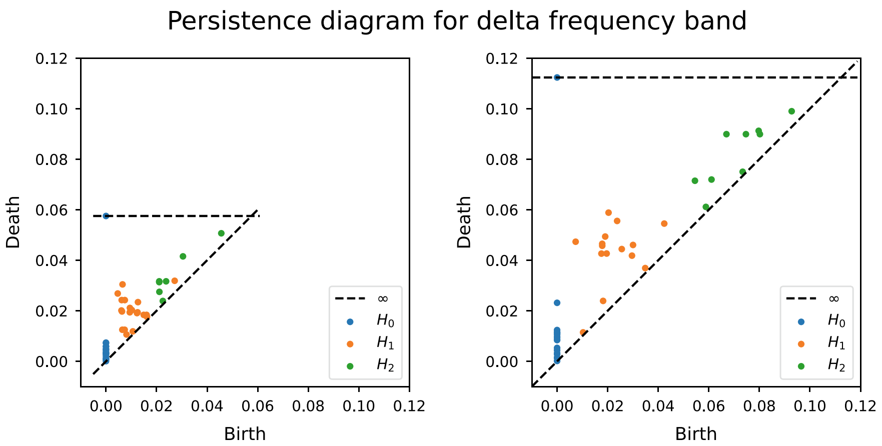


Fig. 17.12 PD based on the anti-symmetric brain component for the delta (0–4 Hz) frequency band. Pre-seizure (Left) and During-seizure (Right). It can be seen that zero- and one-dimensional features (blue and orange dots) are born around the same scale (0.0 and $0.01-0.02$) prior to and during the seizure and die at a much later scale during the seizure (0.11 and 0.06) than prior to the seizure (0.06 and 0.03). For the two-dimensional features (green dots), they seem to be born later and die later prior to and during the seizure

From the above PDs, we notice that the one-dimensional features persist more during the seizure than prior to the seizure mainly for delta, alpha, and beta frequency bands. However, even if the two-dimensional features seem to persist more during the seizure than prior to the seizure, the variability seems larger which makes conclusions uncertain.

In this context of epileptic seizure, the appearance of one-dimensional features means that the epileptic seizure tends to induce an asymmetric circular flow of information within the brain. The appearance of zero-dimensional features indicates the presence of segregated regions where the flow of information is asymmetric. The scale at which these features appear indicates the importance or magnitude of the

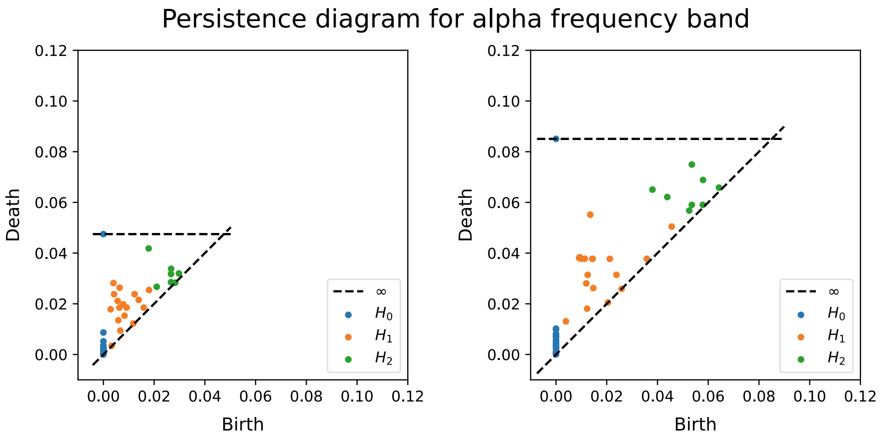


Fig. 17.13 PD based on the anti-symmetric brain component for the alpha (8–12 Hz) frequency band. Pre-seizure (Left) and During-seizure (Right). Similarly, it can be seen that zero- and one-dimensional features (blue and orange dots) are born around the same scale (0.0 and 0.01−0.02) prior to and during the seizure and die at a much later scale during the seizure (0.9 and 0.06) than prior to the seizure (0.05 and 0.03). For the two-dimensional features (green dots), they seem to be born later and die later as well prior to and during the seizure

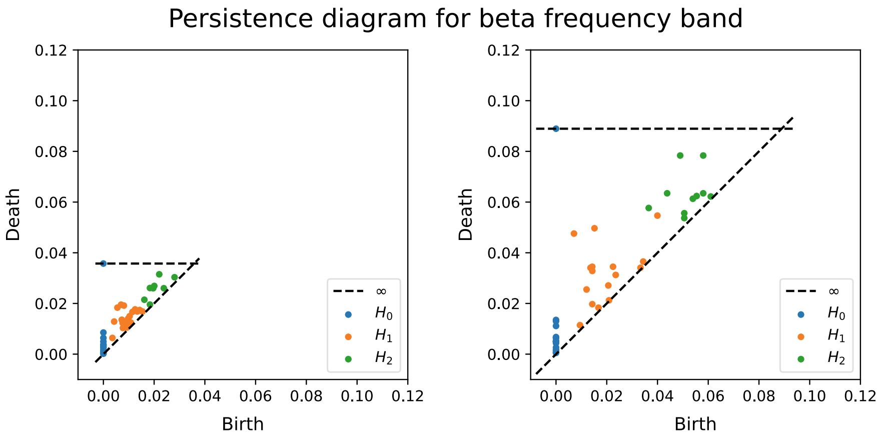


Fig. 17.14 PD based on the anti-symmetric brain component for the beta (12–30 Hz) frequency band. Pre-seizure (Left) and During-seizure (Right). The one-dimensional features (orange dots) seem to be born and die at a later scale respectively at (0.01−0.04) and (0.05) during the seizure than prior to the seizure (0.01−0.02) and (0.01−0.02). Similarly, the two-dimensional features (green dots) seem to be born later and die later during the seizure than prior to the seizure

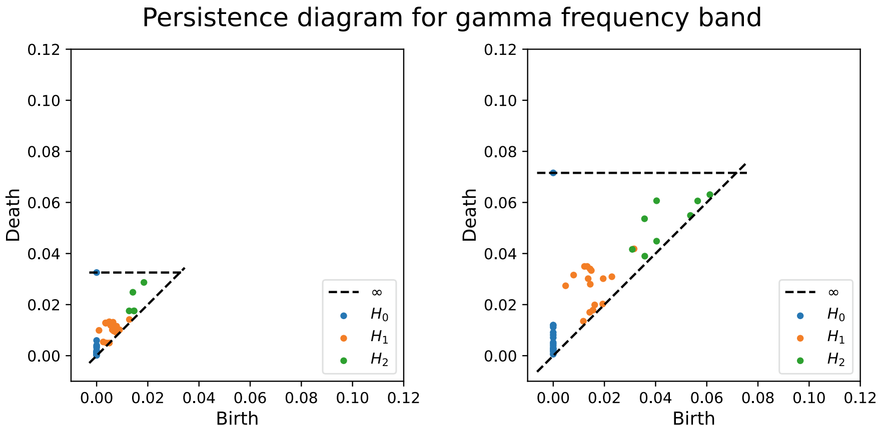


Fig. 17.15 PD based on the anti-symmetric brain component for the gamma (30–50 Hz) frequency band. Pre-seizure (Left) and During-seizure (Right). The zero- and one-dimensional features (blue and orange dots) seem to be born at the same time (0.0) and (0.01−0.02) but die at a later scale (0.04) during the seizure than prior to the seizure (0.03). Similarly, the two-dimensional features (green dots) seem to be born later and die later during the seizure than prior to the seizure

asymmetry, similarly the scale at which these features die indicates the maximum magnitude that the asymmetry within such features might reach.

The fact that these features appear at higher scales indicates that the epileptic seizure induces a large asymmetry in the flow of information within brain regions. Furthermore, since these features persist for larger scales, it means that the epileptic seizure induces a non-homogeneous asymmetry throughout the brain regions.

17.6 Conclusion

Topological data analysis has been built upon the notion of distance, which is symmetric by definition. When applied to brain connectivity networks, current TDA tools encounter serious limitations as they can only analyze brain networks based on symmetric dependence measures such as correlations or coherence. We have presented a new approach to analyze asymmetric brain networks using more general measures of dependence. Our approach relies on network decomposition, using weight matrix projections onto symmetric and anti-symmetric matrix spaces. Using PDC, our approach has been able to provide new insight into the topological alterations induced by an epileptic seizure. Our preliminary analysis provides evidence that epileptic seizure induces asymmetric flow of information across the brain, and that the induced asymmetry is non-homogeneous throughout brain regions. Even if we can detect seizure-induced asymmetry in the flow of information within the brain, our approach cannot detect the origin of this asymmetry. In future work, we plan to combine our oriented TDA approach with graph theoretic measures to detect the nature of the seizure as well as the regions at the origin of the seizure.

References

1. Agami, S. (2021). Comparison of persistence diagrams. *Communications in Statistics-Simulation and Computation* 1–14.
2. Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control* **19** 716–723.
3. Baccala, L. and Sameshima, K. (2001). Partial directed coherence: A new concept in neural structure determination. *Biological Cybernetics* **84** 463–474.
4. Bassett, D. and Bullmore, E. (2009). Human brain networks in health and disease. *Current opinion in neurology* **22** 340–347.
5. Bromfield, E. B., Cavazos, J. E. and Sirven, J. I. (2006). *An Introduction to Epilepsy*. American Epilepsy Society.
6. Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research* **16** 77–102.
7. Bubenik, P. (2020). The persistence landscape and some of its properties. *Topological Data Analysis: The Abel Symposium* **15** 97–117.
8. Bullmore, E. and Sporns, O. (2009). Complex brain networks: Graph theoretical analysis of structural and functional systems. *Nature reviews Neuroscience* **10** 186–98.
9. Cadonna, A., Kottas, A. and Prado, R. (2019). Bayesian spectral modeling for multiple time series. *Journal of the American Statistical Association* **114** 1838–1853.
10. Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society* **46** 255–308.
11. Carlsson, G., Zomorodian, A., Collins, A. and Guibas, L. (2004). Persistence barcodes for shapes. *Association for Computing Machinery* 124–135.
12. Cohen- Steiner, D., Edelsbrunner, H. and Harer, J. (2007). Stability of persistence diagrams. *Discrete and computational geometry* **37** 103–120.
13. Edelsbrunner, H. and Harer, J. (2008). Persistent homology-a survey. *Discrete and Computational Geometry* **453** 257–282.
14. Edelsbrunner, H., Letscher, D. and Zomorodian, A. (2002). Topological persistence and simplification. *Discrete & Computational Geometry* **28** 511–533.
15. El- Yaagoubi, A., Chung, M. K. and Ombao, H. (2022). Topological data analysis for multivariate time series data. *unpublished manuscript.*
16. Fiecas, M., Ombao, H., Linkletter, C., Thompson, W. and Sanes, J. (2010). Functional connectivity: shrinkage estimation and randomization test. *Neuroimage* **4** 15–49.
17. Fiecas, M., Ombao, H., van Lunen, D., Baumgartner, R., Coimbra, A. and Feng, D. (2013). Quantifying temporal correlations: a test-retest evaluation of functional connectivity in resting-state fmri. *Neuroimage* **65** 231–41.
18. Gholizadeh, S. and Zadrozny, W. (2018). A short survey of topological data analysis in time series and systems analysis. *arXiv*: arxiv.org/abs/1809.10745.
19. Ghrist, R. (2008). Barcodes: The persistent topology of data. *Bulletin of the American Mathematical Society* **45** 61–75.
20. Gidea, M. and Katz, Y. (2018). Topological data analysis of financial time series: Landscapes of crashes. *Physica A: Statistical Mechanics and its Applications* **491** 820–834.
21. Giusti, C., Pastalkova, E., Curto, C. and Itskov, V. (2015). Clique topology reveals intrinsic geometric structure in neural correlations. *Proceedings of the National Academy of Sciences* **112** 13455–13460.
22. Gorrostieta, C., Ombao, H., Bedard, P. and Sanes, J. (2012). Investigating stimulus-induced changes in connectivity using mixed effects vector autoregressive models. *NeuroImage* **59** 3347–3355.

23. Gorrostieta, C., Fiecas, M., H Ombao, E. B. and Cramer, S. (2013). Hierarchical vector auto-regressive models and their applications to multi-subject effective connectivity. *Frontiers in Computational Neuroscience* **159** 1–11.
24. Granados- Garcia, G., Fiecas, M., Babak, S., Fortin, N. J. and Ombao, H. (2021). Brain waves analysis via a non-parametric bayesian mixture of autoregressive kernels. *Computational Statistics & Data Analysis* **174** 107409.
25. Granger, C. W. J. (1969). Investigating causal relations by econometric models and cross-spectral methods. *Econometrica* **37** 424–438.
26. Hausmann, J.- C. (2016). *On the Vietoris-Rips complexes and a Cohomology Theory for metric spaces*. vol 138. Princeton University Press.
27. Hilgetag, C. and Goulas, A. (2015). Is the brain really a small-world network? *Brain structure & function* **221** 2361–2366.
28. Hoffmann- Jorgensen, J. and Pisier, G. (1976). The law of large numbers and the central limit theorem in banach spaces. *The Annals of Probability* **4** 587–599.
29. Hu, L., Fortin, N. and Ombao, H. (2019). Vector autoregressive models for multivariate brain signals. *Statistics in the Biosciences* **11** 91–126.
30. Kirch, C., Muhsal, B. and Ombao, H. (2015). Detection of changes in multivariate time series with application to eeg data. *Journal of the American Statistical Association* **110** 1197–1216.
31. Lütkepohl, H. (1991). *New Introduction to Multiple Time Series Analysis*. Springer.
32. Ledoux, M. and Talagrand, M. (2011). *Probability in Banach Spaces. Classics in Mathematics*. Springer-Verlag, Berlin.
33. Leuchter, A. F., Newton, T. F., Cook, I. A., Walter, D. O., Rosenberg-Thompson, S. and Lachenbruch, P. A. (1992). Changes in brain functional connectivity in alzheimer-type and multi-infarct dementia. *Brain* **115** 1543–1561.
34. Merkulov, S. (2003). Algebraic topology. *Proceedings of the Edinburgh Mathematical Society* **46**.
35. Muldoon, S., Bridgeford, E. and Bassett, D. (2016). Small-world propensity and weighted brain networks. *Scientific Reports* **6** 22057.
36. Munkres, J. R. (1984). *Elements of Algebraic Topology*. Addison Wesley Publishing Company.
37. Ombao, H. and Pinto, M. (2021). Spectral dependence. *arXiv*: arxiv.org/abs/2103.17240.
38. Ombao, H. and Van Bellegem, S. (2008). Evolutionary coherence of nonstationary signals. *IEEE Transactions on Signal Processing* **56** 2259–2266.
39. Pessoa, L. (2014). Understanding brain networks and brain organization. *Physics of Life Reviews* **11** 400–435.
40. Prado, R. (2013). Sequential estimation of mixtures of structured autoregressive models. *Computational Statistics & Data Analysis* **58** 58–70.
41. Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6** 461–464.
42. Shumway, R. H. and Stoffer, D. S. (2005). *Time Series Analysis and Its Applications*. Springer-Verlag.
43. Sporns, O. (2013). Structure and function of complex brain networks. *Dialogues in Clinical Neuroscience* **15** 247–262.
44. Wang, Y., Ting, C.- M., Gao, X. and Ombao, H. (2019). Exploratory analysis of brain signals through low dimensional embedding. In *2019 9th International IEEE/EMBS Conference on Neural Engineering (NER)* 997–1002.

Chapter 18
Orthogonal Impulse Response Analysis in Presence of Time-Varying Covariance

Valentin Patilea and Hamdi Raïssi

Abstract In this paper, the orthogonal impulse response functions (OIRFs) are studied in the non-standard but quite common case where the covariance of the error vector is not constant in time. The usual approach for taking such covariance behavior into account consists in applying the standard tools to sub-periods of the whole sample. We underline that such a practice may lead to severe upward bias. We propose a new approach intended to give what we argue to be a more accurate summary of the time-varying OIRFs. This consists in averaging the Cholesky decomposition of nonparametric covariance estimators. In addition, an index is developed to evaluate the heteroscedasticity effect on the OIRFs analysis. The asymptotic behavior of the proposed estimators is investigated.

18.1 Introduction

In time series econometrics, it is common to investigate sub-samples of a full time series in order to capture changes in the data. Reference can be made to [11, 12], who considered rolling windows. In order to accommodate possible regime switches, [5] constituted different sub-periods for measuring monetary policy. The paper [29] split the data considered for the study according to the Federal Reserve operating procedures. The paper [17] proposed a pre-crisis, in-crisis and post-crisis split type to carry out a volatility spillover analysis, while [28] considered pre- and post-1985 financial liberalization samples. The papers [1, 6, 27] used both rolling windows and static periods, to describe non-constant dynamics in the series they studied.

Our main message, focused on orthogonal impulse response functions (OIRFs) analysis, is that if one wishes to work with fixed sub-samples (for comparisons of periods), it is advisable to carry out a pointwise estimation, and then summarize it

V. Patilea (✉)
CREST Ensai, Campus de Ker-Lann, 51 Rue Blaise Pascal, 37203, 35172 BRUZ Cedex, France
e-mail: patilea@ensai.fr

H. Raïssi
Pontificia Universidad Católica de Valparaíso, 2734 Valparaíso, Chile
e-mail: hamdi.raissi@pucv.cl

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_18

using averages according to the periods of interest. This leads us in the following to introduce what we call the *averaged OIRFs*. By doing so, an accurate picture of the non-constant dynamics is obtained. As a matter of fact, applying the standard tools to sub-samples can, in some sense, lead to bias distortions in summarizing the time-varying dynamics of a series. Several available approaches for the pointwise OIRFs estimation could be used for our task (see [14, 22]). In this paper, we develop the above-presented idea in the important case of vector autoregressive (VAR) models with constant AR parameters but with time-varying covariance structure. Indeed, it is often admitted that the conditional mean is constant, while the variance is time-varying (see [4, 9, 15, 20, 25, 26], among others). In addition, it is widely known that a non-constant variance is common for economic variables. For instance, [24] found that more than 80% of the 214 U.S. economic variables they studied have a non-constant variance (see [2] for break detection in the covariance structure). For this reason, the series under study are assumed non-stationary, due to the non-constant variance (i.e., the heteroscedasticity is unconditional).

To illustrate our main idea, let us consider the univariate framework of [30], where the conditional mean is filtered by fitting an AR model. Using this simple framework, we indicate the ways of summarizing the time-varying response functions to a *rescaled*[1] impulse for an univariate series. Let us define by $\sigma_t^2 = g^2(t/T)$, the (unobserved) innovations variance at time $1 \le t \le T$, where $g(\cdot)$ is a function fulfilling some regularity conditions. As the (unobserved) MA coefficients ϕ_i are constant in our case, it suffices to focus on the changes in the variance to capture the evolution of the rescaled impulse response functions (IRF) $\phi_i \sigma_{t-i}$. The usual way to summarize the time-varying IRF over a given period would be to estimate the standard tool that assumes a constant variance. This would lead to estimating $\phi_i (\int_0^1 g^2(r)dr)^{\frac{1}{2}}$ (which will be called the approximate IRF), whereas $\phi_i \int_0^1 g(r)dr$ (which will be called the averaged IRF) is more sound to summarize the IRF. Here, the integrals account for the averaging over given periods of interest. Indeed, if the purpose is to find a summary of the IRF, averaging over the values of fixed periods seems more reasonable than considering a kind of norm such as $(\int_0^1 g^2(r)dr)^{\frac{1}{2}}$. Clearly, the averaged and approximated IRFs are different in general, as long as the variance structure is non-constant. More precisely, the more the variance varies over time, the larger the discrepancy between the averaged and the approximated IRFs. Therefore, in the following, we also propose to build an indicator of the variability of the variance based on the discrepancy between the averaged and approximated IRFs. It is important to underline that robustness/stability studies often rely on the simple (graphical) examination of the different OIRFs. As this way of proceeding is subjective, our indicator is intended to quantify such a kind of analysis.

The rest of this paper is organized as follows. In Sect. 18.2, the VAR model with unconditionally heteroscedastic innovations is presented. Next, different possible concepts of OIRFs that could be considered in our framework are discussed. Moreover, we introduce a scalar variance variability index that measures the departure from the standard constant variance VAR setup. Section 18.3 proposes the estima-

[1] The term *rescaled* is taken from [16, p. 53].

tors and studies their asymptotic properties. The time-varying OIRFs estimator is introduced and its nonparametric rate of convergence is derived. In Sects. 18.3.2 and 18.3.3, the estimators of the approximated and averaged OIRFs are defined. Their asymptotic behavior are also studied. Assumptions and some proofs are given in Sect. 18.5.

18.2 Time-Varying Orthogonal Impulse Response Functions

Following the usual approach for impulse response analysis between variables, consider a VAR model for the series $X_t \in \mathbb{R}^d$:

$$X_t = A_{01}X_{t-1} + \cdots + A_{0p}X_{t-p} + u_t, \tag{18.1}$$

where u_t is the error term and the A_{0i}'s are the AR parameter matrices, such that $\det A(z) \neq 0$ for all $|z| \leq 1$, with $A(z) = I_d - \sum_{i=1}^{p} A_{0i}z^i$. The matrix I_d is the $d \times d$-identity matrix. Here, the covariance of the system is allowed to vary in time. More precisely, the covariance of the process (u_t) is denoted by $\Sigma_t := G(t/T)G(t/T)'$, where $r \mapsto G(r)$, $r \in (0, 1]$, is a $d \times d$-matrix valued function. See Assumption 18.2, given in Sect. 18.5.2. With the rescaling device used by [10], the process (X_t) should be formally written in a triangular form. Herein, instead of writing $X_{t,T}$ and $\Sigma_{t,T}$ (the double subscript), the notation X_t and Σ_t (without the subscript T) is used for simplicity.

The specification we consider allows for commonly observed features as cycles, smooth or abrupt changes for the covariance, and is widely used in the literature (see, e.g., [30] and references therein). In particular, the rescaling device is commonly used to describe long-range phenomena (see [7, 8], among others). In practice, the order p in (18.1) is unknown but can be fixed using the tools proposed in [21, 23] under our assumptions.

Let $\widetilde{X}_t = (X_t', \ldots, X_{t-p+1}')'$. The usual Kronecker product is denoted by $\otimes$. In the following, the model (18.1) is rewritten as follows:

$$X_t = (\widetilde{X}_{t-1}' \otimes I_d)\vartheta_0 + u_t,$$
$$u_t = H_t\epsilon_t,$$

where (ϵ_t) is an iid centered process with $E(\epsilon_t\epsilon_t') = I_d$, and

$$\vartheta_0 = \text{vec}(A_{01}, \ldots, A_{0p}),$$

is the vector of parameters. Herein the vec$(\cdot)$ operator consists of stacking the columns of a matrix into a vector. The matrix H_t is the lower triangular matrix of the Cholesky decomposition of the covariance of the errors, that is, $\Sigma_t = H_tH_t'$. We also define

$$\Phi_0 = I_d, \quad \Phi_i = \sum_{j=1}^{i} \Phi_{i-j} A_{0j}, \tag{18.2}$$

$i = 1, \ldots,$ with $A_{0j} = 0$ for $j > p$. The Φ_i's correspond to the coefficients matrices of the infinite MA representation of (X_t). Under our assumptions the components of the Φ_i's decrease exponentially fast to zero.

If the errors' covariance Σ is assumed constant, then we can define the $d \times d$ matrix of standard OIRFs

$$\theta(i) := \Phi_i H, \qquad i = 1, 2, \ldots,$$

where H is the lower triangular matrix of the Cholesky decomposition of Σ. See [16, p. 59]. Let us denote by $\widehat{\vartheta}_{OLS}$ the ordinary least-squares (OLS) estimator of the AR parameters ϑ_0 and define $\widehat{\Sigma}$, the OLS estimator of the constant errors covariance matrix. Using $\widehat{\vartheta}_{OLS}$ and $\widehat{\Sigma}$, it is easy to see that an estimator of $\theta(i)$ can be built. Under the standard assumptions, it can be shown that such estimators are consistent, $\sqrt{T}$-asymptotically Gaussian. See [16, p. 110]. However, it clearly appears that the classical OIRFs cannot take the time-varying instantaneous effects into account properly, and may be misleading in our non-standard but quite realistic framework.

18.2.1 tv-OIRFs

In the framework of the model (18.1) a common alternative to the classical OIRFs is the time-varying OIRFs (tv-OIRFs hereafter)

$$\theta_r(i) := \Phi_i H(r), \qquad i \geq 1, \tag{18.3}$$

for each $r \in (0, 1]$, where $H(r)$ is the lower triangular matrix of the Cholesky decomposition of $\Sigma(r) = G(r)G(r)'$. The parameter r gives the time over which the impulse response analysis is conducted. In other words, the counterpart of the usual OIRFs in the case of time-varying variance is the two arguments function

$$(r, i) \mapsto \theta_r(i), \qquad (r, i] \in (0, 1] \times \{1, 2, \ldots\}.$$

The form (18.3) implicitly arises when models with constant AR parameters but time-varying variance are used to analyze the data (see [5, 26, 30] among others, for this kind of model). When the covariance of the errors is constant, for each i, the mapping $r \mapsto \theta_r(i)$ is constant, and thus we retrieve the standard case. Although it is interesting to have a pointwise estimator of the OIRFs, in general these mappings are not constant and are typically estimated at nonparametric rates, as will be shown in the following. The papers [14, 22] provided complete tools for estimating (18.3) in general contexts. As a byproduct of our results, we specify the methodological pathway for the pointwise estimation of the OIRFs in the important case where the conditional mean is constant but the variance is time-varying.

A summary of the tv-OIRFs over time could sometimes be needed to compare fixed periods. In many cases, this consists of evaluating the differences between pre- and post-crises situations. In the following, we consider two approaches for summarizing the tv-OIRFs over a given sub-period. First, we replace the matrix $H(r)$ in (18.3) by the lower triangular matrix of the Cholesky decomposition of the *realized variance*, that is the average of the variance, over a given period around r. This will yield what we call the *approximated OIRFs*. Typically, this corresponds to the usual practice which consists in applying the standard method to periods (see, e.g., [3, 27]). Second, keeping in mind that we are looking for a summary of the tv-OIRFs, which is tantamount to looking for a summary of integrated $H(r)$ appearing in (18.3), we introduce the *averaged OIRFs* that is obtained by replacing the matrix $H(r)$ with the average of the lower triangular matrix of the Cholesky decomposition of $\Sigma(\cdot)$ over a given period around r. Both summaries we consider could be estimated at parametric rates and, considering static or rolling periods, could be used for an analysis of the series. However, as argued in Introduction, the averaged OIRFs should be preferred. Before presenting the approximated and averaged approaches, we here point out that, as usual, summarizing the OIRFs does not make the shocks orthogonal pointwise. Note however that such a property is not really needed if we are interested in comparing periods by considering means.

18.2.2 Approximated OIRFs

The usual way to summarize the OIRFs in presence of a non-constant covariance in our framework is to consider the quantities

$$\widetilde{\theta}_r^q(i) = \Phi_i \widetilde{H}(r), \qquad i \geq 1, \tag{18.4}$$

where $\widetilde{H}(r)$ is the lower triangular matrix of the Cholesky decomposition of the positive definite matrix $q^{-1}\int_{r-q/2}^{r+q/2}\Sigma(v)dv$ with $0 < r - q/2 < r + q/2 < 1$. Again the standard case is retrieved if the covariance structure is assumed constant. If r does not correspond to a covariance break, we have $\widetilde{\theta}_r^q(i) \approx \theta_r(i)$ for small enough q. However, as the periods under study are usually somewhat large, we are not aiming in reflecting the evolution of $H(\cdot)$, so we refer (18.4) to the approximate OIRFs in the following. In short, the approximated OIRFs are usually computed to contrast between static periods. For fixed r and q, the quantities $\widetilde{\theta}_r^q(i)$ could be estimated at parametric rates.

18.2.3 Averaged OIRFs

As argued above, by construction, the approximated OIRFs could be misleading in summarizing the time-varying $\theta_r(i)$ over a period. Given the definition of $\theta_r(i)$, a more natural way to approach it would be to average the lower triangular matrix of the Cholesky decomposition over a window around r. We propose a new alternative way to summarize the tv-OIRFs (18.3) based on the quantities

$$\bar{\theta}_r^q(i) := \Phi_i \bar{H}(r) \text{ where } \bar{H}(r) := \frac{1}{q}\int_{r-q/2}^{r+q/2} H(v)dv, \qquad i \geq 1,$$

where $0 < q < 1$ is fixed by the practitioner, and r is such that $0 < r - q/2 < r + q/2 < 1$. The standard case is retrieved if the errors covariance is constant. On the other hand, if r does not correspond to an abrupt break in the covariance structure, we clearly have $\bar{\theta}_r^q(i) \approx \theta_r(i)$ when q is small. However, as noted above, the averaged OIRFs is intended to be applied for a relatively large q.

18.2.4 Variance Variability Indices

In this section we propose an index, that is a scalar, to measure the departure from a constant covariance matrix situation within a given period. We can write

$$\widetilde{\theta}_r^q(i) = \bar{\theta}_r^q(i) I_{r,q}$$

with the $d \times d$-matrix $I_{r,q}$ defined as

$$I_{r,q} = \bar{H}(r)^{-1}\widetilde{H}(r).$$

Let us define

$$i_{r,q} = \left\| I_{r,q} \right\|_2^2,$$

where $\|\cdot\|_2$ denotes the spectral norm of a matrix. In this case, $i_{r,q}$ is equal to the square of the largest eigenvalue of $I_{r,q}$, which has only real, positive eigenvalues. By elementary matrix algebra properties, we also have

$$i_{r,q} = \max_{a \in \mathbb{R}^d, a \neq 0} \frac{Var(a' X_{app})}{Var(a' X_{avg})},$$

where X_{avg} and X_{app} are d-dimensional random vectors with variances $\bar{H}(r)\bar{H}(r)'$ and $\widetilde{H}(r)\widetilde{H}(r)'$, respectively. In the statistical literature, a quantity like $i_{r,q}$ is usually called the first relative eigenvalue of one matrix (here $\widetilde{H}(r)\widetilde{H}(r)'$) with respect to the other matrix (here $\bar{H}(r)\bar{H}(r)'$). See, for instance, [13]. By construction, in our context, the eigenvalues of the matrix $\bar{H}(r)^{-1}\widetilde{H}(r)$ are real numbers greater than or equal to 1, as shown in the following.

The index $i_{r,q}$ is inspired by the OIRFs analysis. It is designed to provide a measure of variability through the contrast between two possible definitions of OIRFs that coincide, in the case of a covariance Σ constant over time. Another simple index could be defined as

$$j_{r,q} = \left\| \frac{1}{q} \int_{r-q/2}^{r+q/2} \Sigma(v)dv - \bar{H}(r)\bar{H}(r)' \right\|_2^2 .$$

By elementary properties of the spectral norm,

$$j_{r,q} = \max_{a \in \mathbb{R}^d, \|a\|=1} \{Var(a'X_{app}) - Var(a'X_{avg})\}.$$

Lemma 18.1 *Under Assumption 18.2, given in Sect. 18.5.2, it holds that*

1. $i_{r,q} \geq 1$ *and* $i_{r,q} = 1$ *if and only if* $v \mapsto H(v)$ *is constant over* $(r - q/2, r + q/2)$*;*
2. $j_{r,q} \geq 0$ *and* $j_{r,q} = 0$ *if and only if* $v \mapsto H(v)$ *is constant over* $(r - q/2, r + q/2)$.

In our context, for any $0 < q < 1$, a mapping $r \mapsto i_{r,q}$ (resp. $r \mapsto j_{r,q}$) constant equal to 1 (resp. 0) means the covariance of the error (u_t) is constant in time. For simplicity, in the following we will focus on the index $i_{r,q}$ which is invariant to multiplication of the errors' covariance matrix by a positive constant. Large values of $i_{r,q}$ indicate a large variability in the variance of the vector series over a given period of interest.

18.3 OIRFs Estimators When the Variance is Varying

Let us first briefly recall the estimation methodology for heteroscedastic VAR models of [20, Sect. 4]. We first consider the OLS estimator of the AR parameters

$$\widehat{\vartheta}_{OLS} = \left\{ \sum_{t=1}^{T} \widetilde{X}_{t-1}\widetilde{X}'_{t-1} \otimes I_d \right\}^{-1} \text{vec}\left(\sum_{t=1}^{T} X_t \widetilde{X}'_{t-1} \right). \tag{18.5}$$

The paper [20] showed that

$$\sqrt{T}(\widehat{\vartheta}_{OLS} - \vartheta_0) \Rightarrow \mathcal{N}(0, \Lambda_3^{-1}\Lambda_2\Lambda_3^{-1}), \tag{18.6}$$

where

$$\Lambda_2 = \int_0^1 \sum_{i=0}^{\infty} \left\{ \tilde{\Phi}_i (\mathbf{1}_{p\times p} \otimes \Sigma(r)) \tilde{\Phi}_i' \right\} \otimes \Sigma(r) dr,$$

$$\Lambda_3 = \int_0^1 \sum_{i=0}^{\infty} \left\{ \tilde{\Phi}_i (\mathbf{1}_{p\times p} \otimes \Sigma(r)) \tilde{\Phi}_i' \right\} \otimes I_d \, dr \tag{18.7}$$

$\mathbf{1}_{p\times p}$ is the $p \times p$ matrix with components equal to one, and $\widetilde{\Phi}_i$ is a block diagonal matrix $\widetilde{\Phi}_i := diag\left(\Phi_i, \Phi_{i-1}, \ldots, \Phi_{i-p+1}\right)$. The matrices Φ_i, $i \geq 0$, are defined in (18.2), and $\Phi_i = 0$ for $i < 0$.

Next, let us consider kernel estimators of the time-varying covariance matrix. Denote by $A \odot B$, the Hadamard (entrywise) product of two matrices A and B of the same dimension. For $t = 1, \ldots, T$, define the symmetric matrices

$$\widehat{\Sigma}_t = \sum_{j=1}^{T} w_{tj} \odot \widehat{u}_j \widehat{u}_j', \tag{18.8}$$

where $\widehat{u}_t = X_t - (\widetilde{X}_{t-1}' \otimes I_d)\widehat{\vartheta}_{OLS}$ is the OLS residual vector. The (k, l)-element, $k \leq l$, of the $d \times d$ matrix of weights w_{tj} is given by

$$w_{tj}(b_{kl}) = (T b_{kl})^{-1} K\left((t-j)/(T b_{kl})\right),$$

where b_{kl} is the bandwidth and $K(\cdot)$ is nonnegative. For any $r \in (0, 1]$, the value $\Sigma(r)$ of the covariance function could be estimated by $\widehat{\Sigma}_{[rT]}$. Here and in the following, for a number a, we denote by $[a]$ the integer part of a, that is the largest integer number less than or equal to a. For all $1 \leq k \leq l \leq d$, the bandwidth b_{kl} belongs to a range $\mathcal{B}_T = [\underline{c} b_T, \bar{c} b_T]$, where $\underline{c}, \bar{c} > 0$ are some constants and $b_T \downarrow 0$ at a suitable rate specified below. In practice the bandwidths b_{kl} can be chosen by minimization of a cross-validation criterion. This estimator is a version of the Nadaraya–Watson estimator considered by [20]. Here, we replace the denominator by the target density, that is the uniform density on the unit interval which is constant equal to 1. A regularization term may be needed to ensure that the matrices $\widehat{\Sigma}_t$ are positive definite (see [20]). Another simple way to circumvent the problem is to select a unique bandwidth $b = b_{kl}$, for all $1 \leq k, l \leq d$.

With an estimator of $\Sigma(r)$ at hand, we could define $\widehat{H}_{[rT]}$, the lower triangular matrix of the Cholesky decomposition of $\widehat{\Sigma}_{[rT]}$, as the estimator of $H(r)$. We establish the convergence rates of these nonparametric estimators below. For $r \in (0, 1)$, let $\Sigma(r-) = \lim_{\tilde{r}\uparrow r} \Sigma(\tilde{r})$ and $\Sigma(r+) = \lim_{\tilde{r}\downarrow r} \Sigma(\tilde{r})$. Moreover, by definition, let $\Sigma(1+) = 0$. Let $H(r-)$ and $H(r+)$ be defined similarly. In the following, $\|\cdot\|_F$ is the Frobenius norm, while $\sup_{\mathcal{B}_T}$ denotes the supremum with respect the bandwidths b_{kl} in $\mathcal{B}_T$.

Proposition 18.1 *Assume that Assumptions 18.1–18.3, given in Sect. 18.5.2 hold. Then, for any* $r \in (0, 1]$,

$$\sup_{\mathcal{B}_T} \left\| \widehat{\Sigma}_{[Tr]} - \frac{1}{2} \{\Sigma(r-) + \Sigma(r+)\} \right\|_F = O_{\mathbb{P}} \left(b_T + \sqrt{\log(T)/Tb_T} \right)$$

and

$$\sup_{\mathcal{B}_T} \left\| \widehat{H}_{[Tr]} - \frac{1}{2} H_{\pm}(r) \right\|_F = O_{\mathbb{P}} \left(b_T + \sqrt{\log(T)/Tb_T} \right),$$

where $H_{\pm}(r)$ *is the lower triangular matrix of the Cholesky decomposition of* $\Sigma(r-) + \Sigma(r+)$.

The convergence rate of $\widehat{\Sigma}_{[Tr]}$ and $\widehat{H}_{[Tr]}$ is given by a bias term, with the standard rate one could obtain when estimating Lipschitz continuous functions nonparametrically, and a variance term which is multiplied by a logarithm factor, the price to pay for the uniformity with respect to the bandwidth.

The above estimation of the non-constant covariance structure could be used to define the adaptive least squares (ALS) estimator

$$\widehat{\vartheta}_{ALS} = \widetilde{\Sigma}_{\underline{\widetilde{X}}}^{-1} \text{vec} \left(\widetilde{\Sigma}_{\underline{X}} \right), \tag{18.9}$$

where

$$\widetilde{\Sigma}_{\underline{\widetilde{X}}} = T^{-1} \sum_{t=1}^{T} \widetilde{X}_{t-1} \widetilde{X}'_{t-1} \otimes \widehat{\Sigma}_t^{-1} \quad \text{and} \quad \widetilde{\Sigma}_{\underline{X}} = T^{-1} \sum_{t=1}^{T} \widehat{\Sigma}_t^{-1} X_t \widetilde{X}'_{t-1}.$$

By minor adaptation of the proofs in [20], in order to take into account the simplified change in the definition of the weights w_{tj}, it could be shown that, uniformly with respect to $b \in \mathcal{B}_T$, $\widehat{\vartheta}_{ALS}$ is consistent in probability and

$$\sqrt{T}(\widehat{\vartheta}_{ALS} - \vartheta_0) \Rightarrow \mathcal{N}(0, \Lambda_1^{-1}),$$

where

$$\Lambda_1 = \int_0^1 \sum_{i=0}^{\infty} \left\{ \widetilde{\Phi}_i (\mathbf{1}_{p\times p} \otimes \Sigma(r)) \widetilde{\Phi}'_i \right\} \otimes \Sigma(r)^{-1} dr.$$

The paper [20] showed that $\Lambda_3^{-1} \Lambda_2 \Lambda_3^{-1} - \Lambda_1$ is a positive semi-definite matrix.

18.3.1 The tv-OIRFs Nonparametric Estimator

In the context of model (18.1), the natural way to build estimators of the time-varying OIRFs defined in (18.3) is to construct plug-in estimators of the Φ_i and $H(r)$. For estimating Φ_i we use $\widehat{\Phi}_i^{als}$ which are obtained as in (18.2), but considering the ALS estimator of the A_{0i}'s. By the arguments used in the proof of Proposition 18.2 below, this estimator has the $O_{\mathbb{P}}(1/\sqrt{T})$ rate of convergence. Using the nonparametric estimator of $H(r)$ we introduced above, we obtain what we call the *ALS estimator of* $\theta_r(i)$, that is,

$$\widehat{\theta}_r(i) := \widehat{\Phi}_i^{als}\widehat{H}_{[rT]}, \qquad r \in (0, 1].$$

Even if $\widehat{\Phi}_i^{als}$ has an improved variance compared to the estimator where the OLS estimator of the A_{0i}'s is used, the estimator $\widehat{\theta}_r(i)$ still inherits the nonparametric rate of the convergence of $\widehat{H}_{[rT]}$ described in Proposition 18.1. Hence, analyzing the variations of the estimated curves $r \mapsto \widehat{\theta}_r(i)$, for various i, suffers from lower, nonparametric convergence rates. In Sect. 18.3.3, we propose to use instead of $\widehat{\theta}_r(i)$ averages over the values in a neighborhood of r, that is a window containing r. In particular, this allows us to recover parametric rate of convergence of the estimators. In practice, this interval corresponds to some period of interest.

18.3.2 Approximated OIRFs Estimators

The results of this part are only stated as they are direct consequences of arguments in [20] and standard techniques (see [16]). Let the usual estimator of (18.4),

$$\widehat{\widetilde{\theta}}_r^{q}(i) := \widehat{\Phi}_i^{ols}\widehat{\widetilde{H}}(r), \tag{18.10}$$

where $\widehat{\widetilde{H}}(r)$ is the lower triangular matrices of the Cholesky decomposition of

$$\widehat{S}_T(r) = \frac{1}{[qT]+1}\sum_{k=-[qT/2]}^{[qT/2]} \widehat{u}_{[rT]-k}\widehat{u}'_{[rT]-k}.$$

Here, $\widehat{u}_{[rT]-k}$ is the OLS residual vector and $\widehat{\Phi}_i^{ols}$'s are the estimators of the MA coefficients obtained from the OLS estimators of the autoregressive parameters. Recall that (18.10) is used to evaluate the OIRFs in the standard homoscedastic case (see [16, Sect. 3.7]), but is also commonly considered to evaluate tv-OIRFs over static periods. The expression (18.10) is suitable at least asymptotically, since, as in the proof of Lemma 18.2 below, it is shown that

$$\begin{aligned}
&\frac{1}{[qT]+1}\sum_{k=-[qT/2]}^{[qT/2]} \widehat{u}_{[rT]-k}\widehat{u}'_{[rT]-k}\\
&\qquad= \frac{1}{[qT]+1}\sum_{k=-[qT/2]}^{[qT/2]} u_{[rT]-k}u'_{[rT]-k} + o_{\mathbb{P}}(1/\sqrt{T})\\
&\qquad\qquad= \frac{1}{q}\int_{r-q/2}^{r+q/2}\Sigma(v)dv + O_{\mathbb{P}}(1/\sqrt{T}).
\end{aligned}$$

In order to specify the asymptotic behavior of $\widetilde{\widehat{\theta}}{}_r^q(i)$, we first state a result which can be proved using similar arguments to those in [18, Lemma 7.4]. Let $\widehat{\zeta}_t := \text{vech}\left(\widehat{u}_t\widehat{u}_t'\right)$, $\zeta_t := \text{vech}\left(u_t u_t'\right)$ and $\Gamma_t := \text{vech}(\Sigma(t/T)) = \text{vech}(\Sigma_t)$, where the vech operator consists of stacking the elements on and below the main diagonal of a square matrix. Define

$$\overline{\Gamma}(r) := \text{vech}\left(q^{-1}\int_{r-q/2}^{r+q/2} \Sigma(v)dv\right) \quad \text{and} \quad \widehat{\overline{\Gamma}}(r) := ([qT]+1)^{-1}\sum_{k=-[qT/2]}^{[qT/2]} \widehat{\zeta}_{[rT]-k},$$

for $r < 1$. Introduce the functions $\Gamma(\cdot)$ and $\Delta(\cdot)$ given by $\Gamma(\cdot) = \text{vech}(\Sigma(\cdot))$ and $\Delta(t/T) = E(\zeta_t\zeta_t')$.

Lemma 18.2 *Under Assumptions 18.1–18.3, given in Sect. 18.5.2 hold, we have*

$$\sqrt{T}\begin{pmatrix} \widehat{\vartheta}_{OLS} - \vartheta_0 \\ \widehat{\overline{\Gamma}}(r) - \overline{\Gamma}(r) \end{pmatrix} \Rightarrow \mathcal{N}\left(0, \begin{pmatrix} \Lambda_3^{-1}\Lambda_2\Lambda_3^{-1} & 0 \\ 0 & \Omega(r) \end{pmatrix}\right), \tag{18.11}$$

where $\widehat{\vartheta}_{OLS}$, and the Λ_i's are defined in (18.7) and (18.5) respectively, and

$$\Omega(r) = \frac{1}{q}\int_{r-q/2}^{r+q/2} \{\Delta(v) - \Gamma(v)\Gamma(v)'\}dv.$$

Now, define the commutation matrix K_d such that $K_d\text{vec}(G) = \text{vec}(G')$, and the elimination matrix L_d such that $\text{vech}(G) = L_d\text{vec}(G)$ for any square matrix G of dimension $d \times d$. Introduce the $pd \times pd$ matrix

$$\mathbb{A} = \begin{pmatrix} A_{01} & \dots & \dots & A_{0p} \\ I_d & 0 & \dots & 0 \\ 0 & \ddots & 0 & \vdots \\ 0 & 0 & I_d & 0 \end{pmatrix} \tag{18.12}$$

and the $d \times pd$ matrix $J = (I_d, 0, \dots, 0)$. We are now in a position to state the asymptotic behavior of the classical approximated OIRFs estimator. Note that this result can be obtained using the same arguments of [16, Proposition 3.6], together with (18.11).

Proposition 18.2 *Under Assumptions 18.1 and 18.2, given in Sect. 18.5.2, we have, for all $r \in (q/2, 1-q/2)$ and as $T \to \infty$,*

$$\sqrt{T}vec\left(\widetilde{\widehat{\theta}}{}_r^q(i) - \widetilde{\theta}_r^q(i)\right) \Rightarrow \mathcal{N}\left(0, C_i(r)\Lambda_3^{-1}\Lambda_2\Lambda_3^{-1}C_i(r)' + D_i(r)\Omega(r)D_i(r)'\right), \tag{18.13}$$

$i = 0, 1, 2, \dots$, where $C_0 = 0$, $C_i(r) = \left(\widetilde{H}(r)' \otimes I_d\right)\left(\sum_{m=0}^{i-1} J(\mathbb{A}')^{i-1-m} \otimes \Phi_m\right)$, $i = 1, 2, \dots$, $\widetilde{H}(r)$ is given in (18.4), and

$$D_i(r) = (I_d \otimes \Phi_i)\,\Xi(r),\ i = 0, 1, 2, \ldots$$

with

$$\Xi(r) = L_d'\left[L_d\left(I_{d^2} + K_d\right)\left(\widetilde{H}(r) \otimes I_d\right)L_d'\right]^{-1}.$$

We propose an alternative approximated OIRFs estimator based on the more efficient estimator $\widehat{\vartheta}_{ALS}$ defined in (18.9) and the estimators $\widehat{\Phi}_i^{als}$ of the coefficients $\widehat{\Phi}_i$ of the infinite MA representation of (X_t). More precisely,

$$\widehat{\widetilde{\theta}}_r^{q,als}(i) := \widehat{\Phi}_i^{als}\,\widehat{\widetilde{H}}(r)$$

is a new approximated OIRFs estimator. We state its asymptotic distribution.

Proposition 18.3 *In addition to the conditions of Proposition 18.2, assume that Assumption 18.3, given in Sect. 18.5.2 holds. With the notation defined in Proposition 18.2, we have, for all $r \in (q/2, 1-q/2)$ and as $T \to \infty$,*

$$\sqrt{T}vec\left(\widehat{\widetilde{\theta}}_r^{q,als}(i) - \widetilde{\theta}_r^q(i)\right) \Rightarrow \mathcal{N}\left(0, C_i(r)\Lambda_1^{-1}C_i(r)' + D_i(r)\Omega(r)D_i(r)'\right), \tag{18.14}$$

$i = 0, 1, 2, \ldots$. *Moreover, the difference between the asymptotic variance given in (18.14) and the asymptotic variance of* $vec\left(\widehat{\widetilde{\theta}}_r^{q,als}(i) - \widetilde{\theta}_r^q(i)\right)$ *is a positive semi-definite matrix.*

The proof of Proposition 18.3 is omitted since it follows the steps of the proof of Proposition 18.2, and uses the results of [20] on the convergence in law of $\widehat{\vartheta}_{ALS}$. In particular, they proved that $\Lambda_3^{-1}\Lambda_2\Lambda_3^{-1} - \Lambda_1^{-1}$ is a positive semi-definite matrix and this implies that $\widehat{\widetilde{\theta}}_r^{q,als}(i)$ is a lower variance estimator of $\widetilde{\theta}_r^q(i)$.

Although the standard $\widehat{\widetilde{\theta}}_r^q(i)$ and more efficient $\widehat{\widetilde{\theta}}_r^{q,als}(i)$ estimators are easy to compute, for the reasons we detailed above, we believe that they are not appropriate tools to summarize the evolution of the tv-OIRFs (18.3). Instead, we propose to use an estimator of the averaged OIRFs. To build such an estimate of the averaged OIRFs with negligible bias, we need a slightly modified kernel estimator of $\Sigma(\cdot)$ that we introduce in the next section.

18.3.3 New OIRFs Estimators With tv-Variance

In this section, we propose an alternative estimator for the approximated OIRFs and an estimator for the averaged OIRFs we introduced in Sect. 18.2.3. To guarantee $\sqrt{T}$-asymptotic normality for these estimators, we implicitly need suitable estimators of integral functionals in the form

$$\frac{1}{q}\int_{r-q/2}^{r+q/2} A(v)\Sigma(v)dv,$$

where $A(\cdot)$ is a given matrix-valued function. The estimator of such an integral, obtained by plugging in the nonparametric estimator of the covariance structure introduced in (18.8), would not be appropriate as it suffers from boundary effects. More details on this problem are provided in Sect. 18.5.1. Therefore, in the following, we construct alternative bias corrected estimators for such integral functionals.

For $-[(q+h)T/2] \le k \le [(q+h)T/2]$, we define

$$\widehat{V}_{[rT]-k} = \frac{1}{T}\sum_{j=[(r-(q-h)/2)T]+1}^{[(r+(q-h)/2)T]} \frac{1}{h} L\left(\frac{[rT]-k-j}{hT}\right)\widehat{u}_j\widehat{u}_j'. \tag{18.15}$$

Hereafter, for simplicity, we use the same bandwidth h for all the d^2 components of the estimated matrix-valued integrals. Note that $\widehat{V}_{[rT]-k}$ is an estimator of $\Sigma_{[rT]-k}$. Next, let $\widehat{H}_{[rT]-k}$ denote the lower triangular matrix of the Cholesky decomposition of $\widehat{V}_{[rT]-k}$, that is,

$$\widehat{V}_{[rT]-k} = \widehat{H}_{[rT]-k}\widehat{H}_{[rT]-k}'.$$

We propose the following adaptive least-squares estimators of the tv-averaged OIRFs:

$$\hat{\bar{\theta}}_r^q(i) = \widehat{\Phi}_i^{als}\ \widehat{\bar{H}}(r),$$

where

$$\widehat{\bar{H}}(r) = \frac{1}{[qT]+1}\sum_{k=-[(q+h)T/2]}^{[(q+h)T/2]} \widehat{H}_{[rT]-k}. \tag{18.16}$$

Proposition 18.4 *If Assumptions 18.1–18.3, given in Sect. 18.5.2, hold, then for all $r \in (q/2, 1-q/2)$ and as $T \to \infty$,*

$$\sqrt{T}vec\left(\hat{\bar{\theta}}_r^q(i) - \bar{\theta}_r^q(i)\right) \Rightarrow \mathcal{N}(0, \overline{C}_i(r)\Lambda_1^{-1}\overline{C}_i(r)' + D_i(r)\Omega(r)D_i(r)'),\quad i = 0, 1, 2, \ldots,$$

where $D_i(r)$ and $\Omega(r)$ are defined in Proposition 18.2 and

$$\overline{C}_i(r) = \left(\frac{1}{q}\int_{r-q/2}^{r+q/2} H(v)'dv \otimes I_d\right)\left(\sum_{m=0}^{i-1} J(\mathbb{A}')^{i-1-m} \otimes \Phi_m\right).$$

18.4 Conclusion

In the context of the VAR modeling, a possible way to measure changes in the relationships between economic variables is to compare the tv-dynamics of the series between periods of interest. For instance, before and after crises or regulatory change periods can be adjusted by practitioners (see, e.g., [17, 28]). A possible way to perform such a comparison is to compute the classical OIRFs within each period (the *approximated* OIRFs). Although the above approach is commonly used because of its easiness, our theoretical study indicates that doing so leads to an erroneous picture of the evolving economic variables dynamics. Indeed, as soon as the OIRFs exhibit marked changes, the classical OIRFs can be a poor guide to locate the general level of response to an impulse for a given period. In order to correct such potentially misleading analyses, we have proposed to first estimate the tv-OIRFs pointwise, and then considered the mean of these estimators (the *averaged* OIRFs). It is found that this alternative tool delivers a more accurate picture from the classical approach when rapid shifts are observed. In view of the above, an index and a formal statistical test for delineating the limits of usefulness between the more sophisticated averaged OIRFs and the simple approximated OIRFs have been proposed. Finally, let us mention that empirical illustrations with simulated and real data, as well as some additional technical arguments, are available in [19].

18.5 Technical Details

18.5.1 Kernel Estimators of the Covariance Function Integrals

As mentioned in Sect. 18.3.3, we need suitable estimators of the integral functionals

$$\frac{1}{q}\int_{r-q/2}^{r+q/2} A(v)\Sigma(v)dv,$$

where $A(\cdot)$ is a given matrix-valued function. The estimator of such integrals obtained by plugging in the nonparametric estimator of the covariance structure introduced in (18.8) would be asymptotically biased due to the boundary effects.

To explain the rationale of the alternative nonparametric estimator, we propose we assume for the moment that the $d \times d$-matrices $u_j u_j'$, $1 \leq j \leq T$, are available. Let us consider the generic real-valued random quantity

$$S_T(r) = \int_{r-q/2}^{r+q/2} a(v) \left[\frac{1}{|\mathcal{J}|} \sum_{j \in \mathcal{J}} \omega_{v,j}(h) u_j^{(k)} u_j^{(l)} \right] dv,$$

where

- $\mathcal{J} = \{j_{min}, \dots, j_{max}\} \subset \{1, \dots, T\}$ is a set of consecutive indices that will be specified below and $|\mathcal{J}|$ is the cardinal of $\mathcal{J}$;
- $\omega_{v,j}(h) = h^{-1}L(h^{-1}(v - j/T))$ where h is a deterministic bandwidth with a rate that will be specified below;
- $L(\cdot)$ is a bounded symmetric density function with support $[-1, 1]$;
- $a(\cdot)$ is a differentiable function with Lipschitz continuous derivative.

Here, $u_j^{(k)}$ and $u_j^{(l)}$ are components of u_j and $E(u_j^{(k)}u_j^{(l)}) = \Sigma^{(k,l)}(j/T)$, that is the (k,l) cell of the matrix $\Sigma(j/T)$.

By a change of variables and the Taylor expansion,

$$\int_{r-q/2}^{r+q/2} a(v)\frac{1}{h}L\left(\frac{v - j/T}{h}\right)dv = \int_{(r-q/2-j/T)/h}^{(r+q/2-j/T)/h} a(j/T + uh)L(u)\,du$$
$$= a(j/T)\int_{(r-q/2-j/T)/h}^{(r+q/2-j/T)/h} L(u)\,du + ha'(j/T)\int_{(r-q/2-j/T)/h}^{(r+q/2-j/T)/h} uL(u)\,du + O(h^2).$$

To avoid a large bias, we aim at using the properties $\int_{-1}^{1} L(u)du = 1$ and $\int_{-1}^{1} uL(u)du = 0$. For this purpose, any $j \in \mathcal{J}$ should satisfy the conditions $(r + q/2 - j/T)/h \geq 1$ and $(r - q/2 - j/T)/h \leq -1$. That is, the indices set $\mathcal{J}$ should be defined such that

$$\forall j \in \mathcal{J}, \quad (r - q/2 + h)T \leq j \leq (r + q/2 - h)T.$$

Let us define

$$j_{min} = [(r - q/2 + h)T] + 1 \qquad \text{and} \qquad j_{max} = [(r + q/2 - h)T].$$

Then, uniformly with respect to $j \in \mathcal{J}$,

$$\int_{r-q/2}^{r+q/2} a(v)\frac{1}{h}L\left(\frac{v - j/T}{h}\right)dv - a(j/T) = O(h^2).$$

Note that $|\mathcal{J}| = [(q - 2h)T]$ and $|\mathcal{J}|/(q - 2h)T = 1 + O(1/T)$.

We can now deduce

$$\begin{aligned} S_T &= \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}} a(j/T)\{u_j^{(k)}u_j^{(l)} - \Sigma^{(k,l)}(j/T)\} \\ &+ \frac{1}{|\mathcal{J}|}\sum_{j\in\mathcal{J}} a(j/T)\Sigma^{(k,l)}(j/T) + O(h^2) \\ &=: \Delta_T + \frac{1}{q - 2h}\int_{r-q/2+h}^{r+q/2-h} a(v)\Sigma^{(k,l)}(v)dv + O(T^{-1}) + O(h^2). \end{aligned}$$

Let us comment on these findings. To make the remainder $O(h^2)$ negligible, we need to impose $Th^4 \to 0$. For instance, we could consider a bandwidth h in the form

$$h = c\frac{q}{2\sqrt{3}}T^{-2/7}, \qquad \text{for some constant } c > 0.$$

The factor $\frac{q}{2\sqrt{3}}$ takes into account the standard deviation of a uniform design on the interval $[r - q/2, r + q/2]$. The term Δ_T is a sum of independent centered variables, having a Gaussian limit. Finally, let us focus on the last integral and notice that

$$\frac{1}{q-2h}\int_{r-q/2+h}^{r+q/2-h} a(v)\Sigma^{(k,l)}(v)dv = \frac{1}{q}\int_{r-q/2}^{r+q/2} a(v)\Sigma^{(k,l)}(v)dv + O(h).$$

Thus S_T preserves a non-negligible bias as an estimator of $q^{-1}\int_{r-q/2}^{r+q/2} a(v)\Sigma^{(k,l)}(v)\, dv$. The solution we will propose to remove this bias is to define estimators like S_T with modified q and thus with modified bounds j_{min} and j_{max} of the set $\mathcal{J}$.

18.5.2 Assumptions

Assumption 18.1 (a) The process (ϵ_t) is iid such that $E(\epsilon_t\epsilon_t') = I_d$ and $\sup_t \| \epsilon_{i,t} \|_\mu < \infty$ for some $\mu > 8$ and for all $i \in \{1, \dots, d\}$ where $\| \,.\, \|_\mu := (E \| \,.\, \|^\mu)^{1/\mu}$ with $\| \,.\, \|$ being the Euclidean norm. Moreover $E\left(\epsilon_t^{(i)}\epsilon_t^{(j)}\epsilon_t^{(k)}\right) = 0$, $i, j, k \in \{1, \dots, d\}$.

(b) The matrix $\mathbb{A}$ given in (18.12) is of full rank.

The covariance of the system (18.1) is allowed to be time-varying, as follows:

Assumption 18.2 We assume that $H_t = G(t/T)$, where the matrix-valued function $G(\cdot)$ is a lower triangular matrix with positive diagonal components. The components $\{g_{k,l}(r) : 1 \le k, l \le d\}$ of the matrix $G(r)$ are measurable deterministic functions on the interval (0, 1], with $\sup_{r\in(0,1]} |g_{k,l}(r)| < \infty$, $1 \le k, l \le d$. The functions $g_{k,l}(\cdot)$ satisfy a Lipschitz condition piecewise on a finite partition of (0, 1] in sub-intervals (the partition may depend on k, l). The matrix $\Sigma(r) = G(r)G(r)'$ is positive definite for all r and $\inf_{r\in(0,1]} \lambda_{min}(\Sigma(r)) > 0$, where $\lambda_{min}(\Gamma)$ denotes the smallest eigenvalue of the symmetric matrix Γ.

The OIRFs estimators we investigate in the following are obtained as products between a functional of the innovation vector u_t, and a centered functional of matrix u_tu_t' (the estimator of some square root matrix built using the covariance structure $\Sigma(\cdot)$). The $\sqrt{T}$-asymptotic normality of the OIRFs estimators is then deduced from the asymptotic behavior of the two factors. The condition $E\left(\epsilon_t^{(i)}\epsilon_t^{(j)}\epsilon_t^{(k)}\right) = 0$,

$i, j, k \in \{1, \ldots, d\}$ is a convenient condition for simplifying the asymptotic variance of our estimators, that is, making it block diagonal. It is in particular fulfilled if the errors (u_t) are Gaussian. The asymptotic results could be also deduced if this condition fails, the asymptotic variance of the estimators would then include some additional covariance terms.

Assumption 18.3 (i) The kernel $K(\cdot)$ is a bounded symmetric density function defined on the real line such that $K(\cdot)$ is non-decreasing on $(-\infty, 0]$ and decreasing on $[0, \infty)$ and $\int_{\mathbb{R}} |v|K(v)dv < \infty$. The function $K(\cdot)$ is differentiable except over a finite number of points and the derivative $K'(\cdot)$ is a bounded integrable function. Moreover, the Fourier transform $\mathcal{F}[K](\cdot)$ of $K(\cdot)$ satisfies $\int_{\mathbb{R}} |s\mathcal{F}[K](s)|\, ds < \infty$.

(ii) The bandwidths b_{kl}, $1 \leq k \leq l \leq d$, are taken in the range $\mathcal{B}_T = [\underline{c}b_T, \bar{c}b_T]$ where $0 < \underline{c} < \bar{c} < \infty$ and $b_T + 1/Tb_T^{2+\gamma} \to 0$ as $T \to \infty$, for some $\gamma > 0$.

(iii) The kernel $L(\cdot)$ is a symmetric bounded Lipschitz, continuous, density function with support in $[-1, 1]$.

(iv) The bandwidth h satisfies the condition $h^4 T + 1/Th^2 \to 0$ as $T \to \infty$.

18.5.3 Proofs

Below, c, c', c'' and C, C', C'' are constants, possibly different from line to line.

Proof of Lemma 18.1. First, note that

$$\widetilde{H}(r)\widetilde{H}(r)' - \bar{H}(r)\bar{H}(r)' = \frac{1}{q}\int_{r-q/2}^{r+q/2} H(v)H(v)'dv - \left[\frac{1}{q}\int_{r-q/2}^{r+q/2} H(v)dv\right]\left[\frac{1}{q}\int_{r-q/2}^{r+q/2} H(v)dv\right]'$$

is a positive semi-definite matrix, irrespective of the values of r and q. Moreover,

$$\widetilde{H}(r)\widetilde{H}(r)' = \bar{H}(r)\bar{H}(r)' \text{ if and only if } H(\cdot) \text{ is constant over } (r - q/2, r + q/2). \tag{18.17}$$

Indeed for any $a \in \mathbb{R}^d$,

$$\begin{aligned} & a'\left\{\widetilde{H}(r)\widetilde{H}(r)' - \bar{H}(r)\bar{H}(r)'\right\}a \\ &= \frac{1}{q}\int_{r-q/2}^{r+q/2} a'\left[H(v) - \frac{1}{q}\int_{r-q/2}^{r+q/2} H(u)du\right]\left[H(v) - \frac{1}{q}\int_{r-q/2}^{r+q/2} H(u)du\right]' a\, dv \\ &= \frac{1}{q}\int_{r-q/2}^{r+q/2} \left\|a'\left[H(v) - \frac{1}{q}\int_{r-q/2}^{r+q/2} H(u)du\right]\right\|^2 dv \geq 0. \end{aligned}$$

This shows that $\widetilde{H}(r)\widetilde{H}(r)' - \bar{H}(r)\bar{H}(r)'$ is positive semi-definite. Next, under our assumptions, for each $a \in \mathbb{R}^d$ the mapping

$$v \mapsto \left\| a' \left[H(v) - \frac{1}{q} \int_{r-q/2}^{r+q/2} H(u)du \right] \right\|^2, \tag{18.18}$$

is piecewise continuous on (0, 1). Thus, if $\widetilde{H}(r)\widetilde{H}(r)' = \bar{H}(r)\bar{H}(r)'$, then necessarily the mapping (18.18) is constant and equal to zero for each a. This implies that $H(\cdot)$ is constant over $(r - q/2, r + q/2)$. Conversely, when $H(\cdot)$ is constant, then $\widetilde{H}(\cdot) = \bar{H}(\cdot)$ and thus $\widetilde{H}(r)\widetilde{H}(r)' = \bar{H}(r)\bar{H}(r)'$. Finally, the two statements in the lemma are direct consequences of (18.17) and the positive semi-definiteness of $\widetilde{H}(r)\widetilde{H}(r)' - \bar{H}(r)\bar{H}(r)'$. □

Proof of Proposition 18.1. See [19]. □

To simplify the reading, before proceeding to the next proofs, let us put the OIRFs notation in a nutshell. First, let $S \mapsto C(S)$ be the operator that maps a positive definite matrix into the lower triangular matrix of the Cholesky decomposition of S. Next, consider a matrix-valued function $r \mapsto A(r)$, $r \in (0, 1]$, and, for any $r \in (0, 1]$, $0 < q < 1$ such that $0 < r - q/2 < r + q/2 < 1$, let

$$Int_{r,q}(A) = \frac{1}{q} \int_{r-q/2}^{r+q/2} A(v)dv, \qquad Int_{T,r,q}(A) = \frac{1}{[qT]+1} \sum_{k=-[qT/2]}^{[qT/2]} A_{[rT]-k}.$$

If $\sup_{r\in(0,1)} \|A(r)\|_F < \infty$ and the components of $A(\cdot)$ are piecewise Lipschitz continuous on each sub-interval of a finite number partition of (0, 1], then there exists a constant c such that

$$\sup_{r,q} \left\| Int_{r,q}(A) - Int_{T,r,q}(A) \right\|_F \le cT^{-1}.$$

We can now rewrite the theoretical OIRFs we introduced above as follows: for any $i \ge 1$,

$$\text{(approximated OIRFs)} \qquad \widetilde{\theta}_r^q(i) = \Phi_i \widetilde{H}(r) = \Phi_i C(Int_{r,q}(\Sigma)),$$

and

$$\text{(averaged OIRFs)} \qquad \bar{\theta}_r^q(i) := \Phi_i \left\{ \frac{1}{q} \int_{r-q/2}^{r+q/2} H(v)dv \right\} = \Phi_i Int_{r,q}(C(\Sigma)).$$

Moreover, the estimators we introduced could be rewritten as follows: with the matrix-valued function $r \mapsto \widehat{U}(r) = \widehat{u}_{[rT]}\widehat{u}'_{[rT]}$, the usual approximated OIRFs estimator is

$$\widehat{\widetilde{\theta}}_r^q(i) = \widehat{\Phi}_i^{ols} \widehat{\widetilde{H}}(r) = \widehat{\Phi}_i^{ols} C\left(Int_{T,r,q}\left(\widehat{U}\right)\right);$$

the new approximated OIRFs estimator is

$$\widehat{\widehat{\theta}}_r^{q,als}(i) = \widehat{\Phi}_i^{als} C\left(Int_{T,r,q}\left(\widehat{U}\right)\right),$$

and the averaged OIRFs estimator is

$$\hat{\hat{\theta}}_r^q(i) = \widehat{\Phi}_i^{als} \frac{q+h}{q} Int_{T,r,q+h}\left(C\left(\widehat{V}\right)\right),$$

where

$$\frac{q+h}{q} Int_{T,r,q+h}\left(C\left(\widehat{V}\right)\right) = \bar{\widehat{H}}(r)\{1+O(1/T)\} = \frac{1+O(1/T)}{[qT]+1} \sum_{k=-[(q+h)T/2]}^{[(q+h)T/2]} \widehat{H}_{[rT]-k}$$

and $\widehat{H}_{[rT]-k}$ is the lower triangular matrix of the Cholesky decomposition of $\widehat{V}_{[rT]-k}$, with $\widehat{V}_{[rT]-k}$ being defined in (18.15).

Proof of Lemma 18.2. First, we use the mean value theorem to get

$$\text{vech}(Int_{T,r,q}(\widehat{U})) = \text{vech}(Int_{T,r,q}(U)) + \frac{\partial \text{vech}(Int_{T,r,q}(U(\vartheta)))}{\partial \vartheta'}|_{\vartheta=\vartheta^*}(\widehat{\vartheta}_{OLS} - \vartheta_0),$$

where $\widehat{U}_t = \hat{u}_t \hat{u}_t'$, $U_t = u_t u_t'$, $U_t(\vartheta) = u_t(\vartheta) u_t(\vartheta)'$, $u_t(\vartheta) = X_t - (\widetilde{X}_{t-1}' \otimes I_d)\vartheta$, for some ϑ^* between $\widehat{\vartheta}_{OLS}$ and ϑ_0. Noting that $\frac{\partial u_t(\vartheta)}{\partial \vartheta'} = -(\widetilde{X}_{t-1}' \otimes I_d)$, the consistency of $\widehat{\vartheta}_{OLS}$ and the fact that u_t is not correlated with $\widetilde{X}_t$, together with basic derivative rules, we have

$$\frac{\partial \text{vech}(Int_{T,r,q}(U(\vartheta))}{\partial \vartheta'}\Big|_{\vartheta=\vartheta^*} = o_p(1).$$

Using the $\sqrt{T}$-convergence of $\widehat{\vartheta}_{OLS}$ (see (18.6), this implies that

$$\sqrt{T}\text{vech}(Int_{T,r,q}(\widehat{U}) - Int_{r,q}(\Sigma)) = \sqrt{T}\text{vech}(Int_{T,r,q}(U) - Int_{r,q}(\Sigma)) + o_p(1), \tag{18.19}$$

where we recall that $Int_{r,q}(\Sigma) = q^{-1} \int_{r-q/2}^{r+q/2} \Sigma(v)dv$.

We next investigate the joint distribution of

$$\sqrt{T}\left[(\widehat{\vartheta}_{OLS} - \vartheta_0)', \left\{\text{vech}(Int_{T,r,q}(U) - Int_{r,q}(\Sigma))\right\}'\right]'. \tag{18.20}$$

We write

$$\begin{pmatrix} \widehat{\vartheta}_{OLS} - \vartheta_0 \\ \text{vech}(Int_{T,r,q}(U) - Int_{r,q}(\Sigma)) \end{pmatrix} = \begin{pmatrix} \left\{Int_{T,0.5,1}(\widetilde{\mathbf{X}}) \otimes I_d\right\}^{-1} & 0 \\ 0 & I_{d(d+1)/2} \end{pmatrix} \begin{pmatrix} \Upsilon_t^1 \\ \Upsilon_t^2 \end{pmatrix},$$

where $\widetilde{\mathbf{X}}_{t-1} = \widetilde{X}_{t-1}\widetilde{X}'_{t-1}$, $\Upsilon_t^1 = \mathrm{vec}(Int_{T,0.5,1}(X^u))$, with $X_t^u = u_t\widetilde{X}'_t$ and

$$\Upsilon_t^2 = \mathrm{vech}(Int_{T,r,q}(U) - Int_{r,q}(\Sigma)).$$

The vector $\Upsilon_t = (\Upsilon_t^{1'}, \Upsilon_t^{2'})'$ is a martingale difference since the process (u_t) is independent. On the other hand, we have $T^{-1}\sum_{t=1}^{T} Int_{T,0.5,1}(\widetilde{\mathbf{X}}) \otimes I_d \to \Lambda_3$, from [20]. Then from the Lindeberg CLT and the Slutsky lemma, (18.20) is asymptotically normally distributed with mean zero. For the asymptotic covariance matrix in (18.11), the top left block is given from the asymptotic normality result (18.6), while the bottom right block can be obtained using the same arguments of [18, Lemmas 7.1–7.4]. The asymptotic covariance matrix is block diagonal since we assumed that $E(u_{it}u_{jt}u_{kt}) = 0, i, j, k \in \{1, \ldots, d\}$ in Assumption 18.2, together considering again that u_t is independent with respect to the past of X_t. Hence the asymptotic covariance matrix of (18.20) is given as in (18.11). The proof of Lemma 18.2 is now complete. □

Next, let us recall the differentiation formula of the Cholesky operator, as follows:

$$\Delta := \frac{\partial \mathrm{vec}(C(\Sigma))}{\partial \mathrm{vec}(\Sigma)} = (I_d \otimes C(\Sigma))Z(C(\Sigma)^{-1} \otimes C(\Sigma)^{-1}), \tag{18.21}$$

where Z is a diagonal matrix such that $Z\mathrm{vec}(A) = \mathrm{vec}(\Phi(A))$ for any $d \times d$-matrix A. Here Φ takes the lower triangular part of a matrix and halves its diagonal, i.e.,

$$\Phi(A)_{ij} = \begin{cases} A_{ij} & i > j \\ \frac{1}{2}A_{ij} & i = j \\ 0 & i < j. \end{cases}$$

Note that

$$(C(\Sigma)^{-1} \otimes C(\Sigma)^{-1})\mathrm{vec}\,(\Sigma) = \mathrm{vec}\left(C(\Sigma)^{-1}\Sigma(C(\Sigma)')^{-1}\right) = \mathrm{vec}(I_d),$$

thus

$$\begin{aligned}\Delta\mathrm{vec}\,(\Sigma) = (I_d \otimes C(\Sigma))Z\mathrm{vec}(I_d) &= (I_d \otimes C(\Sigma))\mathrm{vec}(\Phi(I_d)) \\ &= \mathrm{vec}(C(\Sigma)\Phi(I_d)) = \frac{1}{2}\mathrm{vec}(C(\Sigma)).\end{aligned}$$

Proof of Proposition 18.2. Using our notations, we write

$$\widehat{\widetilde{\theta}}_r^q(i) - \widetilde{\theta}_r^q(i) = \widehat{\Phi}_i^{ols}C(Int_{T,r,q}(\widehat{U})) - \Phi_i C(Int_{r,q}(\Sigma)).$$

From (18.19) and the consistency of the OLS estimator, we have

$$\sqrt{T}(\widehat{\Phi}_i^{ols} C(Int_{T,r,q}(\widehat{U})) - \Phi_i C(Int_{r,q}(\Sigma))) \\ = \sqrt{T}(\widehat{\Phi}_i^{ols} C(Int_{T,r,q}(U)) - \Phi_i C(Int_{r,q}(\Sigma))) + o_p(1).$$

Now let us write

$$\sqrt{T}\text{vec}\left[\widehat{\Phi}_i^{ols} C(Int_{T,r,q}(U) - \Phi_i C(Int_{r,q}(\Sigma))\right] \\ = \sqrt{T}\text{vec}\left[(\widehat{\Phi}_i^{ols} - \Phi_i)C(Int_{r,q}(\Sigma)) + \Phi_i(C(Int_{T,r,q}(U) - C(Int_{r,q}(\Sigma)) \right. \\ \left. + (\widehat{\Phi}_i^{ols} - \Phi_i)(C(Int_{T,r,q}(U)) - C(Int_{r,q}(\Sigma)))\right]. \quad (18.22)$$

Using Lemma 18.2, [16, Rules (8) and (10) in Appendix A.13] and the delta method, both $\sqrt{T}\text{vec}\{\widehat{\Phi}_i^{ols} - \Phi_i\}$ and $\sqrt{T}\text{vec}\{(C(Int_{T,r,q}(U)) - C(Int_{r,q}(\Sigma)))\}$ are asymptotically normal. Hence we have

$$(\widehat{\Phi}_i^{ols} - \Phi_i)(C(Int_{T,r,q}(U)) - C(Int_{r,q}(\Sigma))) = O_p(T^{-1}),$$

and thus from (18.22),

$$\sqrt{T}\text{vec}\left[\widehat{\Phi}_i^{ols} C(Int_{T,r,q}(U) - \Phi_i C(Int_{r,q}(\Sigma))\right] \\ = \sqrt{T}\text{vec}\left[(\widehat{\Phi}_i^{ols} - \Phi_i)C(Int_{r,q}(\Sigma)) \right. \\ \left. + \Phi_i(C(Int_{T,r,q}(U) - C(Int_{r,q}(\Sigma))\right] + o_p(1).$$

Moreover, by the identity $\text{vec}(ABC) = (C' \otimes A)\text{vec}(B)$ for matrices of adequate dimensions, we have, for the right-hand side of the above equation,

$$\text{vec}\left\{(\widehat{\Phi}_i^{ols} - \Phi_i)C(Int_{r,q}(\Sigma))\right\} = (C(Int_{r,q}(\Sigma))' \otimes I_d)\text{vec}(\widehat{\Phi}_i^{ols} - \Phi_i),$$

and

$$\text{vec}\left\{\Phi_i(C(Int_{T,r,q}(U) - C(Int_{r,q}(\Sigma))\right\} \\ = (I_d \otimes \Phi_i)\text{vec}\{C(Int_{T,r,q}(U) - C(Int_{r,q}(\Sigma))\} \\ = (I_d \otimes \Phi_i)\Delta\left\{\text{vec}(Int_{T,r,q}(U)) - \text{vec}(Int_{r,q}(\Sigma)\right\}\{1 + o_{\mathbb{P}}(1)\}.$$

For the last equality, we used (18.21) and the delta method argument. The convergence (18.13) follows by Lemma 18.2 and the CLT. □

Proof of Proposition 18.4. Let us fix $\widetilde{q} \in (0, 1)$ and consider the definitions in (18.15) and (18.16) with the generic $\widetilde{q}$ replacing q. Note that, given a $d \times d$-matrix valued function $A(\cdot)$ defined on $(0, 1]$ with differentiable elements that have Lipschitz continuous derivatives, we have

$$
\begin{aligned}
&Int_{T,r,\widetilde{q}+2h}\left(\text{vec}(A\widehat{V})\right)\\
&\quad = \frac{1}{[(\widetilde{q}+2h)T]+1}\sum_{k=-[(\widetilde{q}+2h)T/2]}^{[(\widetilde{q}+2h)T/2]}\text{vec}(A_{[rT]-k}\widehat{V}_{[rT]-k})\\
&= \frac{1}{\widetilde{q}+2h}\,\frac{1}{T}\sum_{j=[(r-\widetilde{q}/2)T]+1}^{[(r+\widetilde{q}/2)T]}\text{vec}\left(\left[\int_{r-\widetilde{q}/2-h}^{r+\widetilde{q}/2+h}\frac{1}{h}L\left(\frac{v-j/T}{h}\right)A(v)dv\right]\widehat{u}_j\widehat{u}_j'\right)\\
&\qquad + O_{\mathbb{P}}(1/Th)\\
&\quad = \frac{1}{\widetilde{q}+2h}\,\frac{1}{T}\sum_{j=[(r-\widetilde{q}/2)T]+1}^{[(r+\widetilde{q}/2)T]}\text{vec}\left(A(j/T)\widehat{u}_j\widehat{u}_j'\right) + O_{\mathbb{P}}(h^2+1/Th)\\
&\quad = \frac{\widetilde{q}}{\widetilde{q}+2h}\,\frac{1}{[\widetilde{q}T]+1}\sum_{j=-[\widetilde{q}T/2]}^{[\widetilde{q}T/2]}\text{vec}\left(A(([rT]-j)/T)\widehat{u}_{[rT]-j}\widehat{u}_{[rT]-j}'\right)\\
&\qquad + O_{\mathbb{P}}(h^2+1/Th)\\
&\quad = \frac{\widetilde{q}}{\widetilde{q}+2h}\,Int_{T,r,\widetilde{q}}\left(\text{vec}(A\widehat{U})\right) + O_{\mathbb{P}}(h^2+1/Th). \qquad (18.23)
\end{aligned}
$$

The remainder $O_{\mathbb{P}}(1/Th)$ comes from the approximation of $A(\cdot)$ with the values calculated on the grid $j/(Th)$. The remainder $O_{\mathbb{P}}(h^2)$ is given by the first-order Taylor expansion using the Lipschitz continuity of the derivatives of the elements of $A(\cdot)$. Both $O_{\mathbb{P}}(1/Th)$ and $O_{\mathbb{P}}(h^2)$ vanish when $A(\cdot)$ is constant. This will crucially serve to deduce the distribution of the indicator $\widehat{i_{r,q}}$ in the case where $\Sigma(\cdot)$ is constant over the interval $[r-q/2, r+q/2]$.

Moreover, we can write

$$\text{vec}(C(\widehat{V})) - \text{vec}(C(\Sigma)) = \Delta\left[\text{vec}(\widehat{V}) - \text{vec}(\Sigma)\right]\{1+o_{\mathbb{P}}(1)\}.$$

Gathering facts, we now consider $\text{vec}(Int_{T,r,q+h}(C(\widehat{V})))$. We have

$$
\begin{aligned}
&\text{vec}\left(Int_{T,r,q+h}\left(C\left(\widehat{V}\right)\right)\right)\\
&= Int_{T,r,q+h}\left(\text{vec}\left(C\left(\widehat{V}\right)\right)\right)\\
&= Int_{r,q+h}\left(\text{vec}\left(C\left(\Sigma\right)\right)\right) + O_{\mathbb{P}}(1/Th)\\
&+ \left\{Int_{T,r,q+h}(\Delta\text{vec}(\widehat{V})) - Int_{r,q+h}(\Delta\text{vec}(\Sigma)) + O_{\mathbb{P}}(1/Th)\right\}\{1+o_{\mathbb{P}}(1)\}\\
&= \frac{q}{q+h}Int_{r,q}\left(\text{vec}\left(C\left(\Sigma\right)\right)\right) + O_{\mathbb{P}}(1/Th)\\
&+ \frac{1}{q+h}\int_{r-(q+h)/2}^{r-q/2}\text{vec}(C\left(\Sigma(v)\right))dv + \frac{1}{q+h}\int_{r+q/2}^{r+(q+h)/2}\text{vec}(C\left(\Sigma(v)\right))dv\\
&+ \left\{\frac{q-h}{q+h}Int_{T,r,q-h}(\Delta\text{vec}(\widehat{U})) + O_{\mathbb{P}}(h^2+1/Th) - Int_{r,q+h}(\Delta\text{vec}(\Sigma)) + O_{\mathbb{P}}(1/Th)\right\}\\
&\times\{1+o_{\mathbb{P}}(1)\},
\end{aligned}
$$

where we used (18.23) with $q-h$ instead of $\widetilde{q}$, for replacing $Int_{T,r,q+h}(\Delta \text{vec}(\widehat{V}))$. Moreover, since $\Delta \text{vec}\,(\Sigma) = (1/2)\text{vec}(C(\Sigma))$, we also have

$$\begin{aligned}
&Int_{r,q+h}(\Delta \text{vec}(\Sigma)) \\
&= \frac{q-h}{q+h} Int_{r,q-h}(\Delta \text{vec}(\Sigma)) \\
&+ \frac{1}{2}\,\frac{1}{q+h}\int_{r-(q+h)/2}^{r-q/2} \text{vec}(C\,(\Sigma(v)))dv + \frac{1}{2}\,\frac{1}{q+h}\int_{r-q/2}^{r-(q-h)/2} \text{vec}(C\,(\Sigma(v)))dv \\
&+ \frac{1}{2}\,\frac{1}{q+h}\int_{r+q/2}^{r+(q+h)/2} \text{vec}(C\,(\Sigma(v)))dv + \frac{1}{2}\,\frac{1}{q+h}\int_{r+(q-h)/2}^{r+q/2} \text{vec}(C\,(\Sigma(v)))dv \\
&= \frac{q-h}{q+h} Int_{r,q-h}(\Delta \text{vec}(\Sigma)) + \frac{1}{q+h}\int_{r-(q+h)/2}^{r-q/2} \text{vec}(C\,(\Sigma(v)))dv \\
&+ \frac{1}{q+h}\int_{r+q/2}^{r+(q+h)/2} \text{vec}(C\,(\Sigma(v)))dv + O(h^2),
\end{aligned}$$

where we used, for the last equality, the change of variables $v \to v - h/2$ (resp. $v \to v + h/2$) in the integral over the interval $[r - q/2, r - (q-h)/2]$ (resp. $[r + q/2, r + (q+h)/2]$) and the Lipschitz property of the elements on $\Sigma(\cdot)$. Thus, we obtain

$$\begin{aligned}
&\text{vec}\left(Int_{T,r,q+h}\left(C\left(\widehat{V}\right)\right)\right) \\
&= \frac{q}{q+h} Int_{r,q}\left(\text{vec}\left(C\left(\Sigma\right)\right)\right) + \frac{q-h}{q+h}\left\{Int_{T,r,q-h}(\Delta \text{vec}(\widehat{U})) - Int_{r,q-h}(\Delta \text{vec}(\Sigma))\right\} \\
&+ O_{\mathbb{P}}(h^2 + 1/Th).
\end{aligned}$$

That means that

$$\begin{aligned}
&\sqrt{T}\left(\frac{q+h}{q}\text{vec}\left(Int_{T,r,q+h}\left(C\left(\widehat{V}\right)\right)\right) - \text{vec}\left(Int_{r,q}\left(C\left(\Sigma\right)\right)\right)\right) \\
&= \left\{1 - \frac{h}{q}\right\}\sqrt{T}\left\{Int_{T,r,q-h}(\Delta \text{vec}(\widehat{U})) - Int_{r,q-h}(\Delta \text{vec}(\Sigma))\right\} + O_{\mathbb{P}}(\sqrt{Th^4} + 1/\sqrt{Th^2}) \\
&= \sqrt{T}\left\{Int_{T,r,q-h}(\Delta \text{vec}(\widehat{U})) - Int_{r,q-h}(\Delta \text{vec}(\Sigma))\right\} + O_{\mathbb{P}}(h + \sqrt{Th^4} + 1/\sqrt{Th^2}). \quad (18.24)
\end{aligned}$$

It also means that

$$\frac{q+h}{q}\text{vec}\left(Int_{T,r,q+h}\left(C\left(\widehat{V}\right)\right)\right) = \text{vec}\left(Int_{r,q}\left(C\left(\Sigma\right)\right)\right) + O_{\mathbb{P}}(1/\sqrt{T}). \quad (18.25)$$

We now have the ingredients to derive the asymptotic normality of our averaged OIRFs estimator

$$\hat{\hat{\theta}}_r^q(i) = \widehat{\Phi}_i^{als}\frac{q+h}{q} Int_{T,r,q+h}\left(C\left(\widehat{V}\right)\right)$$

of the averaged OIRFs $\bar{\theta}_r^q(i) = \Phi_i Int_{r,q}(C(\Sigma))$. First, note that by (18.25) and the $\sqrt{T}$-convergence of $\text{vec}(\widehat{\Phi}_i^{als})$,

$$\sqrt{T}\text{vec}\left(\hat{\bar{\theta}}_r^q(i) - \bar{\theta}_r^q(i)\right) = \text{vec}\Bigg[\sqrt{T}\left(\widehat{\Phi}_i^{als} - \widehat{\Phi}_i\right) Int_{r,q}(C(\Sigma)) \\ + \widehat{\Phi}_i\sqrt{T}\left\{\frac{q+h}{q} Int_{T,r,q+h}\left(C\left(\widehat{V}\right)\right) - Int_{r,q}(C(\Sigma))\right\}\Bigg] + o_{\mathbb{P}}(1).$$

By (18.24), the $\sqrt{T}$-asymptotic normality of

$$(I_d \otimes \Phi_i)\left\{\frac{q+h}{q}\text{vec}\left(Int_{T,r,q+h}(C(\widehat{V}))\right) - \text{vec}\left(Int_{r,q}\left(C\left(\Sigma\right)\right)\right)\right\} \\ = (I_d \otimes \Phi_i)\left\{\text{vec}\left(\widehat{\bar{H}}(r)\right) - \text{vec}\left(Int_{r,q}\left(C\left(\Sigma\right)\right)\right)\right\}$$

follows from the CLT applied to

$$\begin{aligned} &\sqrt{T}(I_d \otimes \Phi_i)\left\{Int_{T,r,q-h}(\Delta\text{vec}(\widehat{U})) - Int_{r,q-h}(\Delta\text{vec}(\Sigma))\right\} \\ &= \sqrt{T}(I_d \otimes \Phi_i)\Delta\left\{\text{vec}(Int_{T,r,q-h}(\widehat{U})) - \text{vec}(Int_{r,q-h}(\Sigma))\right\} \\ &= \sqrt{T}(I_d \otimes \Phi_i)\Delta\left\{\text{vec}(Int_{T,r,q-h}(U)) - \text{vec}(Int_{r,q-h}(\Sigma))\right\}\{1 + o_{\mathbb{P}}(1)\} \\ &= \sqrt{T}(I_d \otimes \Phi_i)\Delta\left\{\text{vec}(Int_{T,r,q}(U)) - \text{vec}(Int_{r,q}(\Sigma))\right\}\{1 + o_{\mathbb{P}}(1)\}. \end{aligned}$$

The result follows from the $\sqrt{T}$-asymptotic normality of $\text{vec}(\widehat{\Phi}_i^{als})$ and the zero-mean condition for the product of any three components of the error vector, see Assumption 18.2. □

Acknowledgements Valentin Patilea acknowledges support from the research program *Modèles et traitements mathématiques des données en très grande dimension,* of Fondation du Risque and MEDIAMETRIE, and from the grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI—UEFISCDI, project number PN-III-P4-ID-PCE-2020-1112, within PNCDI III. Hamdi Raïssi acknowledges the ANID funding Fondecyt 1201898. The authors are grateful to Fulvio Pegoraro (Banque de France) for comments on the manuscript.

References

1. ALTER, A. AND BEYER, A. (2014). The dynamics of spillover effects during the european sovereign debt turmoil. *Journal of Banking & Finance* **42** 134–153.
2. AUE, A., HÖRMANN, S., HORVÁTH, L. AND REIMHERR, M. (2009). Break detection in the covariance structure of multivariate time series models. *Ann Statist* **37** 4046–4087.
3. BEETSMA, R. AND GIULIODORI, M. (2012). The changing macroeconomic response to stock market volatility shocks. *Journal of Macroeconomics* **34** 281–293.
4. BERNANKE, B. AND MIHOV, I. (1998a). The liquidity effect and long-run neutrality. NBER Working Papers 6608, National Bureau of Economic Research, Inc.

5. Bernanke, B. S. and Mihov, I. (1998b). Measuring monetary policy. *The quarterly journal of economics* **113** 869–902.
6. Blanchard, O. and Simon, J. (2001). The long and large decline in U.S. output volatility. *Brookings Papers on Economic Activity* **32** 135–174.
7. Cavaliere, G. and Taylor, A. M. R. (2008a). Bootstrap unit root tests for time series with nonstationary volatility. *Econometric Theory* **24** 43–71.
8. Cavaliere, G. and Taylor, A. M. R. (2008b). Time-transformed unit root tests for models with non-stationary volatility. *J Time Ser Anal* **29** 300–330.
9. Cavaliere, G., Rahbek, A. and Taylor, A. M. R. (2010). Testing for co-integration in vector autoregressions with non-stationary volatility. *J Econometrics* **158** 7–24.
10. Dahlhaus, R. (1997). Fitting time series models to nonstationary processes. *Ann Statist* **25** 1–37.
11. Dées, S. and Saint- Guilhem, A. (2011). The role of the united states in the global economy and its evolution over time. *Empirical Economics* **41** 573–591.
12. Diebold, F. X. and Y lmaz, K. (2014). On the network topology of variance decompositions: measuring the connectedness of financial firms. *J Econometrics* **182** 119–134.
13. Flury, B. N. (1985). Analysis of linear combinations with extreme ratios of variance. *J Amer Statist Assoc* **80** 915–922.
14. Giraitis, L., Kapetanios, G. and Yates, T. (2018). Inference on multivariate heteroscedastic time-varying random coefficient models. *J Time Series Anal* **39** 129–149.
15. Kew, H. and Harris, D. (2009). Heteroskedasticity-robust testing for a fractional unit root. *Econometric Theory* **25** 1734–1753.
16. Lütkepohl, H. (2005). *New Introduction to Multiple Time Series Analysis*. Springer-Verlag, Berlin.
17. Nazlioglu, S., Soytas, U. and Gupta, R. (2015). Oil prices and financial stress: A volatility spillover analysis. *Energy Policy* **82** 278–288.
18. Patilea, V. and Raïssi, H. (2010). Adaptive estimation of vector autoregressive models with time-varying variance: application to testing linear causality in mean. *arXiv*: arxiv.org/abs/1007.1193.
19. Patilea, V. and Raïssi, H. (2020). Orthogonal impulse response analysis in presence of time-varying covariance. *arXiv*: arxiv.org/abs/2003.11188.
20. Patilea, V. and Raïssi, H. (2012). Adaptive estimation of vector autoregressive models with time-varying variance: application to testing linear causality in mean. *J Statist Plann Inference* **142** 2891–2912.
21. Patilea, V. and Raïssi, H. (2013). Corrected portmanteau tests for VAR models with time-varying variance. *J Multivariate Anal* **116** 190–207.
22. Primiceri, G. E. (2005). Time varying structural vector autoregressions and monetary policy. *Rev Econom Stud* **72** 821–852.
23. Raïssi, H. (2015). Autoregressive order identification for VAR models with non constant variance. *Comm Statist Theory Methods* **44** 2059–2078.
24. Sensier, M. and van Dijk, D. (2004). Testing for volatility changes in U.S. macroeconomic time series. *The Review of Economics and Statistics* **86** 833–839.
25. Sims, C. A. (1999). Drifts and breaks in monetary policy. Tech. rep., Working Paper, Yale University
26. Stock, J. and Watson, M. (2002). Has the business cycle changed and why? NBER Working Papers 9127, National Bureau of Economic Research, Inc.
27. Stock, J. H. and Watson, M. W. (2005). Understanding changes in international business cycle dynamics. *Journal of the European Economic Association* **3** 968–1006.
28. Strohsal, T., Proaño, C. and Wolters, J. (2019). Assessing the cross-country interaction of financial cycles: evidence from a multivariate spectral analysis of the USA and the UK. *Empirical Economics* **57** 385–398.
29. Strongin, S. (1995). The identification of monetary policy disturbances explaining the liquidity puzzle. *Journal of Monetary Economics* **35** 463–497.
30. Xu, K.- L. and Phillips, P. C. B. (2008). Adaptive estimation of autoregressive models with time-varying variances. *J Econometrics* **142** 265–280.

Chapter 19
Robustness Aspects of Optimal Transport

Elvezio Ronchetti

Abstract Optimal transportation is a flourishing area of research and applications in many different fields. We provide an overview of the stability issues related to the use of these techniques by means of a structured discussion of different results available in the literature, which have been developed to face these problems. We point out that important open problems still remain in order to achieve a complete theory of robust optimal transportation.

19.1 Introduction

Professor Masanobu Taniguchi has always been eager to discuss new ideas about challenging problems, and his research has covered several different fields. For the paper of this Festschrift in his honor, we chose to overview and discuss robustness aspects of optimal transportation methods and to outline open challenging problems. Optimal transportation is an important and trending research area with applications in many different fields, whereas robustness is a key issue related to the stability of the results of an analysis obtained using optimal transport techniques in the presence of deviations from assumed models. We provide an overview of the available results in the literature (which are all quite recent) and show the link between penalization and concepts from the classical theory of robust statistics.

The rest of the paper is organized as follows. Section 19.2 summarizes the key elements of optimal transportation, which are needed to discuss its robustness issues. These include Monge's and Kantorovich's formulations and the key concept of the Wasserstein distance. Section 19.3 discusses the basic approach to define robust estimators via the Wasserstein distance and the important link between robust M-estimators and penalized estimators in a saturated linear model. Section 19.4 presents an interesting approach to robust optimal transport, which is based on penalization and establishes the link with familiar ideas from the classical theory of robust

E. Ronchetti (✉)
University of Geneva, Blv. Pont d'Arve 40, CH-1211, Geneva, Switzerland
e-mail: elvezio.ronchetti@unige.ch

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_19

statistics, such as truncation of score functions. Finally, Sect. 19.5 discusses possible alternative approaches to obtain robust estimators through the definition of multivariate ranks and provides an outlook on open problems.

19.2 Optimal Transport

Optimal transportation goes back to [36], who considered the problem of finding the optimal way to move given piles of sand to fill up given holes of the same total volume. Monge's problem was revisited by [28], who considered the economic problem of optimal allocation of resources. Nowadays optimal transportation is a flourishing area of research and applications in many fields, including mathematics, statistics, economics, computer vision, imaging, and machine learning. Book-length presentation are provided by [44, 52]. A statistical perspective can be found in [31].

19.2.1 Basic Formulation

19.2.1.1 Optimal Transport: Monge's Formulation

Let μ and ν be two probability measures over $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the Borel sigma-field and $c : \mathbb{R}^2 \to \mathbb{R}$ a Borel-measurable cost function, i.e. $c(x, y)$ is the cost of transporting x to y. Then, for $X \sim \mu$, $Y \sim \nu$, solve

$$\inf_{\mathcal{T}:X \to Y} \int_{\mathbb{R}} c(x, \mathcal{T}(x)) d\mu. \tag{19.1}$$

The solution to (19.1) is the optimal transportation map $\mathcal{T}$, such that $\mathcal{T}_{\#}\mu = \nu$ (to be read: $\mathcal{T}$ pushes μ forward to ν).

19.2.1.2 Optimal Transport: Kantorovich's Formulation

Monge's problem (defined using $c(x, y) = |x - y|$) remained open until the 1940s,s, when it was revisited by [28], whose objective was to minimize

$$\text{KP}(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X} \times \mathcal{X}} c(x, y) d\gamma(x, y), \tag{19.2}$$

where "KP" stands for Kantorovich's problem and the infimum is over all pairs (X, Y) of (μ, ν), belonging to $\Gamma(\mu, \nu)$, the set of probability measures γ on $\mathcal{X} \times \mathcal{X}$, satisfying $\gamma(A \times \mathcal{X}) = \mu(A)$ and $\gamma(\mathcal{X} \times B) = \nu(B)$, for Borel sets A, B. Kantorovich's problem is more general than Monge's problem, since it allows mass splitting. It can be solved via its dual formulation and under some conditions (see e.g. [44, Theorem 1.40]) there is no duality gap, i.e. the extremum of the primal problem (19.2) equals the extremum of the dual problem.

19.2.2 Wasserstein Distance

The solution to KP(μ, ν) with $c(x, y) = |x - y|^p$ defines the p-Wasserstein distance, for $p \geq 1$,

$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathcal{X}\times\mathcal{X}} |x - y|^p d\gamma(x, y) \right)^{1/p}. \tag{19.3}$$

Let $\mathcal{P}(\mathcal{X})$ be the set of all probability measures defined on $\mathcal{X}$ (a convex subset of $\mathbb{R}$) with finite second moment. Then, the Wasserstein space $(\mathcal{P}(\mathcal{X}), W_2)$ is a metric space and a geodesic space; see [40]. Thus, in $(\mathcal{P}(\mathcal{X}), W_2)$, for $\mu, \nu \in \mathcal{P}(\mathcal{X})$, there exists a continuous path going from μ to ν, such that its length is the distance between the two measures. An excellent discussion of Wasserstein distances and Wasserstein spaces can be found in [39, 40].

The Wasserstein distance has gained importance in machine learning due to its ability to capture the geometric structure of the sample space. This makes it particularly useful in applications for instance in computer vision and image analysis. Moreover, new developments are appearing in risk management; see [29].

From a statistical point of view, the Wasserstein distance can be used to evaluate the accuracy of approximations to the exact distribution of an estimator; see [31]. For instance, let $\{X_i\}, i = 1, \ldots, n$, be iid random variables with $E[X_i] = 0$, $V[X_i] = 1$, and $S_n := \sqrt{n}\bar{X}_n$ with exact CDF P_n, saddlepoint CDF P_{sad} (a higher-order approximation obtained using saddlepoint methods; see [11]), standard normal CDF Φ (obtained by means of the central limit theorem). Then, $W_1(P_n, \Phi) = O(n^{-1/2})$ and $W_1(P_n, P_{\text{sad}}) = O(n^{-1})$, which implies that P_{sad} is closer to the exact P_n than the normal Φ in the Wasserstein distance.

19.3 Robust Approaches

It is a basic tenet of science that models are only approximations to reality. Already in 1960, [51] showed the dramatic loss of efficiency of optimal procedures in the presence of small deviations from the assumed stochastic model.

Robust statistics deals with deviations from ideal models and their dangers for corresponding inference procedures. Its primary goal is the development of procedures which are still reliable and reasonably efficient under small deviations from the model, i.e. when the underlying distribution lies in a neighborhood of the assumed model. Therefore, one can view robust statistics as an extension of parametric statistics, taking into account that parametric models are at best only approximations to reality. Robust statistics (stability) has to play a major role nowadays with the flourishing of (new) procedures to analyze complex data; see the discussions in [22, 55].

From the seminal papers [20, 24, 51] which marked the beginning of the systematic development of the theory and applications of robust statistics, a large and rich literature on robust statistics has emerged in the past decades. An account of the basic general theory can be found in the classical books [21, 25] (2nd edition [26]), [34]. Additional general books include [10, 16, 23, 27, 37, 41, 43, 50], [53, Chap. 5], and the quantile regression approach in [30]. Recent reviews are provided in [1, 42].

19.3.1 Robustifying Estimators via Wasserstein Distance

In this context, it seems natural to consider the Wasserstein distance as the basis of a minimum distance estimator. Specifically, given a family of parametric models $\{\nu_\theta\}$ for the distribution of the data and the empirical measure μ_n of the sample, we can define a minimum Kantorovich estimator (MKE) by

$$\hat{\theta} = \arg\min_{\theta \in \Theta} \mathrm{KP}(\mu_n, \nu_\theta),$$

as proposed in [4] in the setting of independent data. Recently, this idea has been applied by [14] in time series to obtain an estimator of θ which minimizes the Wasserstein distance between the asymptotic theoretical distribution of the periodogram ordinates (an exponential distribution with expectation $f(\cdot;\theta)$, the spectral density) and the empirical distribution.

Minimum distance estimators have been often suggested as robust alternatives to maximum likelihood estimators, but in spite of the large body of work (for a book-length presentation see [5]), the available results on the robustness of minimum distance estimators cannot be applied in a straightforward way in this case; see the discussion in [6]. Indeed results on existence and consistency of the MKE can be found in [4, 6], but the asymptotic distribution of the MKE (see [6, Theorem 2.3]) is not implied by the standard linear form $\hat{\theta} = \theta + \frac{1}{n}\sum_{i=1}^{n} IF(X_i; \hat{\theta}, \mu)$ and this does not allow to read the influence function (IF) and robustify the estimator using the standard robustness theory by imposing a bounded influence function as in [21]. Note however that some results in this direction for the location-scale model are available in [3]. Therefore, in spite of the appealing properties of the Wasserstein distance, no general theory is yet available to claim robustness properties for the MKE. In fact, [54] provides a critical discussion on the properties of the MKE. It indicates its limits and shows its inferiority in terms of mean squared error with respect to minimum distance estimators based on the Kolmogorov distance. Moreover, it stresses its lack of robustness, especially in the presence of data with heavy tail distributions. Section 19.4 will shed more light on possible robustifications of the estimator.

Finally, let us mention a different way to achieve robustness as suggested by [49]. If we rewrite the 1-Wasserstein distance (19.3) using its dual formulation, i.e.

$$W_1(\mu, \nu_\theta) = \sup_{\phi \in B_L} E_\mu[\phi(X)] - E_{\nu_\theta}[\phi(Y)], \tag{19.4}$$

where B_L is the unit ball of the Lipschitz functions space, we can estimate robustly both expectations on the right-hand side of (19.4) by using the Median-of-Means estimator introduced by [8, 33] and this will lead to a robust estimator for θ.

19.3.2 Saturation in Linear Models

A complementary perspective between robustness and sparsity in linear models is provided by the so-called saturated regression model (or mean-shift outlier model):

$$y_i = \sum_{j=1}^{d} x_{ij}\beta_j + \gamma_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $d < n$ and the $\gamma_i' s$ are nonzero when observation i is an outlier.

Minimizing

$$\sum_{i=1}^{n}(y_i - \sum_{j=1}^{d} x_{ij}\beta_j - \gamma_i)^2 + \sum_{i=1}^{n} p_\lambda(|\gamma_i|) \tag{19.5}$$

over the $\beta_j' s$ and the $\gamma_i' s$ for a given penalty $p_\lambda(\cdot)$ with a generic penalty parameter λ, we obtain an estimator of the $\beta_j' s$ matching the one obtained by minimizing

$$\sum_{i=1}^{n} \rho(y_i - \sum_{j=1}^{d} x_{ij}\beta_j)$$

for some loss function $\rho(\cdot)$. This is an M-estimator for the $\beta_j' s$ with score function $\psi(\cdot)$, the derivative of $\rho(\cdot)$. For instance, the Huber estimator is obtained by using the Lasso penalty.

This idea goes back to [45] (in the case of the Huber estimator), [13] (in the context of approximate message passing), [35, 46, 47]. It has been also successfully exploited by David Hendry and coauthors in the econometrics literature (Autometrics) as a variable selection tool and more recently as an outlier detection technique. A very recent development includes a double ($L_0 - L_2$) penalization method by [48].

In the past few years, this approach has become a popular tool in the machine learning community to enforce robustness in available algorithms. We believe that its connection to M-estimation opens the door to a beneficial cross-fertilization between the sparse modeling literature and robust statistics. In the next section, we outline this approach in optimal transport and its link with classical methods developed in the robustness literature.

19.4 Robust Optimal Transport

In this section, we present an interesting approach by [38], which establishes the link between penalization and basic ideas of the classical theory of robust statistics in the spirit of Sect. 19.3.2.

To achieve outlier robustness in MKE, [38] introduced two formulations of a modified optimal transport, which lead to robust versions of MKE. The first one is based on the minimization

$$\text{ROBOT}(\mu, \nu_\theta) := \inf_{\theta \in \Theta, \mu + s \in \mathcal{P}(\mathcal{X})} \text{KP}(\mu + s, \nu_\theta) + \lambda ||s||_{TV}, \tag{19.6}$$

where "ROBOT" stands for robust optimal transport, $\mathcal{P}(\mathcal{X})$ is the set of all probability measures defined on $\mathcal{X}$ (a convex subset of $\mathbb{R}$) with finite second moment, and $|| \cdot ||_{TV}$ is the total variation norm defined as $||\mu||_{TV} = \int \frac{1}{2}|\mu(dx)|$. One can notice the same structure between (19.5) and (19.6).

The second formulation is based on the minimization of (19.2) with a modified cost function

$$\text{KP}_\lambda(\mu, \nu_\theta) = \inf_{\gamma \in \Gamma(\mu, \nu_\theta)} \int_{\mathcal{X} \times \mathcal{X}} c_\lambda(x, y) d\gamma(x, y), \tag{19.7}$$

where $c_\lambda(x, y) = \min\{c(x, y), 2\lambda\}$ is a truncated cost function and λ a generic truncation parameter.

The paper [38] proved two key results. The first one states that (19.6) and (19.7) lead to the same optimal transport, linking penalization with a basic idea in the theory of robust statistics, namely, the truncation of score functions. This corresponds to the result presented in Sect. 19.3.2. The second one states that

$$\begin{aligned} &\text{ROBOT}(\tilde{\mu}, \nu_\theta) \\ &\leq \min\{\text{KP}(\mu, \nu_\theta) + \lambda\epsilon ||\mu - \mu_c||_{TV}, \text{KP}(\tilde{\mu}, \nu_\theta), \lambda ||\tilde{\mu} - \nu_\theta||_{TV}\}, \end{aligned} \tag{19.8}$$

where $\tilde{\mu} = (1 - \epsilon)\mu + \epsilon\mu_c$ for some $\epsilon \in [0, 1)$.

Since the total variation norm is bounded by 1, (19.8) implies that $\text{ROBOT}(\tilde{\mu}, \nu_\theta)$ is bounded for any measure $\tilde{\mu}$ in a ϵ-neighborhood of μ defined by the total variation norm. In practice this means that ROBOT$(\tilde{\mu}, \nu_\theta)$ cannot go to infinity in the presence of a small amount of contamination in the data.

A final remark is made about the truncated cost function $c_\lambda(x, y)$. If we use the familiar quadratic function for the original cost function, the modified cost function c_λ corresponds in the language of robust statistics to the so-called "hard rejection," which is known to provide robustness with an important loss of efficiency at the model; see [21]. Therefore, a Huber loss function could be more appropriate than c_λ because it would lead to a small loss of efficiency at the model. However, no proof of robustness in the spirit of (19.8) is available in this case.

19.5 Related Techniques and Outlook

There are several recent papers where optimal transport is used to define a new notion of multivariate ranks. This leads to new rank estimators, which are known to have robustness properties and can be viewed as robust alternatives to classical estimators in a variety of models; see [9, 12, 15, 17–19].

Robust optimal transport methods in generative models were discussed in [2] and various related robustness issues in [7, 32].

In spite of the promising recent work discussed in this overview, it is fair to say that a full-fledged optimality theory as in [21, 25] is still lacking for robust optimal transport. For instance, concepts of local stability such as the influence function and of global reliability such as the breakdown point still have to be developed. Moreover, minimax estimators obtained with respect to neighborhoods of the model distribution as in [25] would be important steps toward such a theory.

Acknowledgements The author would like to thank Davide La Vecchia and Yannis Yatracos for useful discussions and references on this topic, and the editorial team for many remarks that improved the clarity of the paper.

References

1. Avella Medina, M. and Ronchetti, E. (2015). Robust statistics: A selective overview and new directions. *WIREs Comput Stat* **7** 372–393.
2. Balaji, Y., Chellappa, R. and Feizi, S. (2020). Robust optimal transport with applications in generative modeling and domain adaptation. *Advances in Neural Information Processing Systems* **33** 12934 –12944.
3. Bassetti, F. and Regazzini, E. (2005). Asymptotic properties and robustness of minimum dissimilarity estimators of location-scale parameters. *Teor Veroyatnost i Primenen* **50** 312–330.
4. Bassetti, F., Bodini, A. and Regazzini, E. (2006). On minimum Kantorovich distance estimators. *Statistics & Probability Letters* **76** 1298–1302.
5. Basu, A., Shioya, H. and Park, C. (2011). *Statistical Inference: The Minimum Distance Approach*. Chapman and Hall/CRC, London.

6. Bernton, E., Jacob, P. E., Gerber, M. and Robert, C. P. (2019). On parameter estimation with the Wasserstein distance. *Information and Inference: A Journal of the IMA* **8** 657–676.
7. Blanchet, J., Murthy, K. and Zhang, F. (2022). Optimal transport-based distributionally robust optimization: Structural properties and iterative schemes. *Mathematics of Operations Research* **47** 1500–1529.
8. Catoni, O. (2012). Challenging the empirical mean and empirical variance: A deviation study. *Annales de l'Istitut Henri Poincaré, Probabilités et Statistiques* **48** 1148–1185.
9. Chernozhukov, V., Galichon, A., Hallin, M. and Henry, M. (2017). Monge–Kantorovich depth, quantiles, ranks and signs. *The Annals of Statistics* **45** 223–256.
10. Clarke, B. (2018). *Robustness Theory and Application*. Wiley, New York.
11. Daniels, H. E. (1954). Saddlepoint approximations in statistics. *Annals of Mathematical Statistics* **25** 631–650.
12. Deb, N. and Sen, B. (2022). Multivariate rank-based distribution-free nonparametric testing using measure transportation. *Journal of the American Statistical Association*. *To appear*.
13. Donoho, D. and Montanari, A. (2016). High dimensional robust M-estimation: Asymptotic variance via approximate message passing. *Probability Theory and Related Fields* **166** 935–969.
14. Felix, M. and La Vecchia, D. (2022). Semiparametric estimation for time series: A frequency domain approach based on optimal transportation theory. *Technical Report*.
15. Ghosal, P. and Sen, B. (2022). Multivariate ranks and quantiles using optimal transport: Consistency, rates, and nonparametric testing. *The Annals of Statistics* **50** 1012–1037.
16. Greco, L. and Farcomeni, A. (2015). *Robust Methods for Data Reduction*. Chapman and Hall/CRC, Boca Raton, London, New York.
17. Hallin, M., del Barrio, E., Cuesta- Albertos, J. and Matran, C. (2022a). Center-outward distribution and quantile functions, ranks, and signs in dimension d: A measure transportation approach. *The Annals of Statistics* **49** 1139–1165.
18. Hallin, M., La Vecchia, D. and Liu, H. (2022b). Center-outward R-estimation for semiparametric VARMA models. *Journal of the American Statistical Association*. *To appear*.
19. Hallin, M., La Vecchia, D. and Liu, H. (2022c). Rank-based testing for semiparametric VAR models: A measure transportation approach. arXiv:arxiv.org/abs/2011.06062.
20. Hampel, F. (1968). Contribution to the theory of robust estimation. *PhD Thesis, University of California, Berkeley*.
21. Hampel, F. R., Ronchetti, E., Rousseeuw, P. J. and Stahel, W. A. (1986). *Robust Statistics: The Approach Based on Influence Functions*. Wiley, New York.
22. Hanin, L. (2021). Cavalier use of inferential statistics is a major source of false and irreproducible scientific findings. *Mathematics* **9** 603.
23. Heritier, S., Cantoni, E., Copt, S. and Victoria- Feser, M. (2009). *Robust Methods in Biostatistics*. Wiley, New York.
24. Huber, P. J. (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics* **35** 73–101.
25. Huber, P. J. (1981). *Robust Statistics*. Wiley, New York.
26. Huber, P. J. and Ronchetti, E. (2009). *Robust Statistics*. 2nd edn. Wiley, New York.
27. Jurecková, J. and Picek, J. (2006). *Robust Statistical Methods*. Chapman & Hall/CRC, New York.
28. Kantorovich, L. V. (1942). On the translocation of masses. *(Dokl) Acad Sci URSS* **37** 199–201.
29. Kiesel, R., Rühlicke, R., Stahl, G. and Zheng, J. (2016). The Wasserstein metric and robustness in risk management. *Risks* **4** 1–14.
30. Koenker, R. (2005). *Quantile Regression*. Cambridge University Press.
31. La Vecchia, D., Ronchetti, E. and Ilievski, A. (2022). On some connections between Esscher's tilting, saddlepoint approximations, and optimal transportation: A statistical perspective. *Statistical Science*. *To appear*.

32. Le, K., Nguyen, H., Nguyen, Q. M., Pham, T., Bui, H. and Ho, N. (2021). On robust optimal transport: Computational complexity and barycenter computation. *Technical Report*.
33. Lecué, G. and Lerasle, M. (2020). Robust machine learning by median-of-means: Theory and practice. *The Annals of Statistics* **48** 906–931.
34. Maronna, R. A., Martin, D. R. and Yohai, V. J. (2006). *Robust Statistics: Theory and Methods*. Wiley, New York.
35. McCann, L. and Welsch, R. E. (2007). Robust variable selection using least angle regression and elemental set sampling. *Computational Statistics & Data Analysis* **52** 249–257.
36. Monge, G. (1781). Mémoire sur la théorie des déblais et des remblais. *Mem. Math. Phys. Acad. Royale Sci.* 666-704.
37. Morgenthaler, S. and Tukey, J. (1991). *Configural Polysampling: A Route to Practical Robustness*. Wiley, New York.
38. Mukherjee, D., Guha, A., Solomon, J., Sun, Y. and Yurochkin, M. (2021). Outlier-robust optimal transport. arXiv: arxiv.org/abs/2012.07363.
39. Panaretos, V. M. and Zemel, Y. (2019). Statistical aspects of Wasserstein distances. *Annual Review of Statistics and its Application* **6** 405–431.
40. Panaretos, V. M. and Zemel, Y. (2020). *An Invitation to Statistics in Wasserstein Space*. Springer Nature.
41. Rieder, H. (1994). *Robust Asymptotic Statistics*. Springer, New York.
42. Ronchetti, E. (2021). The main contributions of robust statistics to statistical science and a new challenge. *METRON* **79** 127–135.
43. Rousseeuw, P. J. and Leroy, A. M. (1987). *Robust Regression and Outlier Detection*. Wiley, New York.
44. Santambrogio, F. (2015). *Optimal Transport for Applied Mathematicians*. Springer.
45. Sardy, S., Tseng, P. and Bruce, A. (2001). Robust wavelet denoising. *IEEE Transactions on Signal Processing* **49** 1146–1152.
46. She, Y. and Chen, K. (2017). Robust reduced-rank regression. *Biometrika* **104** 633–647.
47. She, Y. and Owen, A. (2011). Outlier detection using nonconvex penalized regression. *Journal of the American Statistical Association* **106** 626–639.
48. She, Y., Wang, Z. and Shen, J. (2022). Gaining outlier resistance with progressive quantiles: Fast algorithms and theoretical studies. *Journal of the American Statistical Association* **117** 1282–1295.
49. Staerman, G., Laforgue, P., Mozharovskyi, P. and d'Alché Buc, F. (2022). When OT meets MoM: Robust estimation of Wasserstein distance. *Technical Report*.
50. Staudte, R. and Sheather, S. (1990). *Robust Estimation and Testing*. Wiley, New York.
51. Tukey, J. W. (1960). A survey of sampling from contaminated distributions. *Contributions to Probability and Statistics, I. Olkin (Ed.)* Stanford University Press, 448–485.
52. Villani, C. (2009). *Optimal Transport: Old and New*. vol 338. Springer Science & Business Media.
53. Welsh, A. H. (1996). *Aspects of Statistical Inference*. Wiley, New York.
54. Yatracos, Y. G. (2022). Limitations of the Wasserstein MDE for univariate data. *Technical Report*.
55. Yu, B. (2013). Stability. *Bernoulli* **19** 1484–1500.

Chapter 20
Estimating Finite-Time Ruin Probability of Surplus with Long Memory via Malliavin Calculus

Shota Nakamura and Yasutaka Shimizu

Abstract We consider a surplus process of a drifted fractional Brownian motion with the Hurst index $H > 1/2$, which appears as a functional limit of drifted compound Poisson risk models with correlated claims. This is a kind of representation of a surplus with a long memory. Our interest is to construct confidence intervals of the ruin probability of the surplus when the volatility parameter is unknown. We obtain the derivative of the ruin probability w.r.t. the volatility parameter via the Malliavin calculus, and apply the delta method to identify the asymptotic distribution of an estimated ruin probability.

Keywords Finite-time ruin probability · Long memory surplus · Fractional Brownian motion · Malliavin calculus

MSC2020 **60G22**; 60H07; 62P05

20.1 Introduction

In the classical ruin theory initiated by [9], the insurance surplus is described by a drifted compound Poisson process such as

$$X_t = x + ct - \sum_{i=1}^{N_t} U_i, \tag{20.1}$$

where $x, c > 0$, N is a Poisson process, and U_i's are IID random variables with mean μ, representing claim sizes. One of the directions to extend the model is the following drifted Lévy surplus $(R_t)_{0<t<T}$:

S. Nakamura · Y. Shimizu (✉)
Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
e-mail: shimizu@waseda.jp

S. Nakamura
e-mail: nakamurashota@akane.waseda.jp

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_20

$$R_t = u + dt + \sigma W_t - V_t, \tag{20.2}$$

where $u, d, \sigma > 0$, W_t is a standard Brownian motion and V_t is a Lévy subordinator. The model (20.2) is a natural extension of (20.1) and considers claim sizes with stationary independent increment. Statistical inference for ruin probability based on the model (20.2) has been studied by many authors; see, e.g., [1, 14] and the references therein. However, such an independent assumption is often unrealistic in a certain insurance contract because large claims can be successive once a large claim has occurred. Therefore, it would be better to assume that $(U_i)_{i\in\mathbb{N}}$ are correlated. The paper [10, Theorem 3] assumed that there exists a constant $D \in (0, 1)$ and a slowly varying function L: $L(tx) \sim L(x)$ as $x \to \infty$ for any $t > 0$, such that

$$\mathbb{E}[U_i U_{i+k}] \sim k^{-D} L(k), \quad k \to \infty, \tag{20.3}$$

under which the process $(U_i)_{i\in\mathbb{N}}$ has a *long memory*, i.e., $\sum_{k=1}^{\infty} Cov(U_1, U_k) = \infty$.

Let us consider a sequence of such a long memory surplus processes $X^n = \left(X_t^n\right)_{t\in[0,T]}$ indexed by $n = 1, 2, \ldots$:

$$X_t^n = x_n + c_n t - \sum_{i=0}^{N_t^n} U_i,$$

where x_n, c_n are positive sequences, $\left(N_t^n\right)_{t\in[0,T]}$ is a Poisson process with the intensity n and U_i's are correlated random variables (see (20.3)). Then, according to [10, Theorem 3], there exists a norming sequence $(\eta_n)_{n\in\mathbb{N}}$ and some constants u and θ such that the process X^n/η_n converges weakly in a functional space $D[0, \infty)$, a space of càdlàg functions with the Skorokhood topology, i.e.,

$$\frac{X_t^n}{\eta_n} \xrightarrow{d} u + \theta t - W_t^H \quad \text{in} \quad \mathcal{D}[0, \infty) \quad (n \to \infty),$$

where W^H is a fractional Brownian motion with the Hurst parameter $H \in (\frac{1}{2}, 1)$. In such a way, the surplus driven by a fractional Brownian motion naturally appears as a limit of a Poissonian model with a long memory.

Some earlier works model a surplus by fractional Brownian motions. The paper [8] discussed the surplus of insurance and reinsurance companies as the two-dimensional fractional Brownian motion and derived asymptotic of the ruin probability when the initial capital tends to infinity. The paper [3] considered a drifted mixed fractional Brownian motion as a surplus model and estimated the ruin probability with an unknown drift parameter. In this paper, we are interested in the following drifted fractional Brownian motion as a surplus model:

$$X_t = u + \sigma\theta t - \sigma W_t^H, \tag{20.4}$$

where $\theta > 0$ and $H \in (\frac{1}{2}, 1)$ are known parameters and $\sigma > 0$ is an unknown parameter. Since our model is a normalized limit of a classical type surplus with a known premium rate c_n, the drift will be known under a suitable scaling. Therefore we assume θ is known although the scaling parameter σ is unknown.

Our interest is to estimate the finite-time ruin probability, which is defined as, for any $T \in (0, \infty)$,

$$\Psi_\sigma(u, T) := \mathbb{P}\left(\inf_{0 \le t \le T} X_t < 0\right).$$

The rest of the paper is organized as follows. In Sect. 20.2, we prepare some notation and give a brief review of the Malliavin calculus. In Sect. 20.3, we provide a result on estimating the volatility parameter σ and the ruin probability by the delta method. In this procedure, the partial derivative $\frac{\partial}{\partial \sigma}\Psi_\sigma(u, T)$ is required to obtain confidence intervals of the ruin probability, so we derive its explicit form using the integration by parts formula in the Malliavin calculus in Sect. 20.4.

20.2 Preliminaries

20.2.1 Notation

We use the following notations.

- $A \lesssim B$ means that there exists a universal constant $c > 0$ such that $A \le cB$.
- The partial derivative of the function f at the point $x \in \mathbb{R}^n$ with respect to the ith variable is denoted by $\partial_i f(x)$.
- Let $\overline{A}^{||\cdot||}$ be the closure of the subset A of the norm space $(B, ||\cdot||)$.
- Let $\otimes$ be the tensor product of the norm space.
- Let $H_n(\cdot)$ denote the n-th order *Hermite polynomial*, which is defined by

$$H_n(x) = \frac{(-1)^n}{n!} e^{\frac{x^2}{2}} \frac{d^n}{dx^n}\left(e^{-\frac{x^2}{2}}\right) (n \ge 1).$$

- Let $\mathcal{D}(A)$ be the Skorokhod space on the set $A \subset \mathbb{R}_+$.
- Let $C_\uparrow^\infty(\mathbb{R}^n)$ be the set of all infinitely continuously differentiable functions $f : \mathbb{R}^n \to \mathbb{R}$ such that f and all of its partial derivatives are of polynomial growth.
- Let $C_b^\infty(\mathbb{R}^n)$ be the set of all infinitely continuously differentiable functions $f : \mathbb{R}^n \to \mathbb{R}$ such that f and all of its partial derivatives are bounded.
- For any $p > 1$ and $f, g \in L^p(\mathbb{R})$ we define the convolution $f * g$ as

$$f * g(x) := \int_{\mathbb{R}} f(x - y) g(y) dy.$$

- Let $\Gamma(\cdot)$ and $\mathrm{B}(\cdot, \cdot)$ be the gamma function and the beta function, i.e.,

$$\Gamma(x) = \int_0^\infty t^{x-1} e^t dt \quad (x > 0)$$
$$\mathrm{B}(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} dt \quad (x, y > 0).$$

- We denote the left and right-sided fractional integrals $I^\alpha_{a\pm} f(\cdot)$ and derivatives $D^\alpha_{a\pm} f(\cdot)$ by

$$I^\alpha_{a+} f(x) = \frac{1}{\Gamma(\alpha)} \int_a^x (x-y)^{\alpha-1} f(y) dy,$$
$$I^\alpha_{b-} f(x) = \frac{1}{\Gamma(\alpha)} \int_x^b (y-x)^{\alpha-1} f(y) dy,$$
$$D^\alpha_{a+} g(x) = \frac{1}{\Gamma(1-\alpha)} \left(\frac{g(x)}{(x-a)^\alpha} + \alpha \int_a^x \frac{g(x) - g(y)}{(x-y)^{\alpha+1}} dy \right),$$
$$D^\alpha_{b-} g(x) = \frac{1}{\Gamma(1-\alpha)} \left(\frac{g(x)}{(b-x)^\alpha} + \alpha \int_x^b \frac{g(x) - g(y)}{(y-x)^{\alpha+1}} dy \right),$$

for any $0 < \alpha < 1, x \in (a, b), f \in L^1(a, b)$ and $g \in I^\alpha_{a+}(L^p)$ $\big($resp. $g \in I^\alpha_{b-}(L^p)\big)$ where $p > 1$ (see [13] for details).

20.2.2 Malliavin Calculus

We briefly introduce the Malliavin calculus; see, e.g., [12, Chaps. 1 and 5]. In the sequel, we denote by G a real separable Hilbert space.

20.2.2.1 Malliavin Calculus on a Real Separable Hilbert Space

Definition 20.1 We say that a stochastic process $(W_g)_{g \in G}$ is an isonormal Gaussian process associated with the real separable Hilbert space G if W is a centered Gaussian family of random variables such that $\mathbb{E}[W_h W_g] = \langle h, g \rangle_G$ for any $h, g \in G$.

In the following, we assume that the isonormal Gaussian $W(\cdot)$ is defined on the complete probability space $(\Omega, \mathcal{G}, \mathbb{P})$, where $\mathcal{G}$ is the σ-algebra generated by W throughout in this paper.

Definition 20.2 Let $\mathcal{G}$ be the σ-algebra generated by an isonormal Gaussian W. If a random variable $F : \Omega \to \mathbb{R}$ satisfies

$$F = f(W(g_1), \ldots, W(g_n)) \quad (f \in C^\infty_\uparrow(\mathbb{R}^n), g_1, \ldots, g_n \in G), \tag{20.5}$$

F is called a smooth random variable, and the set of all such random variables is denoted by $\mathcal{S}_G$.

Definition 20.3 The derivative of a smooth random variable F of the form (20.5) is the G-valued random variable given by

$$D^G F = \sum_{i=1}^{n} \partial_i f\,(W(g_1), \ldots, W(g_n))\, g_i.$$

Since the operator D^G defined in Definition 20.3 is a closable operator, we can extend D^G as a closed operator on $\mathbb{D}_G^{1,p} := \overline{\mathcal{S}_G}^{\|\cdot\|_{1,p}}$ where the seminorm $\|\cdot\|_{1,p}$ on $\mathcal{S}_G$ is defined by

$$\|F\|_{1,p} := \left(\mathbb{E}[F^p] + \mathbb{E}\left[\|D^G F\|_G^p\right]\right)^{\frac{1}{p}},$$

for any $p \geq 1$.

The above definitions can be extended to Hilbert-valued random variables. Consider the family $\mathcal{S}_G(V)$ of V-valued smooth random variables of the form

$$F = \sum_{i=1}^{n} F_i v_i \quad (v_i \in V, F_i \in \mathcal{S}_G).$$

Define $D^G F := \sum_{i=1}^{n} D^G F_i \otimes v_j$. Then D^G is a closable operator from $\mathcal{S}_G(V)$ into $L^p(\Omega; G \otimes V)$ for any $p \geq 1$. Therefore, D^G is a closed operator on $\mathbb{D}_G^{1,p}(V) = \overline{\mathcal{S}_G(V)}^{\|\cdot\|_{1,p,V}}$ for the seminorm $\|\cdot\|_{1,p}$, defined by

$$\|F\|_{1,p,V} := \left(\mathbb{E}\left[\|F\|_V^p\right] + \mathbb{E}\left[\|D^G F\|_{G\otimes V}^p\right]\right)^{\frac{1}{p}},$$

on $\mathcal{S}_G(V)$. In particular, we define $\mathbb{D}_G^{1,\infty}$ and $\mathbb{D}_{\mathbb{G}}{}^{1,\infty}(V)$ by

$$\mathbb{D}_G^{1,\infty} := \bigcap_{p=1}^{\infty} \mathbb{D}_G^{1,p}, \quad \mathbb{D}_{\mathbb{G}}{}^{1,\infty}(V) := \bigcap_{p=1}^{\infty} \mathbb{D}_{\mathbb{G}}{}^{1,p}(V).$$

The following proposition is the chain rule for D^G.

Proposition 20.1 *Suppose that $F = (F^1, \ldots, F^m)$ is a random vector whose components belong to $\mathbb{D}_{\mathbb{G}}{}^{1,\infty}$. Let $f \in C_p^\infty(\mathbb{R}^m)$. Then $f(F) \in \mathbb{D}_{\mathbb{G}}{}^{1,\infty}$, and*

$$D^G(f(F)) = \sum_{i=1}^{m} \partial_i f(F) D^G F^i.$$

Next, we consider the divergence operator.

Definition 20.4 The divergence operator δ^G is an unbounded operator on $L^2(\Omega; G)$ with values in $L^2(\Omega)$ such that:

(1) The domain of δ^G, denoted by Domδ, is the set of stochastic processes $u \in L^2(\Omega; G)$ such that

$$\left|\mathbb{E}\left[\left\langle D^G F, u.\right\rangle_G\right]\right| \leq c(u)\|F\|_{L^2(\Omega)},$$

for any $F \in \mathbb{D}_G^{1,2}$, where $c(u)$ is some constant depending on u.

(2) If $u.$ belongs to Domδ^G, then $\delta^G(u)$ is characterized by

$$\mathbb{E}\left[F\delta^G(u)\right] = \mathbb{E}\left[\left\langle D^G F, u\right\rangle\right],$$

for any $F \in \mathbb{D}_G^{1,2}$.

The following proposition allows us to factor out a scalar random variable in a divergence.

Proposition 20.2 *Let $F \in \mathbb{D}_G^{1,2}$ and $u \in$ Dom δ^G such that $Fu \in L^2(\Omega; G)$. Then $Fu \in$ Dom δ^G and*

$$\delta^G(Fu) = F\delta^G(u) - \left\langle D^G F, u\right\rangle_G .$$

20.2.2.2 Malliavin Calculus for the Fractional Brownian Motion

We introduce the fractional Brownian motion and the Hilbert space associated with the fractional Brownian motion.

Definition 20.5 A vented Gaussian process $(W_t^H)_{t\geq 0}$ is called a fractional Brownian motion of the Hurst index $H \in (0, 1)$ if it has the covariance function

$$R_H(t, s) := \mathbb{E}[W_t^H W_s^H] = \frac{1}{2}(s^{2H} + t^{2H} - |t - s|^{2H}).$$

In this paper, we will only use the fractional Brownian motions with the Hurst index $H > \frac{1}{2}$. We denote by $\mathcal{E}$ the set of step functions on [0, T]. Let $\mathcal{H}$ be the Hilbert space defined as the closure of $\mathcal{E}$ with respect to the scalar product

$$\left\langle 1_{[0,t]}, 1_{[0,s]}\right\rangle_{\mathcal{H}} = R_H(t, s),$$

which yields for any $\phi, \psi \in \mathcal{H}$,

$$\langle \psi, \phi\rangle_{\mathcal{H}} := H(2H - 1)\int_0^T \int_0^T |r - u|^{2H-2}\psi(r)\phi(u)drdu.$$

It is easy to see that the covariance of the fractional Brownian motion can be written as

$$\begin{aligned}\left\langle 1_{[0,t]}, 1_{[0,s]}\right\rangle_{\mathcal{H}} &= H(2H-1)\int_0^t \int_0^s |r-u|^{2H-2} du dr \\ &= R_H(t,s).\end{aligned}$$

Therefore, the fractional Brownian motion W^H can be expressed as $W_t^H = W(1_{[0,t]})$ for the isonormal Gaussian W associated with the Hilbert space $\mathcal{H}$. Consider the square integrable kernel

$$K_H(t,s) := c_H s^{\frac{1}{2}-H} \int_s^t (u-s)^{H-\frac{3}{2}} u^{H-\frac{1}{2}} du,$$

where $c_H = \left[\frac{H(2H-1)}{B(2-2H, H-\frac{1}{2})}\right]^{\frac{1}{2}}$ and $t > s$. Define the isometric function $K_H^* : \mathcal{E} \to L^2(0,T)$ by

$$(K_H^* \phi)(s) := \int_s^T \phi(t) \frac{\partial K_H}{\partial t}(t,s) dt.$$

Then, the operator K_H^* is an isometry between $\mathcal{E}$ and $L^2(0,T)$ that can be extended to the Hilbert space $\mathcal{H}$. The operator K_H^* can be expressed in terms of the fractional integrals, i.e.,

$$(K_H^* \phi)(s) = c_H \Gamma(H - \frac{1}{2}) s^{\frac{1}{2}-H} (I_{T-}^{H-\frac{1}{2}} \cdot^{H-\frac{1}{2}} \phi(\cdot))(s).$$

Finally, we consider the Malliavin calculus on $\mathcal{H}$. For the sake of simplicity, we will use the notation D^{W^H}, $\mathbb{D}_{W^H}^{1,p}$, and δ^{W^H} as the derivative operator, the domain of the derivative, and the divergence operator associated with the Hilbert space $\mathcal{H}$, respectively. In the sequel, we present two results on the derivative operator for fractional Brownian motion (for the proof, see [5, Lemma 3.2]).

Proposition 20.3 *For any* $F \in \mathbb{D}_{W_H}^{1,2} = \mathbb{D}_{L^2(0,T)}^{1,2}$

$$K_H^* D^{W_H} F = D^{L^2([0,T])} F.$$

Proposition 20.4 $\sup_{0 \le t \le T} (W_t^H - \theta t)$ *belongs to* $\mathbb{D}_{W_H}^{1,2}$ *and* $D_t^{W_H} \sup_{0 \le t \le T} (W_t^H - \theta t) = 1_{[0,\tau]}(t)$*, for any* $t \in [0,T]$*, where* τ *is the point where the supremum is attained.*

20.3 Statistical Problems

Suppose that we have the past surplus data in $[0, T_0]$-interval at discrete time points $\frac{[kT_0]}{n}$ $(k = 0, 1, \dots, n)$. Our goal is to estimate the finite-time ruin probability for each $T \in (0, \infty]$ from the discrete data $(X_{\frac{[kT_0]}{n}})_{k\in\{0,1,\dots,n\}}$.

20.3.1 Estimation of σ

Fix $0 < H < \frac{3}{4}$. We use the results in [4] to construct the estimator for the true value of σ by using a power variation of the order $p > 0$. We define the power variation of (X_t) (see [4]), as follows:

$$V_p^n(X)_t := \sum_{i=1}^{[nt]} \left| X_{\frac{i}{n}} - X_{\frac{i-1}{n}} \right|^p .$$

Let

$$\widehat{\sigma}_{t,n} := \left(\frac{V_p^n(X)_t}{c_p n^{1-pH} t} \right)^{\frac{1}{p}} \tag{20.6}$$

be an estimator $\hat{\sigma}_{t,n}$ of σ_0, where $c_p = \frac{2^{\frac{p}{2}} \Gamma\left(\frac{p+1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)}$. We obtain the asymptotic normality of $\widehat{\sigma}_{t,n}$ as in the following theorem.

Theorem 20.1 *Let $p > 1$ and $0 < H < \frac{3}{4}$. Then*

$$\sqrt{n} \left(\left(\widehat{\sigma}_{t,n}\right)^p - \sigma^p \right) \xrightarrow{\mathcal{L}} \frac{v_1 \sigma^p}{c_p} W_t \quad (n \to \infty),$$

in law in the space $\mathcal{D}[0, T_0]$ equipped with the Skorohod topology, where

$$v_1^2 = \mu_p + 2 \sum_{j \geq 1} \left(\gamma_p \left(\rho_H(j) \right) - \gamma_p(0) \right),$$

$$\mu_p = 2^p \left(\frac{1}{\sqrt{\pi}} \Gamma\left(p + \frac{1}{2} \right) - \frac{1}{\pi} \Gamma\left(\frac{p+1}{2} \right)^2 \right),$$

$$\gamma_p(x) = (1 - x^2)^{\frac{p+1}{2}} 2^p \sum_{k=0}^{\infty} \frac{(2x)^{2k}}{\pi (2k)!} \Gamma\left(\frac{p+1}{2} + k \right)^2,$$

$$\rho_H(n) = \frac{1}{2} \left((n+1)^{2H} + (n-1)^{2H} - 2n^{2H} \right),$$

and $(W_t)_{t\in[0,T_0]}$ *is a Brownian motion independent of the fractional Brownian motion* W_t^H.

Proof It is sufficient to prove that

$$n^{-\frac{1}{2}+pH} V_p^n(u+\sigma\theta T_0) \to 0 \quad (n\to\infty), \tag{20.7}$$

in probability, uniformly on $[0, T_0]$, by [4, p. 272]. Thus, we show that

$$n^{-\frac{1}{2}+pH} V_p^n(u+\sigma\theta T_0) = n^{-\frac{1}{2}+pH} \sum_{i=1}^{[nT_0]} \left| u+\sigma\theta\frac{i}{n} - \left(u+\sigma\theta\frac{i-1}{n}\right)\right|^p \to 0 \quad \text{a.s.,}$$

as $n\to\infty$. We note that $(1-H)p > \frac{1}{2}$, hence, we have

$$\begin{aligned} n^{-\frac{1}{2}+pH} \sum_{i=1}^{[nT_0]} \left| u+\sigma\theta\frac{i}{n} - \left(u+\sigma\theta\frac{i-1}{n}\right)\right|^p &= n^{-\frac{1}{2}+pH} \sum_{i=1}^{[nT_0]} \left|\sigma\theta(\frac{i}{n} - \frac{i-1}{n})\right|^p \\ &= n^{-\frac{1}{2}+pH} |\sigma\theta|^p \frac{[nT_0]}{n^p} \\ &\to 0, \end{aligned}$$

as $n\to\infty$ and this proof is completed. □

20.3.2 Simulation-Based Inference for $\Psi_\sigma(u, T)$

Using the estimator $\widehat{\sigma}_{t,n}$ of σ_0 given in (20.6), we can estimate $\Psi_\sigma(u, T)$ by

$$\widehat{\Psi}_{t,n}(u, T) := \Psi_{\widehat{\sigma}_{t,n}}(u, T),$$

and, with the aid of the delta method (c.f., [15, p. 374]), it follows that

$$\sqrt{n}\left(\widehat{\Psi}_{t,n}(u,T) - \Psi(u,T)\right) \to \partial_\sigma \Psi_{\sigma_0}(u,T)\frac{1}{p}\frac{v_1\sigma_0}{c_p} W_t \quad (n\to\infty),$$

in law in the space $\mathcal{D}[0, T_0]$ equipped with the Skorohod topology, if $\Psi_\sigma(u, T)$ is differentiable at the true volatility parameter σ. This leads us an α-confidence interval for $\Psi_\sigma(u, T)$ such as

$$I_\alpha(\Psi) := \left[\widehat{\Psi}_{T_0,n}(u,T) \pm \frac{z_{\alpha/2}}{\sqrt{n}} |\partial_\sigma \Psi_{\widehat{\sigma}_{T_0,n}}(u,T)| \frac{\sqrt{T_0} v_1 \widehat{\sigma}_{T_0,n}}{p c_p}\right], \tag{20.8}$$

where $[a \pm b]$ stands for the interval $[a - b, a + b]$ for $a, b > 0$, and z_α stands for the upper α-quantile of the standard normal distribution.

Now, the problem is to compute the following quantity:

$$\partial_\sigma^k \Psi_\sigma(u, T) = \left(\frac{\partial}{\partial \sigma}\right)^k \mathbb{P}\left(\inf_{0 \le t \le T} X_t^\sigma < 0\right), \quad k = 0, 1. \tag{20.9}$$

- $k = 0$: Since $\Psi_\sigma(u, T)$ does not have a closed expression, we will compute it by the Monte Carlo simulation for a given value of σ, that is, we generate sample paths of $X_t = u + \sigma\theta t - \sigma W_t^H$ for given σ, say $\left(X^{(k)}\right)_{k=1,2,\ldots,m}$ independent each other, to obtain

$$\widetilde{\Psi}(u, T) = \frac{1}{m}\sum_{k=1}^m \mathbf{1}_{\{\tau^{(k)} \le T\}}, \quad \tau^{(k)} := \inf\{t > 0 | X_t^{(k)} < 0\},$$

  which goes to the true $\Psi_\sigma(u, T)$ almost surely as $m \to \infty$ by the strong law of large numbers. However, the event of ruin in $[0, T]$ is often very rare and most of the indicators of summand will be zero, which may lead to an underestimation of the true value $\Psi_\sigma(u, T)$. Changing the measure $\mathbb{P}$ into a suitable one, more efficient sampling procedure (importance sampling) will be proposed.
- $k = 1$: Computing $\partial_\sigma \Psi_\sigma(u, T)$ is not straightforward because the integrand of the righthand side of (20.9) is not differentiable in σ, and we cannot differentiate it under the expectation sign $\mathbb{E}$. Moreover, computing numerically, e.g., for small $\epsilon > 0$,

$$\frac{\Psi_{\sigma+\epsilon}(u, T) - \Psi_\sigma(u, T)}{\epsilon} \quad \text{or} \quad \frac{\Psi_{\sigma+\epsilon}(u, T) - \Psi_{\sigma-\epsilon}(u, T)}{2\epsilon}, \tag{20.10}$$

  we have to compute $\Psi_{\sigma+\epsilon}$ and $\Psi_{\sigma-\epsilon}(u, T)$ separately, which usually takes much time. In addition, since the accuracy of the calculation in (20.10) depends on ε, the problem of determining the value of ε also arises. Importance sampling can give the fast convergence with the variance reduction.

20.4 Differentiability of Ψ_σ

Fix $H > \frac{1}{2}$. We discuss the differentiability of $\Psi_\sigma(u, T)$ w.r.t. σ after the ideas of [7, 16].

Theorem 20.2 *The finite-time ruin probability $\Psi_\sigma(u, T)$ is differentiable w.r.t. σ and*

$$\partial_\sigma \Psi_\sigma(u,T) = \mathbb{E}\left[1_{[\sigma \sup\limits_{0\le t\le T}(W_t^H-\theta t)<u]} \delta^{W^H}\left(\frac{u_A(\cdot)\sup\limits_{0\le t\le T}(W_t^H-\theta t)}{\sigma\int_0^T \psi(Y_t)dt} \right)\right], \tag{20.11}$$

where

$$\begin{aligned} u_A(t) &:= \frac{d_H}{c_H\Gamma(H-\frac{1}{2})} t^{\frac{1}{2}-H} D_{T-}^{H-\frac{1}{2}}\left[(\cdot)^{2H-1} D_{0+}^{H-\frac{1}{2}}\left((\cdot)^{\frac{1}{2}-H}\psi(Y_\cdot)\right)(\cdot)\right](t) \qquad (20.12)\\ &= \frac{d_H}{c_H B\left(H-\frac{1}{2},\frac{3}{2}-H\right)\Gamma\left(\frac{3}{2}-H\right)} t^{\frac{1}{2}-H} \\ &\times \left\{ \frac{1}{(T-t)^{H-\frac{1}{2}}}\left[\psi(Y_t) + \left(H-\frac{1}{2}\right) t^{2H-1}\int_0^t \frac{t^{\frac{1}{2}-H}\psi(Y_t) - s^{\frac{1}{2}-H}\psi(Y_s)}{(t-s)^{H+\frac{1}{2}}} ds \right]\right. \\ &+ \left(H-\frac{1}{2}\right)\int_t^T \left\{ \frac{\psi(Y_t) + \left(H-\frac{1}{2}\right) t^{2H-1}\int_0^t \frac{t^{\frac{1}{2}-H}\psi(Y_t)-u^{\frac{1}{2}-H}\psi(Y_u)}{(t-u)^{H+\frac{1}{2}}}du}{(s-t)^{H+\frac{1}{2}}} \right. \\ &- \left.\left. \frac{\psi(Y_s) + \left(H-\frac{1}{2}\right) s^{2H-1}\int_0^s \frac{s^{\frac{1}{2}-H}\psi(Y_s)-u^{\frac{1}{2}-H}\psi(Y_u)}{(s-u)^{H+\frac{1}{2}}}du}{(s-t)^{H+\frac{1}{2}}} \right\} ds \right\}, \end{aligned}$$

$$Y_t := 8\left(4\int_0^T\int_0^T \frac{|W_s^H-\theta s-(W_u^H-\theta u)|^r}{|s-u|^{m+2}} ds du \right)^{\frac{1}{r}} \frac{m+2}{m} t^{\frac{m}{r}}, \tag{20.13}$$

$$d_H = \left(c_H \Gamma\left(H-\frac{1}{2}\right)\right)^{-1}, \tag{20.14}$$

for any even integers r, m such that $rH > m+2$ *and* $\psi \in C_b^\infty(\mathbb{R}_+)$ *satisfies*

$$\psi(x) = \begin{cases} 1 & (x \le \frac{u}{2\sigma}) \\ 0 & (x > \frac{u}{\sigma}). \end{cases}$$

In the proof of (20.11), it suffices to show that

$$\begin{aligned} &\partial_\sigma \mathbb{E}\left[\phi(\sigma \sup_{0\le t\le T}(W_t^H-\theta t))\right] \\ &= \mathbb{E}\left[\phi\left(\sigma\sup_{0\le t\le T}(W_t^H-\theta t)\right)\delta^{W^H}\left(\frac{u_A(\cdot)\sup\limits_{0\le t\le T}(W_t^H-\theta t)}{\sigma\int_0^T\psi(Y_t)dt}\right)\right] \qquad (20.15) \end{aligned}$$

holds for a sufficiently smooth function $\phi : \mathbb{R}_+ \to \mathbb{R}$ instead of $1_{(-\infty,0)}(\cdot)$ in (20.11) from the density argument. In the following, we impose the following conditions on ϕ:

(1) $\phi \in C_b^\infty(\mathbb{R}_+)$.
(2) The function ϕ is constant on $[0, u]$.

We prove the following properties of Y_t.

Lemma 20.1 *For Y_t defined in (20.13), the following (i)–(iv) hold.*

(i) $|W_t^H - \theta t| \leq Y_t \quad (t \in [0, T])$.
(ii) $\psi(Y_t) \in \mathbb{D}_{W^H}^{1,\infty} \quad (t \in [0, T])$.
(iii) *There exists a function* $\alpha : \mathbb{R} \to \mathbb{R}_+$, *with* $\lim_{q\to\infty} \alpha(q) = \infty$, *such that, for any* $q \geq 1$, *one has:* $\forall t \in [0, T] \quad E[Y_t^q] \leq C_q t^{\alpha(q)}$.
(iv) $\left(\int_0^T \psi(Y_t)dt\right)^{-1} \in L^p(\Omega) \quad (p \geq 1)$.

Proof We can show (i) and (ii) similarly as in [7, proof of Lemma 2.1]. Note (iii) follows immediately from (ii). To prove (iv), it is sufficient to show that

$$\mathbb{P}\left(\int_0^T \psi(Y_t)dt \leq \varepsilon\right) = O(\varepsilon^p) \quad (\varepsilon \downarrow 0)$$

holds from [12, p. 133]. Since

$$\int_0^T \psi(Y_t)dt = \int_0^{\frac{\epsilon}{\sigma}} \psi(Y_t)dt + \int_{\frac{\epsilon}{\sigma}}^T \psi(Y_t)dt \geq \frac{\epsilon}{\sigma}$$

holds on $\left[\frac{u}{2\sigma} > Y_{\frac{\epsilon}{\sigma}}\right]$, we obtain

$$\left[\sigma \int_0^T \psi(Y_t)dt \leq \epsilon\right] \subset \left[\frac{u}{2\sigma} \leq Y_{\frac{\epsilon}{\sigma}}\right].$$

Therefore, for any $q \geq 1$ such that $\alpha(q) \geq p$ we have

$$\begin{aligned}
\mathbb{P}\left(\int_0^T \sigma\psi(Y_t)dt < \epsilon\right) &\leq \mathbb{P}\left(\frac{u}{2\sigma} \leq Y_{\frac{\epsilon}{\sigma}}\right) \\
&\leq \left(\frac{u}{2\sigma}\right)^{-q} \mathbb{E}\left[Y_{\frac{\epsilon}{\sigma}}^q\right] \\
&\leq \left(\frac{u}{2\sigma}\right)^{-q} \left(\frac{\epsilon}{\sigma}\right)^{\alpha(q)} \\
&= O(\epsilon^p).
\end{aligned}$$

□

Proposition 20.5 *For Y_t defined in (20.13), we have*

$$\left(D_s^{W_H}\phi\left(\sigma \sup_{0\leq t\leq T}(W_t^H - \theta t)\right)\right)\psi(Y_t) = \phi'(\sigma \sup_{0\leq t\leq T}(W_t^H - \theta t))\sigma\psi(Y_t).$$

Proof The proof is analogous to the proof of [7]. Let

$$A := [0 \le \sigma \sup_{0\le t\le T} (W_t^H - \theta t) \le u].$$

Since $\phi'(\sigma \sup_{0\le t\le T} (W_t^H - \theta t)) = 0$ on A from the assumption of the function ϕ, we get

$$\begin{aligned}\left(D_s^{W_H}\phi\left(\sigma \sup_{0\le t\le T}(W_t^H - \theta t)\right)\right)\psi(Y_t) &= \phi'\left(\sigma \sup_{0\le t\le T}(W_t^H - \theta t)\right)\sigma 1_{[s\le\tau]}\psi(Y_t)\\ &= 0\\ &= \phi'\left(\sigma \sup_{0\le t\le T}(W_t^H - \theta t)\right)\sigma\psi(Y_t)\end{aligned}$$

on A. On the other hand, on $A^c = [\sigma \sup_{0\le t\le T} (W_t^H - \theta t) > u]$, for any $\omega \in A^c \cap [\phi(Y_t) \ne 0]$ we have

$$\sigma \sup_{0\le t\le T} (W_t^H(\omega) - \theta t) > u, \quad Y_t(\omega) < \frac{u}{\sigma}.$$

Therefore, we get

$$Y_t(\omega) < \frac{u}{\sigma} < \sup_{0\le t\le T} (W_t^H(\omega) - \theta t)) = W_\tau^H(\omega) - \theta\tau(\omega) \le Y_\tau(\omega),$$

so we have $t \le \tau$ since Y_t is an non-decreasing process. □

In the sequel, we consider the smoothness of u_A in the Malliavin sense. Let $\tilde{K}_H^*$ be the restriction of K_H^* to $L^2(0,T)$, and let $\tilde{K}_H^{*,adj}$ be the adjoint operator of $\tilde{K}_H^*$ in $L^2(0,T)$. Then, by [16], we have

$$\begin{aligned}&\left(\tilde{K}_H^{*,adj}\right)^{-1}(\psi(Y_\cdot)(t)\\ &= d_H t^{H-\frac{1}{2}} D_{0+}^{H-\frac{1}{2}} t^{\frac{1}{2}-H}\psi(Y_t)\\ &= \frac{d_H}{\Gamma(\frac{3}{2}-H)}\left(t^{\frac{1}{2}-H}\psi(Y_t) - \left(H-\frac{1}{2}\right)t^{H-\frac{1}{2}}\int_0^t \frac{t^{\frac{1}{2}-H}\psi(Y_t) - s^{\frac{1}{2}-H}\psi(Y_s)}{(t-s)^{H+\frac{1}{2}}}ds\right),\end{aligned}$$

and

$$\begin{aligned}
&\tilde{K}_H^{*-1}(u.)(t)\\
&= \frac{1}{c_H\Gamma\left(H-\frac{1}{2}\right)} t^{\frac{1}{2}-H} D_{T-}^{H-\frac{1}{2}}\left((\cdot)^{H-\frac{1}{2}}u(\cdot)\right)(t)\\
&= \frac{1}{c_H B\left(H-\frac{1}{2},\frac{3}{2}-H\right)}\left\{\frac{u_t}{(T-t)^{H-\frac{1}{2}}}+\left(H-\frac{1}{2}\right)t^{\frac{1}{2}-H}\int_t^T \frac{s^{H-\frac{1}{2}}u_s - t^{H-\frac{1}{2}}u_t}{(s-t)^{H+\frac{1}{2}}}ds\right\},
\end{aligned}$$

where $d_H = \left(c_H\Gamma\left(H-\frac{1}{2}\right)\right)^{-1}$. So we get $u_A = \tilde{K}_H^{*-1}\circ \tilde{K}_H^{*,adj-1}(\psi(Y.))$. Since $\tilde{K}_H^* : L^2(0,T) \to K_H^*[L^2(0,T)]$ is the isometric isomorphism, the following lemma holds.

Lemma 20.2 *For any stochastic process u_t, if*

$$u. \in \mathbb{D}^{1,p}_{K_H^*[L^2(0,T)]}(K_H^*[L^2(0,T)]),$$

then

$$\tilde{K}_H^{*-1}(u.) \in \mathbb{D}^{1,p}_{W^H}(\mathcal{H}).$$

Proof When $u. = Fv$ holds for $F \in \mathbb{D}^{1,p}_{K_H^*[L^2(0,T)]}$ and $v \in K_H^*[L^2(0,T)] \subset L^2(0,T)$, we have

$$\begin{aligned}
D^{W^H}(\tilde{K}_H^{*-1}(u.)) &= D^{W^H}(F\tilde{K}_H^{*-1}(v.))\\
&= D^{W^H}F \otimes \tilde{K}_H^{*-1}(v)\\
&= \tilde{K}_H^{*-1}(D^{K_H^*[L^2(0,T)]}F)\otimes \tilde{K}_H^{*-1}(v)\\
&= \tilde{K}_H^{*-1}\otimes\tilde{K}_H^{*-1}(D^{K_H^*[L^2(0,T)]}F\otimes v)\\
&= \tilde{K}_H^{*-1}\otimes\tilde{K}_H^{*-1}(D^{K_H^*[L^2(0,T)]}u.).
\end{aligned}$$

When $u \in \mathbb{D}^{1,p}_{K_H^*[L^2(0,T)]}(K_H^*[L^2(0,T)])$, there exists $u_n \in S_{K_H^*[L^2(0,T)]}(K_H^*[L^2(0,T)])$ such that $u_n \to u\ in\ \mathbb{D}^{1,p}_{K_H^*[L^2(0,T)]}(K_H^*[L^2(0,T)])$, and we have

$$\tilde{K}_H^{*-1}u_n \to \tilde{K}_H^{*-1}u\ in\ L^p(\Omega;\mathcal{H}),$$

since $\tilde{K}_H^*$ is the isometric isomorphism. Thus we have

$$\begin{aligned}
&\left\|D^{W^H}(\tilde{K}_H^{*-1}(u_n)) - \tilde{K}_H^{*-1}\otimes\tilde{K}_H^{*-1}(D^W(u)\right\|_{L^p(\Omega;\mathcal{H}\otimes\mathcal{H})}\\
&= \left\|\tilde{K}_H^{*-1}\otimes\tilde{K}_H^{*-1}(D^{K_H^*[L^2(0,T)]}u_n - D^{K_H^*[L^2(0,T)]}u)\right\|_{L^p(\Omega;\mathcal{H}\otimes\mathcal{H})}\\
&\le \left\|D^{K_H^*[L^2(0,T)]}u_n - D^{K_H^*[L^2(0,T)]}u\right\|_{L^p(\Omega;L^2(0,T)\otimes L^2(0,T))}\\
&\to 0,
\end{aligned}$$

and so we obtain $K_H^{*-1}(u_\cdot) \in \mathbb{D}^{1,p}_{W^H}(\mathcal{H})$. □

Proposition 20.6 *For u_A given in (20.12),*

$$u_A \in \mathbb{D}^{1,\infty}_{W^H}(\mathcal{H}).$$

Proof It suffices to show that $\int_0^t \frac{t^{\frac{1}{2}-H}\psi(Y_t)-s^{\frac{1}{2}-H}\psi(Y_s)}{(t-s)^{H+\frac{1}{2}}}ds \in \mathbb{D}^{1,\infty}_{W^H}\left(\tilde{K}_H^*\left(L^2(0,T)\right)\right)$ from Lemma 20.2 and $\psi(Y_t) \in \mathbb{D}^{1,p}_{W^H}(\mathcal{H})$. Let

$$a_n := \sum_{i=0}^{n-1} \frac{t^{\frac{1}{2}-H}\psi(Y_t) - \left(\frac{t}{2}+\frac{it}{2n}\right)^{\frac{1}{2}-H}\psi(Y_{\frac{t}{2}\frac{it}{2n}})}{\left(t-\left(\frac{t}{2}+\frac{it}{2n}\right)\right)^{H+\frac{1}{2}}}\frac{t}{2n},$$

$$b_n := \sum_{i=0}^{n-1} \frac{t^{\frac{1}{2}-H}\psi(Y_t) - \left(\frac{it}{2n}\right)^{\frac{1}{2}-H}\psi(Y_{\frac{it}{2n}})}{\left(t-\left(\frac{it}{2n}\right)\right)^{H+\frac{1}{2}}}\frac{t}{2n}.$$

We can show

$$a_n \to \int_{\frac{t}{2}}^t \frac{t^{\frac{1}{2}-H}\psi(Y_t) - s^{\frac{1}{2}-H}\psi(Y_s)}{(t-s)^{H+\frac{1}{2}}}ds \quad \text{a.s.},$$

$$b_n \to \int_0^{\frac{t}{2}} \frac{t^{\frac{1}{2}-H}\psi(Y_t) - s^{\frac{1}{2}-H}\psi(Y_s)}{(t-s)^{H+\frac{1}{2}}}ds \quad \text{a.s.},$$

as $n \to \infty$. We first show that

$$a_n \xrightarrow{L^p(\Omega)} \int_{\frac{t}{2}}^t \frac{t^{\frac{1}{2}-H}\psi(Y_t) - s^{\frac{1}{2}-H}\psi(Y_s)}{(t-s)^{H+\frac{1}{2}}}ds \quad (n \to \infty),$$

and that $D^{W_H}a_n$ converges in $L^p(\Omega)$. Defining

$$A_T := 8\left(4\int_0^T\int_0^T \frac{|W_s^H - \theta s - (W_u^H - \theta u)|^r}{|s-u|^{m+2}}dsdu\right)^{\frac{1}{r}}\frac{m+2}{m},$$

we have

$$|a_n| \le \left|\sum_{i=0}^{n-1} \frac{t^{\frac{1}{2}-H}\psi(Y_t) - \left(\frac{t}{2}+\frac{it}{2n}\right)^{\frac{1}{2}-H}\psi(Y_t)}{\left(t-\left(\frac{t}{2}+\frac{it}{2n}\right)\right)^{H+\frac{1}{2}}}\frac{t}{2n}\right|$$
$$+\left|\sum_{i=0}^{n-1} \frac{\left(\frac{t}{2}+\frac{it}{2n}\right)^{\frac{1}{2}-H}\psi(Y_t) - \left(\frac{t}{2}+\frac{it}{2n}\right)^{\frac{1}{2}-H}\psi(Y_{\frac{t}{2}+\frac{it}{2n}})}{\left(t-\left(\frac{t}{2}+\frac{it}{2n}\right)\right)^{H+\frac{1}{2}}}\frac{t}{2n}\right|$$

$$
\begin{aligned}
&\lesssim \left| \sum_{i=0}^{n-1} \frac{\left(\frac{t}{2}\right)^{-\frac{1}{2}-H} \psi(Y_t)}{\left(t-\left(\frac{t}{2}+\frac{it}{2n}\right)\right)^{H-\frac{1}{2}}} \frac{t}{2n} \right| + \left| \sum_{i=0}^{n-1} \frac{\left(\frac{t}{2}\right)^{\frac{m}{r}-\frac{1}{2}-H} A_T}{\left(t-\left(\frac{t}{2}+\frac{it}{2n}\right)\right)^{H-\frac{1}{2}}} \frac{t}{2n} \right| \\
&\lesssim \left| \sum_{i=0}^{n-1} \frac{\psi(Y_t)}{\left(t-\left(\frac{t}{2}+\frac{it}{2n}\right)\right)^{H-\frac{1}{2}}} \frac{t}{2n} \right| + \left| \sum_{i=0}^{n-1} \frac{A_T}{\left(t-\left(\frac{t}{2}+\frac{it}{2n}\right)\right)^{H-\frac{1}{2}}} \frac{t}{2n} \right| \\
&=: f_n^1(t) + f_n^2(t).
\end{aligned}
$$

Note that $\left(f_n^1(t)\right)^p \xrightarrow{L^1(\Omega)} \left(\int_{\frac{t}{2}}^t \frac{\psi(Y_t)}{(t-s)^{H-\frac{1}{2}}} ds\right)^p$. Indeed, we have

$$
\begin{aligned}
&\left\| f_n^1(t) - \int_{\frac{t}{2}}^t \frac{\psi(Y_t)}{(t-s)^{H-\frac{1}{2}}} ds \right\|_{L^p(\Omega)} \\
&\leq \|\psi(Y_t)\|_{L^p(\Omega)} \left| \sum_{i=0}^{n-1} \frac{1}{\left(t-\left(\frac{t}{2}+\frac{it}{2n}\right)\right)^{H-\frac{1}{2}}} \frac{t}{2n} - \int_{\frac{t}{2}}^t \frac{1}{(t-s)^{H-\frac{1}{2}}} ds \right| \\
&\to 0 \quad (n \to \infty).
\end{aligned}
$$

Similarly, $\left(f_n^2\right)^p \xrightarrow{L^1(\Omega)} \left(\int_{\frac{t}{2}}^t \frac{A_T}{(t-s)^{H-\frac{1}{2}}} ds\right)^p$, so we obtain

$$
a_n \xrightarrow{L^p(\Omega)} \int_{\frac{t}{2}}^t \frac{t^{\frac{1}{2}-H}\psi(Y_t) - s^{\frac{1}{2}-H}\psi(Y_s)}{(t-s)^{H+\frac{1}{2}}} ds \quad (n \to \infty). \tag{20.16}
$$

Also, since

$$
D^{W^H} a_n = \sum_{i=0}^{n-1} \frac{D^{W^H} A_T \left(t^{\frac{1}{2}+\frac{m}{r}-H} \psi'(Y_t) - \left(\frac{t}{2}+\frac{it}{2n}\right)^{\frac{1}{2}+\frac{m}{r}-H} \psi'(Y_{\frac{t}{2}+\frac{it}{2n}})\right)}{\left(t-\left(\frac{t}{2}+\frac{it}{2n}\right)\right)^{H+\frac{1}{2}}} \frac{t}{2n},
$$

we have

$$
\begin{aligned}
\left| D^{W^H} a_n \right| &\lesssim \left| \sum_{i=0}^{n-1} \frac{D^{W^H} A_T \left(\frac{t}{2}\right)^{\frac{m}{r}-H-\frac{1}{2}} \psi'(Y_t)}{\left(t-\left(\frac{t}{2}+\frac{it}{2n}\right)\right)^{H-\frac{1}{2}}} \right| + \left| \sum_{i=0}^{n-1} \frac{D^{W^H} A_T \left(\frac{t}{2}\right)^{2\frac{m}{r}-H-\frac{1}{2}} A_T}{(t-s)^{H-\frac{1}{2}}} \right| \\
&\lesssim \left| \sum_{i=0}^{n-1} \frac{D^{W^H} A_T \psi'(Y_t)}{\left(t-\left(\frac{t}{2}+\frac{it}{2n}\right)\right)^{H-\frac{1}{2}}} \right| + \left| \sum_{i=0}^{n-1} \frac{A_T D^{W^H} A_T}{(t-s)^{H-\frac{1}{2}}} \right|.
\end{aligned}
$$

Thus we get

$$D^{W^H} a_n \xrightarrow{L^p(\Omega)} \int_{\frac{t}{2}}^{t} \frac{D^{W^H} A_T \left(t^{\frac{m}{r}+\frac{1}{2}-H} \psi'(Y_t) - s^{\frac{m}{r}+\frac{1}{2}-H} \psi'(Y_s) \right)}{(t-s)^{H+\frac{1}{2}}} ds \quad (n \to \infty)$$

in the same way as in (20.16). Second, we show that

$$b_n \xrightarrow{L^p(\Omega)} \int_0^{\frac{t}{2}} \frac{t^{\frac{1}{2}-H} \psi(Y_t) - s^{\frac{1}{2}-H} \psi(Y_s)}{(t-s)^{H+\frac{1}{2}}} ds \quad (n \to \infty),$$

and that $D^{W_H} b_n$ converges in $L^p(\Omega)$. Now, we have

$$\begin{aligned} |b_n| &\le \left(\frac{2}{t}\right)^{H+\frac{1}{2}} \left\{ \sum_{i=0}^{n-1} \left| t^{\frac{1}{2}-H} \psi(Y_t) - \left(\frac{it}{2n}\right)^{\frac{1}{2}-H} \psi(Y_{\frac{it}{2n}}) \right| \right\} \frac{t}{2n} \\ &\lesssim \left(\frac{2}{t}\right)^{H+\frac{1}{2}} \left\{ t^{\frac{1}{2}-H} \psi(Y_t) + \sum_{i=0}^{n} \left(\frac{it}{2n}\right)^{\frac{m}{r}+\frac{1}{2}-H} A_T \frac{t}{2n} \right\} \\ &=: \left(\frac{2}{t}\right)^{H+\frac{1}{2}} f_n^3(t), \end{aligned} \tag{20.17}$$

where

$$\begin{aligned} &\left\| f_n^3(t) - \int_0^{\frac{t}{2}} \left(t^{\frac{1}{2}-H} \psi(Y_t) - s^{\frac{1}{2}-H} A_T \right) ds \right\|_{L^p(\Omega)} \\ &\lesssim \|A_T\|_{L^p(\Omega)} \left| \sum_{i=0}^{n-1} \left(\frac{it}{2n}\right)^{\frac{m}{r}+\frac{1}{2}-H} \frac{t}{2n} - \int_0^{\frac{t}{2}} s^{\frac{m}{r}+\frac{1}{2}-H} ds \right| \\ &\to 0 \quad (n \to \infty). \end{aligned}$$

Thus we have $b_n \xrightarrow{L^p(\Omega)} \int_0^{\frac{t}{2}} \frac{t^{\frac{1}{2}-H}\psi(Y_t) - s^{\frac{1}{2}-H}\psi(Y_s)}{(t-s)^{H+\frac{1}{2}}} ds$. From (20.17), we obtain

$$D^{W^H} b_n \xrightarrow{L^p(\Omega)} \int_0^{\frac{t}{2}} \frac{D^{W^H} A_T \left(t^{\frac{m}{r}+\frac{1}{2}-H} \psi'(Y_t) - s^{\frac{m}{r}+\frac{1}{2}-H} \psi'(Y_s) \right)}{(t-s)^{H+\frac{1}{2}}} ds \quad (n \to \infty)$$

in the same way as for a_n. Therefore, since D^{W_H} is a closable operator, we have

$$\int_0^t \frac{t^{\frac{1}{2}-H} \psi(Y_t) - s^{\frac{1}{2}-H} \psi(Y_s)}{(t-s)^{H+\frac{1}{2}}} ds \in \mathbb{D}^{1,\infty}_{W^H}.$$

$\square$

***Proof of Theorem* 20.2** From Proposition 20.1 we have

$$
\begin{aligned}
&\phi'(\sigma \sup_{0\le t\le T}(W_t^H-\theta t))\sigma \left\langle 1_{[0,\tau]}, u_A\right\rangle_{\mathcal{H}} \\
&= \phi'(\sigma \sup_{0\le t\le T}(W_t^H-\theta t))\sigma \left\langle \tilde{K}_H^*(1_{[0,\tau]}), \tilde{K}_H^{*,adj-1}(\psi(Y_\cdot))\right\rangle_{L^2(0,T)} \\
&= \phi'(\sigma \sup_{0\le t\le T}(W_t^H-\theta t))\sigma \left\langle 1_{[0,\tau]}, \psi(Y_\cdot)\right\rangle_{L^2(0,T)} \\
&= \phi'(\sigma \sup_{0\le t\le T}(W_t^H-\theta t))\sigma \int_0^T \psi(Y_t)dt.
\end{aligned}
$$

Thus, since

$$
\begin{aligned}
&\partial_\sigma \mathbb{E}\left[\phi\left(\sigma \sup_{0\le t\le T}(W_t^H-\theta t)\right)\right] \\
&= \mathbb{E}\left[\phi'(\sigma \sup_{0\le t\le T}(W_t^H-\theta t)) \sup_{0\le t\le T}(W_t^H-\theta t)\right] \\
&= \mathbb{E}\left[\phi'(\sigma \sup_{0\le t\le T}(W_t^H-\theta t)) \sup_{0\le t\le T}(W_t^H-\theta t)\frac{\int_0^T \psi(Y_t)dt}{\int_0^T \psi(Y_t)dt}\right] \\
&= \mathbb{E}\left[\phi'(\sigma \sup_{0\le t\le T}(W_t^H-\theta t))\sigma \sup_{0\le t\le T}(W_t^H-\theta t)\frac{\left\langle 1_{[0,\tau]}, u_A(\cdot)\right\rangle_{\mathcal{H}}}{\sigma \int_0^T \psi(Y_t)dt}\right] \\
&= \mathbb{E}\left[\left\langle D^{W^H}(\phi(\sigma \sup_{0\le t\le T}(W_t^H-\theta t))), \frac{u_A(\cdot) \sup_{0\le t\le T}(W_t^H-\theta t)}{\sigma \int_0^T \psi(Y_t)dt}\right\rangle_{\mathcal{H}}\right],
\end{aligned}
$$

and $\frac{u_A(\cdot) \sup_{0\le t\le T}(W_t^H-\theta t)}{\sigma \int_0^T \psi(Y_t)dt} \in \mathrm{Dom}\delta^{W^H}$, we get (20.15). Next we show that (20.11) holds. Defining $g_n : \mathbb{R}_+ \to \mathbb{R}$ by

$$
g_n := 1_{[u+\frac{1}{n}, u+n+\frac{1}{n}]} * \rho_{\frac{1}{n}} \quad (n \ge 2,\ \rho : \text{molifier}),
$$

then we have $g_n \in C_b^\infty(\mathbb{R})$ and $g_n(x) = 0$ on $[0, u]$. Also we get $g_n \to 1_{(u,\infty)}$ as $n \to \infty$ and

$$
\begin{aligned}
f_n(\sigma) &:= \mathbb{E}[g_n(\sigma \sup_{0\le t\le T}(W_t^H-\theta t))] \\
&\to \mathbb{E}[1_{(u,\infty)}(\sigma \sup_{0\le t\le T}(W_t^H-\theta t))] \\
&= \mathbb{E}[1_{[u,\infty)}(\sigma \sup_{0\le t\le T}(W_t^H-\theta t))].
\end{aligned}
$$

Therefore, for any compact set $K \subset \mathbb{R}_+$ and $V_T^* := \sup_{0 \le t \le T} (W_t^H - \theta t)$, we have

$$\begin{aligned}
&\left| \sup_{\sigma \in K} \left(\partial_\sigma f_n(\sigma) - \mathbb{E}\left[1_{[u,\infty)}(\sigma \sup_{0\le t\le T}(W_t^H - \theta t)) \delta^{W^H} \left(\frac{u_A(\cdot) \sup_{0\le t\le T}(W_t^H - \theta t)}{\int_0^T \sigma \psi(Y_t) dt} \right) \right] \right) \right| \\
&\le \frac{1}{\inf_{\sigma \in K} \sigma} \left\| \delta^{W^H} \left(\frac{u_A(\cdot) \sup_{0\le t\le T}(W_t^H - \theta t)}{\int_0^T \sigma \psi(Y_t) dt} \right) \right\|_{L^2(\Omega)} \sup_{\sigma\in K} \mathbb{E}\left[\left(g_n(\sigma V_T^*) - 1_{[u,\infty)}(\sigma V_T^*) \right)^2 \right]^{\frac{1}{2}} \\
&\lesssim \sup_{\sigma \in K} \left\{ \mathbb{P}\left(\sigma V_T^* \in [u, u + \frac{2}{n}] \right) + \mathbb{P}\left(\sigma V_T^* \in [u+n, \infty) \right) \right\} \\
&=: \sup_{\sigma \in K} \left\{ h_n^1(\sigma) + h_n^2(\sigma) \right\}. \qquad (20.18)
\end{aligned}$$

In considering the compact uniform convergence of $h_n^1(\sigma)$ and $h_n^2(\sigma)$ w.r.t. σ, it suffices to show only the continuity of $h_n^1(\sigma)$ and $h_n^2(\sigma)$ w.r.t. σ by the Dini theorem. The problem here is that when σ changes, the interval $\left[\frac{u}{\sigma}, \frac{u+\frac{2}{n}}{\sigma}\right]$ also moves, so the continuity of the measure $\mathbb{P}$ cannot be exploited. Therefore, for any $(\sigma_m) \subset K$ such that $\sigma_m \downarrow \sigma\,(m \to \infty)$, we shall show that for any $m \in \mathbb{N}$ there exists a fixed point that in $\left[\frac{u}{\sigma_m}, \frac{u+\frac{2}{n}}{\sigma_m}\right]$. For $(\sigma_m) \in K$ such that $\sigma_m \downarrow \sigma\,(m \to \infty)$, take $\varepsilon > 0$ satisfying $\varepsilon < \frac{\frac{2}{n\sigma}}{2u - \frac{2}{n}}$ and $N \in \mathbb{N}$ large enough to satisfy $\left|\frac{1}{\sigma_N} - \frac{1}{\sigma}\right| < \varepsilon$. Then,

$$\frac{u}{\sigma_N} < u\left(\frac{1}{\sigma} + \varepsilon\right) < \left(u + \frac{2}{n}\right)\left(\frac{1}{\sigma} - \varepsilon\right) < \frac{u + \frac{2}{n}}{\sigma_N},$$

since $\left(2u + \frac{2}{n}\right)\varepsilon < \frac{2}{n\sigma}$ and $\left|\frac{1}{\sigma_N} - \frac{1}{\sigma}\right| < \varepsilon$. Therefore, we can show that $h_n^1(\sigma)$ is right-continuous w.r.t. σ because we can take $a := u\left(\frac{1}{\sigma} + \varepsilon\right)$ independent of $N \in \mathbb{N}$ such that $\frac{u}{\sigma_N} < a < \frac{u+\frac{2}{n}}{\sigma_N}$. In the same way, we can show the case of $\sigma_m \uparrow \sigma\,(m \to \infty)$, so we obtain the continuity of $h_n^1(\sigma)$ and $h_n^2(\sigma)$. Thus, (20.18) converges to 0 as $n \to \infty$, and this proof is complete. □

Acknowledgements The authors express sincere thanks to Prof. A. Kohats-Higa for the valuable discussion related to the part of Malliavin calculus. This research was partially supported by JSPS KAKENHI Grant-in-Aid for Scientific Research (C) #21K03358.

References

1. ASMUSSEN, S. AND ALBRECHER, H. (2010). *Ruin Probabilities*. 2nd ed. World Scientific Publishing Co. Pte. Ltd., Hackensack, NJ.
2. BILLINGSLEY, P. (1999). *Convergence of Probability Measures*. 2nd ed. John Wiley & Sons, New York.
3. CAI, C. AND XIAO, W. (2021). Simulation of an integro-differential equation and application in estimation of ruin probability with mixed fractional Brownian motion. *J. Integral Equations Applications* **33** 1–17.
4. CORCUERA, J. M., NUALART, D AND WOERNER, J.H.C. (2006). Power variation of some integral fractional processes. *Bernoulli* **12**, 713–735.
5. FLORIT, C. AND NUALART, D. (1995). A local criterion for smoothness of densities and application to the supremum of the Brownian sheet. *Statistics and Probability Letters* **22** 25–31.
6. GERBER, H.U., SHIU, E.S.W. (1998). On the time value of ruin. *N. Am. Actuar.* **2** 48–72.
7. GOBET, E. AND KOHATSU- HIGA, A. (2003). Computation of Greeks for barrier and look-back options using Malliavin calculus. *Electron. Comm. Probab* **8** 51–62.
8. JI, L. AND ROBERT, R. (2018). Ruin problem of a two-dimensional fractional Brownian motion risk process. *Stoch. Models* **34** 73–97.
9. LUNDBERG, F. (1903). Approximerad framställning av sannolikhetsftmktionen. Aterforsäknng av kollektivrisker. *Akad Afhandling. Almqvist och Wiksell, Uppsala.*
10. MICHNA, Z. (1998). Self-similar processes in collective risk theory, *J. Appl. Math. Stochastic Anal.* **11** 429–448.
11. NORROS, I., VALKEILA, E. AND VIRTAMO, J. (1999). An elementary approach to a Girsanov formula and other analytical results on fractional Brownian motion. *Bernoulli* **5** 571–587.
12. NUALART, D. (1995). *Malliavin Calculus and Related Topics*. (Probability and its Applications). Berlin Heidelberg New York Springer.
13. SAMKO, S. G., KILBAS, A. A. AND MARICHEV, O. I. (1994). *Fractional Integrals and Derivatives. Theory and Applications.* Gordon and Breach Science, Yverdon.
14. SHIMIZU, Y. (2021), *Asymptotic Statistics in Insurance Risk Theory.* Springer Briefs in Statistics.
15. VAN DER VAART, A. W. AND WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. With Applications to Statistics. Springer Series in Statistics. Springer-Verlag, New York.
16. ZAÏDI, N.L. AND NUALART, D. (2003). Smoothness of the law of the supremum of the fractional Brownian motion. *Electron. Comm. Probab.* **8** 102–111.

Chapter 21
Complex-Valued Time Series Models and Their Relations to Directional Statistics

Takayuki Shiohama

Abstract The fluctuation of stationary time series often shows a certain periodic behavior and this pattern is usually summarized via a spectral density. Since spectral density is a periodic function, it can be modeled by using a circular distribution function. In this paper, several time series models are studied in relation to a circular distribution. As an introduction, we illustrate how to model bivariate time series data using complex-valued time series in the context of circular distribution functions. These models are extended to have a skewed spectrum by incorporating a sine-skewing transformation. Two parameter estimation methods are considered and their asymptotic properties are investigated. These theoretical results are verified via a Monte Carlo simulation. Real data analyses illustrate the applicability of the proposed model.

Keywords Circular statistics · Maximum likelihood estimation · Sine-skewed model · Spectral density · Time series analysis

21.1 Introduction

The spectral density of a time series is a periodic function, and it is closely related to the circular density function used in the field of directional statistics. The relationship between time series spectra and circular density was investigated by [24]. The well-known circular distributions include a wrapped Cauchy, cardioid, and various extended families of circular distributions [12, 13] are related to the spectral density of AR(1), MA(1), AR(2), and ARMA(1, 1) models. The circular distributions have been characterized by the location and concentration parameters (μ, ρ), whereas the spectral densities of real-valued AR(1) and MA(1) models have peaks at frequencies of 0 or π, which indicates the location is restricted at $\mu = 0$ or $\mu = \pi$. A phase rotation or shift in spectral densities can be realized by multiplying $e^{i\mu}$ to the AR or MA coefficients of time series models; the resulting spectral density function has a

T. Shiohama (✉)
Nanzan University, 18 Yamazato-cho, Showa, Nagoya 466-8673, Japan
e-mail: shiohama@nanzan-u.ac.jp

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_21

mode or a location at μ. This fact has motivated us to consider complex-valued time series modeling.

In directional statistics, angular-valued time series models have been widely proposed for various geometric manifold structures such as a circle, torus, or hyper-torus. Some circular time series models have been considered in the literature; for example, the circular Markov models [2] and the higher-order circular Markov model [19]. In addition, these studies showed the relationship between a circular time series analysis and the time series spectral analysis. Unlike [24] who revisited the circular distributions from the point of view of time series spectral analysis, we focus on complex-valued time series modeling. That is, the observed process is on the complex plane when considering a stationary time series with complex-valued AR and/or MA coefficients.

Complex-valued random processes are conventional approaches for modeling phenomena of fluctuations, including waves in oceanography, electro-magnetics, and communications [21, 22]. A complex-valued random variable could be replaced by a pair of real random variables, and these real and imaginary components are assumed to be independent and identically distributed (i.i.d.). This assumption is said to be statistically proper. Early works of the proper complex statistical modeling include [8, Sect. 6.5] and [16, 18]. The maximum likelihood estimation of the proper complex-valued AR(1) was investigated by [14]. In contrast to proper complex-valued processes, the theory of improper complex-valued processes differs from the techniques of proper cases, see [20–22] for more details. Several approaches for modeling complex-valued random walk include [9, 11]. Recently, [7] studied complex-valued time series using a spectral density characterized by a generalized von Mises distribution.

We introduce several bivariate models using complex-valued AR models in the stationary time series; for example, bivariate random walks whose increments are complex-valued AR or MA models. We consider an exponential model because its spectral density function has a form of von Mises density on the circle. Further, we consider the complex-valued process whose spectral density is given by a sine-skewed circular density function [1].

The spectral analysis of a bivariate time series is performed to investigate the cross-spectrum and the spectra of both time series, which are summarized by a spectral density matrix. The asymmetric characteristics of the bivariate time series are summarized in this function because the cross-spectrum function involves complex-values; therefore, quantities such as the phase spectrum, gain spectrum, and coherency can provide useful frequency domain information about the data. In contrast, our complex-valued time series models do not have these cross-spectrum functions, and the asymmetric characteristics of the bivariate time series are summarized by a single asymmetric spectral density. Thus, we hope that this asymmetric spectrum provides useful information on the asymmetric characteristics of the bivariate process.

The complex-valued time series models introduced in this paper are applied to the two real datasets. Therefore, methods of moments estimator and Whittle estimator are considered, together with their asymptotic properties. Note that some time series models do not necessarily impose an identifiability assumption (see [17]).

The rest of this paper is organized as follows. Section 21.2 introduces some basic definitions for the complex-valued time series model. The sine-skewed extension of the complex-valued wrapped Cauchy and von Mises processes is also given. The methods for estimating unknown parameters are investigated in Sect. 21.3. Section 21.4 provides some Monte Carlo simulations. Section 21.5 analyzes two real datasets. Finally, this paper is summarized in Sect. 21.6.

21.2 Complex-Valued Time Series

A complex-valued random variable Z is a map from some probability space into the field of complex numbers whose real and imaginary parts are random variables. For complex-valued random variables $Z(= X + iY)$, Z_1 and Z_2, we define

$$E(Z) = E(X) + iE(Y), \quad \mathrm{Var}(Z) = E|Z - E(Z)|^2, \tag{21.1}$$
$$\mathrm{Cov}(Z_1, Z_2) = E(Z_1 - E(Z_1))\overline{(Z_2 - E(Z_2))}.$$

The following properties are useful for evaluating autocovariance functions:

$$\mathrm{Cov}(\alpha Z_1, \beta Z_2) = \alpha\overline{\beta}\mathrm{Cov}(Z_1, Z_2), \quad \mathrm{Cov}(Z_1, Z_2) = \overline{\mathrm{Cov}(Z_2, Z_1)}.$$

A complex-valued time series is a sequence of complex-valued random variables Z_t. For a stationary complex-valued process $\{Z_t\}$, the autocovariance function is defined by $\gamma_Z(h) = \mathrm{Cov}(Z_{t+h}, Z_t)$. Recall that the autocovariance function is a conjugate symmetric such that $\gamma_Z(h) = \overline{\gamma_Z(-h)}$. The autocorrelation function is defined by $\rho_Z(h) = \gamma_Z(h)/\gamma_Z(0)$, and the sample autocovariance function for a zero mean complex-valued process is defined by

$$\hat{\gamma}_Z(h) = \frac{1}{n - |h|} \sum_{t=1}^{n-|h|} Z_{t+h}\overline{Z_t}.$$

The sample autocorrelation function is defined by $\hat{\rho}_Z(h) = \hat{\gamma}_Z(h)/\hat{\gamma}_Z(0)$.

For the zero mean random variables X and Y with correlation coefficient ρ_{XY}, the variance given in (21.1) becomes

$$\mathrm{Var}(Z) = E|Z|^2 = E[(X + iY)(X - iY)] = \sigma_X^2 + \sigma_Y^2,$$

where σ_X^2 and σ_Y^2 are the variance of X and Y, respectively. Since this $\mathrm{Var}(Z)$ does not carry information about the correlation coefficient between the pairs (X, Y), we need another complex second-order moment as

$$\widetilde{\mathrm{Var}}(Z) = E(Z)^2 = E[(X + iY)^2] = \sigma_X^2 - \sigma_Y^2 + 2i\rho_{XY}\sigma_X\sigma_Y,$$

which is referred to as the complementary variance or pseudo-variance (see [18, 21]). If $\sigma_X^2 = \sigma_Y^2$ and $\rho_{XY} = 0$, the complementary variance becomes zero. This is the so-called proper or circular case; all others are called improper cases. For the improper case, the complementary autocovariance function is defined by $\tilde{\gamma}_Z(h) = E(Z_{t+h}Z_t)$. In order to investigate the complementary autocovariance structure of the process, we need an augmented process as $\boldsymbol{Z}_t = (Z_t, \overline{Z}_t)^T$, which is beyond the focus of this paper. Hereafter, we restrict our attention to the proper complex-valued process $\{Z_t\}$ and assume that the process has zero mean.

21.2.1 Wrapped Cauchy Process

We consider the complex-valued AR(1) process $\{Z_t\}$ defined by

$$Z_t = \rho e^{-i\mu} Z_{t-1} + \varepsilon_t, \tag{21.2}$$

where $\{\varepsilon_t\}$ is a complex-valued white noise with zero mean, variance $E|\varepsilon_t|^2 = \sigma_\varepsilon^2$ and $E(\varepsilon_t^2) = 0$. The parameters ρ and μ are the amplitude and argument of a complex-valued AR coefficient of order 1 that satisfy $0 \le \rho < 1$ and $\mu \in [0, 2\pi)$. Recall that for $\mu = 0$ and $\mu = \pi$, the AR coefficient corresponds to ρ and $-\rho$, respectively. The process $\{Z_t\}$ has the spectral density

$$f_Z(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi} \frac{1}{1 + \rho^2 - 2\rho\cos(\lambda - \mu)}. \tag{21.3}$$

In directional statistics, the spectral density (21.3) corresponds to the wrapped Cauchy distribution when $\sigma_\varepsilon^2 = 1 - \rho^2$ due to the normalization $\int_{-\pi}^{\pi} f_Z(\lambda)d\lambda = 1$.

21.2.2 von Mises Process

The paper [5] considered the spectrum of the time series has the form

$$f(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi} \exp\{\theta\cos(\lambda)\}.$$

This spectrum can be archived by taking the characteristic polynomial function of the MA process as $\alpha(z) = e^{\theta z}$. Then by setting $\theta = \kappa e^{i\mu}$, the corresponding spectral density becomes

$$f(\lambda) = \frac{\sigma_\varepsilon^2}{2\pi} \exp\{2\kappa\cos(\lambda - \mu)\},$$

which is the density function of the von Mises distribution on the circle when $\sigma_\varepsilon^2 = 1/I_0(2\kappa)$, where $I_p(\cdot)$ is the p-th order modified Bessel function of the first kind. Since the Taylor expansion for the $e^{\theta z}$ is given by $e^{\theta z} = \sum_{j=0}^{\infty} \frac{(\theta z)^j}{j!}$, this indicates that the von Mises process corresponds to the MA(∞) process that is satisfied with $|\theta| < 1$ for an invertibility condition. Hence, the von Mises process is given by

$$Z_t = \sum_{j=0}^{\infty} \frac{\left(\kappa e^{i\mu}\right)^j}{j!} \varepsilon_{t-j}, \tag{21.4}$$

where $\{\varepsilon_t\}$ is a complex-valued white noise with zero mean, variance $E|\varepsilon_t|^2 = \sigma_\varepsilon^2$ and $E(\varepsilon_t^2) = 0$. The model parameters are $\kappa \in (0, 1]$ and $\mu \in [0, 2\pi)$.

21.2.3 Sine-Skewed Process

The following filtered time series spectrum representation is useful to define a sine-skewed process. For this, we need to define a transfer function. The transfer function with a filter coefficient ψ_j is given by

$$\psi(\lambda) = \sum_{j=-\infty}^{\infty} \psi_j e^{ij\lambda}.$$

Let Z_t be a stationary time series with spectral density $f_Z(\lambda)$ and let $\sum_{j=-\infty}^{\infty} |\psi_j| < \infty$. Then $Y_t = \sum_{j=-\infty}^{\infty} \psi_j Z_{t-j}$ has a spectral density $f_Y(\lambda)$ defined as

$$f_Y(\lambda) = |\psi(\lambda)|^2 f_Z(\lambda).$$

Consequently, setting the transfer function as

$$\psi(\lambda) = \sqrt{1 + \eta \sin(\lambda - \mu)} = \sum_{j=-\infty}^{\infty} \psi_j e^{ij(\lambda-\mu)},$$

then, the time series $\{Y_t\}$ defined by $Y_t = \sum_j \psi_j \overline{Z_{t-j}}$ has the spectral density

$$f_Y(\lambda) = (1 + \eta \sin(\lambda - \mu)) f_Z(\lambda),$$

which corresponds to the sine-skewed wrapped Cauchy (SSWC) and sine-skewed von Mises (SSVM) density functions when $f_Z(\lambda - \mu)$ is the wrapped Cauchy and von Mises process, respectively. Denote $\boldsymbol{\varphi}^{(\mathrm{SSWC})} = (\mu, \rho, \eta)^T$ and $\boldsymbol{\varphi}^{(\mathrm{SSVM})} = (\mu, \kappa, \eta)^T$ by the parameter vectors of the SSWC and SSVM processes, respectively.

The following theorem states the autocorrelation structures for the SSWC processes.

Theorem 21.1 *We assume that* $Var(\varepsilon_t) = 1 - \rho^2$ *in model (21.2) and* $\psi(\lambda) = \sqrt{1 + \eta \sin(\lambda - \mu)}$. *Then, the SSWC process* $\{Y_t\}$ *has the following autocorrelation function:*

$$\gamma_Y^{(SSWC)}(h) = \rho_h^{(SSWC)} e^{ih\mu_h^{(SSWC)}}, \tag{21.5}$$

where $\rho_h^{(SSWC)} = \sqrt{\alpha_h^2(\boldsymbol{\varphi}^{(SSWC)}) + \beta_h^2(\boldsymbol{\varphi}^{(SSWC)})}$ *and* $\mu_h^{(SSWC)} = \arg(\alpha_h(\boldsymbol{\varphi}^{(SSWC)}) + i\beta_h(\boldsymbol{\varphi}^{(SSWC)}))$ *with*

$$\alpha_h(\boldsymbol{\varphi}^{(SSWC)}) = \cos(h\mu)\rho^{|h|} - \sin(h\mu)\frac{\eta}{2}(\rho^{|h-1|} - \rho^{|h+1|}),$$
$$\beta_h(\boldsymbol{\varphi}^{(SSWC)}) = \sin(h\mu)\rho^{|h|} + \cos(h\mu)\frac{\eta}{2}(\rho^{|h-1|} - \rho^{|h+1|}).$$

When $h = 1$, *it reduces to*

$$\gamma_Y^{(SSWC)}(1) = \sqrt{\rho^2 + \eta^2(1-\rho^2)^2/4}\{\cos\mu_1(\boldsymbol{\varphi}^{(SSWC)}) + i\sin\mu_1(\boldsymbol{\varphi}^{(SSWC)})\}.$$

The following theorem is the autocorrelation structures for the SSVM processes.

Theorem 21.2 *We assume that* $Var(\varepsilon_t) = 1/I_0(2\kappa)$ *in model (21.4) and* $\psi(\lambda) = \sqrt{1 + \eta \sin(\lambda - \mu)}$. *Then, the SSVM process* $\{Y_t\}$ *has the following autocorrelation function:*

$$\gamma_Y^{(SSVM)}(h) = \rho_h^{(SSVM)} e^{ih\mu_h^{(SSVM)}}, \tag{21.6}$$

where $\rho_h^{(SSVM)} = \sqrt{\alpha_h^2(\boldsymbol{\varphi}^{(SSVM)}) + \beta_h^2(\boldsymbol{\varphi}^{(SSVM)})}$ *and* $\mu_h(\boldsymbol{\varphi}) = \arg(\alpha_h(\boldsymbol{\varphi}^{(SSVM)}) + i\beta_h(\boldsymbol{\varphi}^{(SSVM)}))$ *with*

$$\alpha_h(\boldsymbol{\varphi}^{(SSVM)}) = \frac{I_h(2\kappa)}{I_0(2\kappa)}\left\{\cos(h\mu) - \frac{h\eta}{2\kappa}\sin(h\mu)\right\},$$
$$\beta_h(\boldsymbol{\varphi}^{(SSVM)}) = \frac{I_h(2\kappa)}{I_0(2\kappa)}\left\{\sin(h\mu) + \frac{h\eta}{2\kappa}\cos(h\mu)\right\}.$$

When $h = 1$, *it reduces to*

$$\gamma_Y^{(SSVM)}(1) = \frac{I_1(2\kappa)}{2\kappa I_0(2\kappa)}\sqrt{4\kappa^2 + \eta^2}\{\cos\mu_1(\boldsymbol{\varphi}^{(SSVM)}) + i\sin\mu_1(\boldsymbol{\varphi}^{(SSVM)})\}.$$

For the proof of Theorems 21.1 and 21.2, we refer the reader to [1, 17] for detailed derivations of the trigonometric moments of the sine-skewed circular distributions. Notice that the Fourier coefficients of the series expansion of the circular density

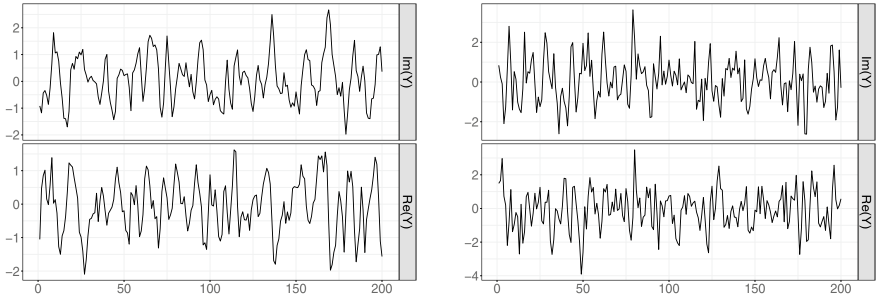


Fig. 21.1 Sample paths of the SSWC process (left panel) and the SSVM process (right panel). The skewness parameters are $\eta = 0.5$ and $\eta = -0.5$ for the SSWC and SSVM processes, respectively

function corresponds to the autocorrelation function of a time series with spectral density $f_Y(\lambda)$.

For the SSVM process, the absolute value of the autocorrelation at lag 1 is bounded above by $|\gamma_Y^{(SSVM)}(1)| \leq |I_1(2)/I_0(2)| \approx 0.6978$ when $\eta = 0$ due to the restriction $|\kappa e^{i\mu}| < 1$. This fact shows the limitation to fit a time series whose autocorrelation at lag 1 has strong dependent structures.

The simulated sample paths for the SSWC and SSVM processes are plotted in Fig. 21.1. For the SSWC models, we choose $\mu = \pi/6$, $\rho = 0.8$, and $\eta = 0.5$, and we set $\mu = -\pi/6$, $\kappa = 0.6$, and $\eta = 0.5$ for the SSVM process. A cyclical pattern in the wrapped Cauchy process was apparent; however, it was not confirmed for the von Mises processes. This was verified via the fact that the autocorrelation at lag 1 became $\gamma_Y^{(SSWC)}(1) = 0.648 + 0.478i$ and $\gamma_Y^{(SSVM)}(1) = 0.337 - 0.441i$, as apparent by the opposite sign for the imaginary part in the autocorrelation function.

The autocorrelation functions for the SSWC and SSVM processes are plotted in the complex plane in Fig. 21.2. The parameters are the same as those used in Fig. 21.1. We see the conjugate symmetric property of the complex-valued autocorrelation functions because the plots show symmetry with respect to the x axis. The wrapped Cauchy process shows strong persistence with fluctuational patterns.

The spectral density functions of the SSWC and SSVM processes are plotted in Fig. 21.3. The parameters μ, ρ, and κ are the same as those used in Figs. 21.1 and 21.2; the skewness parameters are changed based on $\eta = 0, 0.5, 0.9$ for the SSWC and $\eta = 0, -0.5, -0.9$ for the SSVM processes. As indicated by [1], the degree of skewness weakened when the concentration of the spectral distribution became large. This is the case for the SSWC process with $\rho = 0.8$, whereas for the SSVM cases, the larger η, the more apparent is the effect of the skewness as the location shifts to the left.

Finally, we plot the log periodogram of the simulated time series with its smoothed estimators. The true log-spectral density with the non-skewed spectral density is shown in Fig. 21.4. The length of the simulated sample path is 500, and we can see that the estimated smoothed periodograms are well-fitted to the true sine-skewed

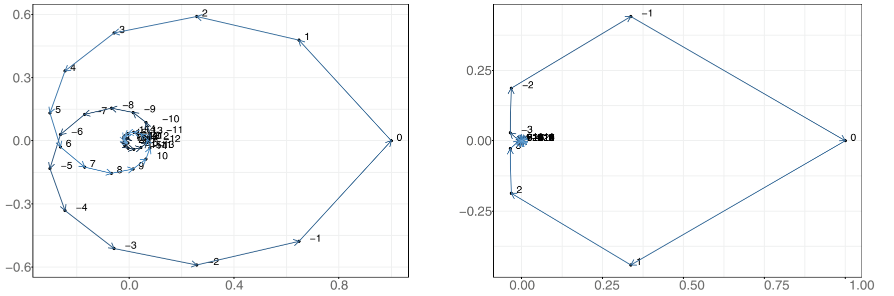


Fig. 21.2 Theoretical autocorrelation functions for the SSWC (left panel) and SSVM processes (right panel). The parameters are the same as those used in Fig. 21.1. The numbers in the plot represent the lag indicators

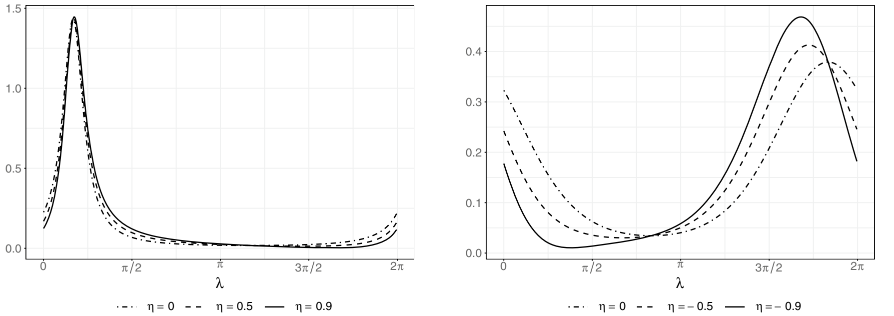


Fig. 21.3 Spectral density functions of the SSWC (left panel) and SSVM (right panel) processes. The skewness parameters are set as $\eta = 0, 0.5, 0.9$ for the SSWC and $\eta = 0, -0.5, -0.9$ for the SSVM process

spectral density; the symmetric spectral density fails to capture the structure of the skewness of the spectra.

21.2.4 Random Walks

The stochastic processes proposed so far can be extended to the random walk. We consider the complex-valued random walk plus complex-valued noise as follows:

$$X_t = X_{t-1} + Z_t,$$

where $\{Z_t\}$ represents the complex-valued processes defined so far. These models can provide useful information of the trajectories of the object in geographical science when considering the bivariate data as the complex plane. The applications can be

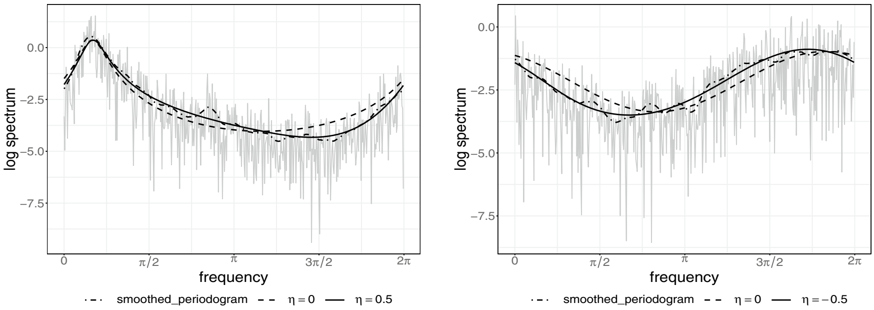


Fig. 21.4 Log spectral density functions of the SSWC (left panel) and SSVM (right panel) processes. The periodograms are computed using 500 samples of the simulated series

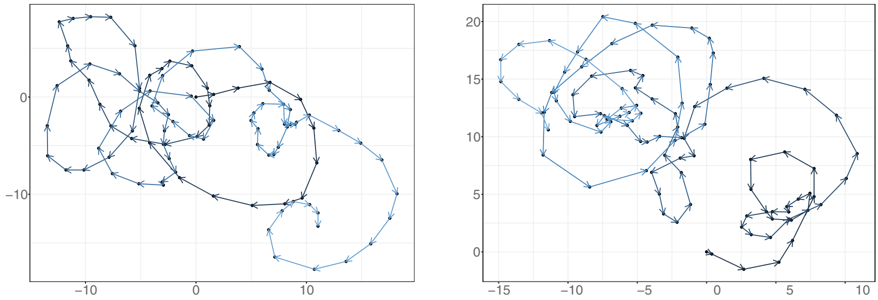


Fig. 21.5 Sample paths of the bivariate random walk with complex-valued wrapped Cauchy process with $\mu = \pi/6$ (left panel) and $\mu = -\pi/6$ (right panel)

found in various fields in natural science, environmental science, ecology, and so forth.

Figure 21.5 plots the trajectories of the complex-valued random walks with wrapped Cauchy noise with $\rho = 0.8$, $\mu = \pi/6$, and $\mu = -\pi/6$. The initial value of the process is set as $X_0 = (0, 0)^T$. These figures show that the parameter μ controls the direction of the movement; for $\mu > 0$, the sample trajectories have clockwise rotations and vice versa.

This specification indicates that the complex-valued local linear trend model is given as

$$X_t = m_t + \varepsilon_{1,t}, \quad m_{t+1} = m_t + v_t + \varepsilon_{2,t}, \quad v_{t+1} = \rho^{-i\mu} v_t + \varepsilon_{3,t},$$

where $\varepsilon_{i,t},\ i = 1, 2, 3$, represent the complex-valued random variables whose real and imaginary variables follow a normal distribution. Recall that the process $\{v_t\}$ follows a complex-valued wrapped Cauchy process; this model can provide smoothed trends in the two-dimensional plane.

21.3 Parameter Estimation

We discusses parameter estimation for the sine-skewed processes in the complex plane. We have three parameters $\boldsymbol{\varphi}^{(\mathrm{SSWC})} = (\mu, \rho, \eta)^T$ to be estimated for the SSWC process and $\boldsymbol{\varphi}^{(\mathrm{SSVM})} = (\mu, \kappa, \eta)^T$ for the SSVM process. For the sake of simplicity, we denote the parameters of the model by $\boldsymbol{\varphi} = (\mu, \beta, \eta)^T$, where β denotes the concentration of the spectral density function. The corresponding parameter space is defined by

$$\Gamma = \left\{(\mu, \beta, \eta) \middle| \, 0 \le \mu < 2\pi, \; 0 < \beta < 1, \; -1 \le \eta \le 1\right\}. \tag{21.7}$$

Estimation methods considered here are a method of moments and the Whittle likelihood method, instead of a conditional least squares type method. We recommend these methods in practice, because of the computational efficacy.

Let the parameter spaces of $\boldsymbol{\varphi}^{(\mathrm{SSWC})}$ and $\boldsymbol{\varphi}^{(\mathrm{SSVM})}$ be

$$\Gamma_{\mathrm{WC}} = \left\{(\mu, \rho, \eta) \middle| \, 0 \le \mu < 2\pi, \; 0 < \rho < 1, \; -1 \le \eta \le 1\right\}$$

and

$$\Gamma_{\mathrm{VM}} = \left\{(\mu, \kappa, \eta) \middle| 0 \le \mu < 2\pi, 0 < \kappa \le 1, -1 \le \eta \le 1\right\},$$

respectively. The following lemma states the identifiability of the sine-skewed spectral densities.

Lemma 21.1 ([17, Propositions 1 and 2])

(a) The family $\{f_{SSWC}(\lambda|\mu, \rho, \eta)| \, (\mu, \rho, \eta)^T \in \Gamma_{WC}\}$ *of the SSWC spectral density function is identifiable.*
(b) The family $\{f_{SSVM}(\lambda|\mu, \kappa, \eta)| \, (\mu, \kappa, \eta)^T \in \Gamma_{VM}\}$ *of the SSVM spectral density function is identifiable.*

It can be shown that the SSWC and SSVM processes are expressed by the linear process using complex-valued coefficient $\alpha_j(\boldsymbol{\varphi}) \in \mathbb{C}, \;\; j = 1, \ldots, \infty$. Hence, we consider a scalar complex-valued process $\{Z_t; t \in \mathbb{Z}\}$ given by

$$Z_t = \sum_{j=0}^{\infty} \alpha_j(\boldsymbol{\varphi})\varepsilon_{t-j},$$

where $\{\varepsilon_t\}$ is an i.i.d. complex-valued Gaussian process with zero mean, variance $E(\varepsilon_t \overline{\varepsilon_t}) = \sigma_\varepsilon^2$, and complementary variance $E(\varepsilon_t^2) = 0$. The last condition of zero complementary variance indicates that the process is proper. The proper assumption indicates that the white noise process $\varepsilon_t = u_t + i v_t$ satisfies $\sigma_u^2 = \sigma_v^2$ and $\mathrm{Cov}(u_t, v_t) = 0$. We call the process defined above a *proper complex-valued Gaussian process*. Recall that if the process is proper, complementary autocovariance function satisfies $\widetilde{\gamma}_Z(h) = E[Z_{t+h} Z_t] = 0$ for all $h \in \mathbb{Z}$.

21.3.1 Method of Moments Estimation

The sample autocorrelation function should be matched to the theoretical one because we can evaluate the theoretical autocorrelation function of the sine-skewed process; this provides the method of moments (MM) estimator. For the SSWC process, the estimating function is given by

$$\sum_{i=3}^{n} \boldsymbol{g}^{(\mathrm{SSWC})}(\boldsymbol{Z}_i, \boldsymbol{\varphi}) = \begin{pmatrix} \sqrt{\rho^2 + \eta^2(1-\rho^2)^2/4} \cdot \cos(\mu_1(\boldsymbol{\varphi})) - \mathrm{Re}(\hat{\gamma}_Z(1)) \\ \sqrt{\rho^2 + \eta^2(1-\rho^2)^2/4} \cdot \sin(\mu_1(\boldsymbol{\varphi})) - \mathrm{Im}(\hat{\gamma}_Z(1)) \\ \sqrt{\rho^4 + \eta^2(\rho-\rho^3)^2/4} \cdot \cos(\mu_2(\boldsymbol{\varphi})) - \mathrm{Re}(\hat{\gamma}_Z(2)) \end{pmatrix} = \mathbf{0}.$$

The third equation above can be replaced by

$$\sqrt{\rho^4 + \eta^2(\rho-\rho^3)^2/4} \cdot \sin(\mu_2(\boldsymbol{\varphi})) - \mathrm{Im}(\hat{\gamma}_Z(2)) = 0.$$

Then the MM estimator $\hat{\boldsymbol{\varphi}}^{(MM)}$ is given by the solution of the equations

$$\sum_{i=3}^{n} \boldsymbol{g}^{(\mathrm{SSWC})}(\boldsymbol{Z}_i, \hat{\boldsymbol{\varphi}}^{(MM)}) = 0.$$

Let $\boldsymbol{\varphi}_0$ be the solution of

$$E\left[\boldsymbol{g}^{(\mathrm{SSWC})}(\boldsymbol{Z}; \boldsymbol{\varphi}_0)\right] = 0,$$

where the expectation is taken under the true joint distribution of $\boldsymbol{Z} = (Z_1, Z_2, Z_3)^T$. A similar estimating function is constructed for the SSVM process as

$$\sum_{i=3}^{n} \boldsymbol{g}^{(\mathrm{SSVM})}(\boldsymbol{Z}_i, \boldsymbol{\varphi}) = \begin{pmatrix} I_1(2\kappa)/(2\kappa I_0(2\kappa))\sqrt{4\kappa^2 + \eta^2} \cdot \cos(\mu_1(\boldsymbol{\varphi})) - \mathrm{Re}(\hat{\gamma}_Z(1)) \\ I_1(2\kappa)/(2\kappa I_0(2\kappa))\sqrt{4\kappa^2 + \eta^2} \cdot \sin(\mu_1(\boldsymbol{\varphi})) - \mathrm{Im}(\hat{\gamma}_Z(1)) \\ I_2(2\kappa)/(\kappa I_0(2\kappa))\sqrt{\kappa^2 + \eta^2} \cdot \cos(\mu_2(\boldsymbol{\varphi})) - \mathrm{Re}(\hat{\gamma}_Z(2)) \end{pmatrix} = \mathbf{0}.$$

The third equation above can be replaced by

$$I_2(2\kappa)/(\kappa I_0(2\kappa))\sqrt{\kappa^2 + \eta^2} \cdot \sin(\mu_2(\boldsymbol{\varphi})) - \mathrm{Im}(\hat{\gamma}_Z(2)) = 0.$$

We are ready to establish the asymptotic normality of the MM estimator with an additional assumption stated below.

Assumption 21.1

(a) The process satisfies $E|Z_t|^4 < \infty$ and $\sum_{h=1}^{\infty} h|\gamma_Z(h)|^2 < \infty$.
(b) The parameter spaces Γ_{WC} and Γ_{VM} are compact and the true parameter vector $\boldsymbol{\varphi}_0$ belongs to the interior of Γ_{WC} or Γ_{VM}.
(c) The autocovariance functions defined by (21.5) and (21.6) are twice continuously differentiable and their absolute values are bounded by some constant for all $\boldsymbol{\varphi}$.

(d) $E[\sup_{\varphi\in\Gamma} \|\boldsymbol{g}^{(\text{Model})}(Z, \boldsymbol{\varphi})\|] < \infty$ for Model $\in$ {SSWC, SSVM}.

(e) The matrix $\boldsymbol{A}(\boldsymbol{\varphi}_0) = E\left[-\frac{\partial}{\partial \boldsymbol{\varphi}^T} \boldsymbol{g}^{(\text{Model})}(\boldsymbol{Z}, \boldsymbol{\varphi}_0)\right]$ is nonsingular.

Theorem 21.3 *Suppose that Assumption 21.1 holds. Then, the following asymptotic normality holds for the SSWC and SSVM processes:*

$$\sqrt{n}(\hat{\boldsymbol{\varphi}}^{(MM)} - \boldsymbol{\varphi}_0) \to_d N\left(\boldsymbol{0}, \boldsymbol{A}(\boldsymbol{\varphi}_0)^{-1}\boldsymbol{B}(\boldsymbol{\varphi}_0)\boldsymbol{A}(\boldsymbol{\varphi}_0)^{-1}\right),$$

where

$$\boldsymbol{A}(\boldsymbol{\varphi}_0) = E\left[-\frac{\partial}{\partial \boldsymbol{\varphi}^T} \boldsymbol{g}^{(Model)}(\boldsymbol{Z}, \boldsymbol{\varphi}_0)\right],$$
$$\boldsymbol{B}(\boldsymbol{\varphi}_0) = E\left[\boldsymbol{g}^{(Model)}(\boldsymbol{Z}, \boldsymbol{\varphi}_0)\boldsymbol{g}^{(Model)}(\boldsymbol{Z}, \boldsymbol{\varphi}_0)^T\right],$$

for Model $\in \{SSWC, SSVM\}$.

Proof Here we provide a sketch of the proof. The Taylor expansion of $\boldsymbol{G}_n(\boldsymbol{\varphi}) = \frac{1}{n}\sum_{i=1}^n \boldsymbol{g}^{(\text{Model})}(\boldsymbol{Z}_i, \boldsymbol{\varphi})$ gives

$$\boldsymbol{0} = \boldsymbol{G}_n(\hat{\boldsymbol{\varphi}}) = \boldsymbol{G}_n(\boldsymbol{\varphi}_0) + \boldsymbol{G}_n'(\boldsymbol{\varphi}_0)(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}_0) + \boldsymbol{R}_n, \quad \boldsymbol{G}_n'(\boldsymbol{\varphi}_0) = \frac{\partial}{\partial \boldsymbol{\varphi}^T} G_n(\boldsymbol{\varphi})\bigg|_{\varphi=\varphi_0}.$$

By Assumptions 21.1(a, c), the absolute value of the second derivative of $\boldsymbol{G}_n(\boldsymbol{\varphi})$ is bounded above. Hence, we can show that

$$\sqrt{n}(\hat{\boldsymbol{\varphi}} - \boldsymbol{\varphi}_0) = \left[-\boldsymbol{G}_n'(\boldsymbol{\varphi}_0)\right]^{-1}\sqrt{n}\boldsymbol{G}_n(\boldsymbol{\varphi}_0) + \sqrt{n}\boldsymbol{R}_n.$$

From the identifiability result given in Lemma 21.1, we observe that $E[\|\boldsymbol{g}(\boldsymbol{Z}, \boldsymbol{\varphi})\|]$ has a unique minimum at $\boldsymbol{\varphi}_0$. This together with Assumptions 21.1(b–d) implies the consistency of the MM estimator, such that $\hat{\boldsymbol{\varphi}}^{(MM)} \to_p \boldsymbol{\varphi}_0$. In addition, we observe that, as $n \to \infty$,

$$-\boldsymbol{G}_n'(\boldsymbol{\varphi}_0) = \frac{1}{n}\sum_{i=1}^n\left[-\frac{\partial}{\partial \boldsymbol{\varphi}^T}\boldsymbol{g}(\boldsymbol{Z}_i, \boldsymbol{\varphi}_0)\right] \to_p E\left[-\frac{\partial}{\partial \boldsymbol{\varphi}^T}\boldsymbol{g}(\boldsymbol{Z}, \boldsymbol{\varphi}_0)\right] = \boldsymbol{A}(\boldsymbol{\varphi}_0),$$
$$\sqrt{n}\boldsymbol{G}_n'(\boldsymbol{\varphi}_0) \to_d N(\boldsymbol{0}, \boldsymbol{B}(\boldsymbol{\varphi}_0)), \quad \boldsymbol{B}(\boldsymbol{\varphi}_0) = E\left[\boldsymbol{g}(\boldsymbol{Z}, \boldsymbol{\varphi}_0)\boldsymbol{g}(\boldsymbol{Z}, \boldsymbol{\varphi}_0)^T\right],$$
$$\sqrt{n}\boldsymbol{R}_n \to_p \boldsymbol{0}.$$

Here the existence of the matrix $\boldsymbol{B}(\boldsymbol{\varphi}_0)$ follows from Assumption 21.1(a), which together with Assumption 21.1(e) yields the asymptotic normality. □

The MM estimator is related to the Yule–Walker estimator in the AR models. We consider the wrapped Cauchy (WC) process for simplicity. Multiplying $\overline{Z_{t-1}}$ from both sides of (21.2) and taking the expectation, we get

$$E(Z_t\overline{Z_{t-1}}) = \rho^{-i\mu}E(Z_{t-1}\overline{Z_{t-1}}) + E(\varepsilon_t\overline{Z_{t-1}}) \Longleftrightarrow \gamma_Z(1) = \rho^{-i\mu}\gamma_Z(0).$$

Then, $\hat{\rho}^{-i\hat{\mu}} = \hat{\rho}_Z(1)$. Rewriting the sample autocorrelation at lag 1 as $\hat{\rho}_Z(1) = A + Bi$, the MM estimator for ρ and μ are expressed as

$$\sum_{t=2}^{n} \boldsymbol{g}(z_t, \boldsymbol{\eta}) = \begin{pmatrix} \rho\cos\mu - \mathrm{Re}\left(\frac{\sum_{t=2}^{n} Z_t\overline{Z_{t-1}}}{\sum_{t=1}^{n} Z_t\overline{Z_t}}\right) \\ \rho\sin\mu - \mathrm{Im}\left(\frac{\sum_{t=2}^{n} Z_t\overline{Z_{t-1}}}{\sum_{t=1}^{n} Z_t\overline{Z_t}}\right) \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.$$

The explicit expressions for the estimators are given by

$$\hat{\mu}_{\mathrm{WC}}^{(\mathrm{MM})} = \mathrm{atan2}(B, A), \quad \hat{\rho}_{\mathrm{WC}}^{(\mathrm{MM})} = \frac{A}{\cos\hat{\mu}_{\mathrm{WC}}^{(\mathrm{MM})}} = \frac{B}{\sin\hat{\mu}_{\mathrm{WC}}^{(\mathrm{MM})}}.$$

Similarly, the MM estimator for the von Mises (VM) process is given by

$$\hat{\mu}_{\mathrm{VM}}^{(\mathrm{MM})} = \mathrm{atan2}(B, A), \quad \frac{I_1(2\hat{\kappa}_{\mathrm{VM}}^{(\mathrm{MM})})}{I_0(2\hat{\kappa}_{\mathrm{VM}}^{(\mathrm{MM})})} = \frac{A}{\cos\hat{\mu}_{\mathrm{VM}}^{(\mathrm{MM})}} = \frac{B}{\sin\hat{\mu}_{\mathrm{VM}}^{(\mathrm{MM})}}.$$

Note that $\hat{\kappa}_{\mathrm{VM}}^{(\mathrm{MM})}$ must be solved numerically.

21.3.2 Whittle Estimation

We now apply the Whittle likelihood method to estimate the unknown parameter vector $\boldsymbol{\varphi}$. Let $\boldsymbol{Z}_n = (Z_1, \ldots, Z_n)^T$ and its conjugate transpose be $\boldsymbol{Z}_n^* = (\overline{Z_1}, \ldots, \overline{Z_n})$. Let the $n \times n$ Hermitian covariance matrix $\boldsymbol{R}_Z$ be

$$\boldsymbol{R}_Z(f_\varphi) = [\gamma_Z(k-j)]_{j,k=1,\ldots,n},$$

whose (j, k)-th element is

$$\gamma_Z(j-k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} e^{i(j-k)\lambda} f_\varphi(\lambda) d\lambda,$$

with $\gamma_Z(k-j) = \overline{\gamma_Z(j-k)}$. The proper normal probability density function of the joint complex-valued random variables is given by

$$p(\boldsymbol{z}) = \frac{1}{\pi^n \det \boldsymbol{R}_Z(f_\varphi)} \exp\left\{-\boldsymbol{Z}_n^* \boldsymbol{R}_Z(f_\varphi)^{-1} \boldsymbol{Z}_n\right\},$$

see, for example, [21, p. 39]. Then the log-likelihood function of the observed sequence $\boldsymbol{Z}_n$ becomes

$$\ell(\boldsymbol{\varphi}) = -\log\det\boldsymbol{R}_Z(f_\varphi) - \boldsymbol{Z}_n^*\boldsymbol{R}_Z(f_\varphi)^{-1}\boldsymbol{Z}_n.$$

The periodogram of the complex-valued processes is defined by

$$I_n(Z_t;\lambda) = \frac{1}{2\pi n}\left|\sum_{t=1}^{n} Z_t e^{it\lambda}\right|^2.$$

By similar argument of [4, Sect. 10.8], we can see that

$$\log\det\boldsymbol{R}_Z(f_\varphi) + \boldsymbol{Z}_n^*\boldsymbol{R}_Z(f_\varphi)^{-1}\boldsymbol{Z}_n = \sum_{j=1}^{n}\left(\log f_\varphi(\lambda_j) + \frac{I_n(Z_t;\lambda_j)}{f_\varphi(\lambda_j)}\right) + O(1).$$

Then, the Whittle estimator for the parameter vector $\boldsymbol{\varphi}$ is obtained by minimizing

$$D(f_\varphi, I_n) = \int_{-\pi}^{\pi}\left\{\log f_\varphi(\lambda) + \frac{I_n(Z_t;\lambda)}{f_\varphi(\lambda)}\right\}d\lambda,$$

i.e.,

$$\hat{\boldsymbol{\varphi}}^{(W)} = \underset{\varphi\in\Gamma}{\text{argmin}}\, D(f_\varphi, I_n),$$

where the parameter space Γ is given in (21.7). Let $\boldsymbol{\varphi}_0$ be the minimizer of the function $D(f_\varphi, f)$ such that $\boldsymbol{\varphi}_0 = \text{argmin}_{\varphi\in\Gamma} D(f_\varphi, f)$.

The Whittle estimator is asymptotically efficient under proper complex-valued Gaussian assumptions because the function $D(f_\varphi, I_n)$ is an approximation to the proper complex-valued Gaussian likelihood function. In addition, the Whittle estimator is a quasi-likelihood estimator for the case when the complex-valued error terms are improper and we misspecify the underlying complex-valued density functions or time series models.

The following assumption is required for the asymptotic results of the Whittle estimator.

Assumption 21.2

(a) The spectral density $f_\varphi(\lambda)$ and $f_\varphi^{-1}(\lambda)$ are continuous for $\boldsymbol{\varphi}$ and λ. In addition, $f_\varphi(\lambda)$ is twice differentiable with respect the φ and derivatives are continuous with respect to $\lambda \in [-\pi, \pi]$ and $\boldsymbol{\varphi}$.

(b) There exists constants $M > 0$ and $m >$ such that for any $\lambda \in [0, 2\pi)$,

$$\sup_{\varphi\in H} f_\varphi(\lambda) < M \quad \text{and} \quad \inf_{\varphi\in H} f_\varphi(\lambda) > m.$$

(c) The matrix $\mathcal{I}(\boldsymbol{\varphi}_0) = \frac{1}{4\pi}\int_{-\pi}^{\pi}\frac{\partial}{\partial\boldsymbol{\varphi}}\log f_{\varphi_0}(\lambda)\frac{\partial}{\partial\boldsymbol{\varphi}^T}\log f_{\varphi_0}(\lambda)d\lambda$ is nonsingular.

The following asymptotic normality holds for $\hat{\boldsymbol{\varphi}}^{(W)}$.

Theorem 21.4 *We assume that the process* $\{Z_t\}$ *represents a proper complex-valued Gaussian stationary process and Assumptions 21.1(a, b) and 21.2 hold. In addition, the true spectral density* f *is specified by* f_{φ_0}. *Then,*

$$\sqrt{n}(\hat{\boldsymbol{\varphi}}^{(W)} - \boldsymbol{\varphi}_0) \to_d N\left(\mathbf{0}, \mathcal{I}^{-1}(\boldsymbol{\varphi}_0)\right),$$

where

$$\mathcal{I}(\boldsymbol{\varphi}_0) = \frac{1}{4\pi}\int_{-\pi}^{\pi} \frac{\partial}{\partial \boldsymbol{\varphi}} \log f_{\varphi_0}(\lambda) \frac{\partial}{\partial \boldsymbol{\varphi}^T} \log f_{\varphi_0}(\lambda) d\lambda.$$

Proof From the Kullback–Leibler inequality and the identifiability lemma given by Lemma 21.1, $D(f_\varphi, f_{\varphi_0})$ is minimized by $\boldsymbol{\varphi}_\mathbf{0}$. For the Whittle estimator $\hat{\boldsymbol{\varphi}}^{(W)}$ such that $D(f_\varphi, f_{\varphi_0}) \le D(f_\varphi, f_{\hat{\varphi}^{(W)}})$ and Assumption 21.2(b), we observe $D(f_\varphi, f_{\hat{\varphi}^{(W)}}) \to D(f_\varphi, f_{\varphi_0})$ and this implies $\hat{\boldsymbol{\varphi}}^{(W)} \to \boldsymbol{\varphi}_0$ in probability. From Assumption 21.2(a) and $f = f_{\varphi_0}$, we have

$$\frac{\partial}{\partial \boldsymbol{\varphi}} D(f_\varphi, I_n)\bigg|_{\boldsymbol{\varphi}=\hat{\boldsymbol{\varphi}}^{(W)}} = 0$$

and

$$\sqrt{n}(\hat{\boldsymbol{\varphi}}^{(W)} - \boldsymbol{\varphi}_0)$$
$$= \left(-\frac{\partial}{\partial \boldsymbol{\varphi}\partial \boldsymbol{\varphi}^T} D(f_\varphi, I_n)\bigg|_{\boldsymbol{\varphi}=\tilde{\boldsymbol{\varphi}}}\right)^{-1} \sqrt{n}\left(\frac{\partial}{\partial \boldsymbol{\varphi}} D(f_\varphi, I_n)\bigg|_{\boldsymbol{\varphi}=\boldsymbol{\varphi}_0} - \frac{\partial}{\partial \boldsymbol{\varphi}} D(f_\varphi, f_{\varphi 0})\bigg|_{\boldsymbol{\varphi}=\boldsymbol{\varphi}_0}\right),$$

where $\tilde{\boldsymbol{\varphi}}$ satisfies $\left\|\tilde{\boldsymbol{\varphi}} - \boldsymbol{\varphi}_0\right\| \le \left\|\hat{\boldsymbol{\varphi}}^{(W)} - \boldsymbol{\varphi}_0\right\|$. We have

$$\sqrt{n}\frac{\partial}{\partial \boldsymbol{\varphi}} D(f_\varphi, I_n)\bigg|_{\boldsymbol{\varphi}_0} \to_d N(\mathbf{0}, \mathcal{I}(\boldsymbol{\varphi}_0)).$$

In addition, we have

$$-\frac{\partial}{\partial \boldsymbol{\varphi}\partial \boldsymbol{\varphi}^T} D(f_\varphi, I_n)\bigg|_{\tilde{\boldsymbol{\varphi}}} \to_p \frac{1}{4\pi}\int_{-\pi}^{\pi} (f(\lambda) - f_{\varphi_0}(\lambda)) \frac{\partial^2}{\partial \boldsymbol{\varphi}\partial \boldsymbol{\varphi}^T} f_{\varphi_0}(\lambda)^{-1} d\lambda + \mathcal{I}(\boldsymbol{\varphi}_0),$$

where the first term vanishes from the corrected model assumption. Then we obtain the desired result. □

The non-Gaussian extension of the process $\{Z_t\}$ is possible; for example, [6, 10, 23]. In such a case, it must be investigated the fourth-order cumulant spectra of the complex-valued process $\{Z_t\}$, and we leave this to future research.

Table 21.1 Simulation results for the SSWC process. The true parameters were set at $\mu = \pi/6$, $\rho = 0.8$, and $\eta = 0.5$

	MM			Whittle		
	$\hat{\mu}^{(\mathrm{MM})}$	$\hat{\rho}^{(\mathrm{MM})}$	$\hat{\eta}^{(\mathrm{MM})}$	$\hat{\mu}^{(\mathrm{W})}$	$\hat{\rho}^{(\mathrm{W})}$	$\hat{\eta}^{(\mathrm{W})}$
$n = 250$						
Mean	0.5229	0.7957	0.4979	0.4985	0.7967	0.4909
RMSE	0.0444	0.0309	0.1019	0.0402	0.0149	0.0860
$n = 500$						
Mean	0.5237	0.7984	0.5001	0.5125	0.7986	0.4931
RMSE	0.0316	0.0197	0.0699	0.0287	0.0102	0.0600
$n = 1000$						
Mean	0.5238	0.7984	0.4980	0.5177	0.7994	0.4957
RMSE	0.0202	0.0137	0.0505	0.0183	0.0067	0.0437

21.4 Monte Carlo Simulations

We simulated the SSWC and SSVM processes with different sample size $n = 250, 500, 1000$. The model parameters were the same as those used in Sect. 21.2: $\boldsymbol{\varphi}_{\mathrm{SSWC}} = (\mu, \rho, \eta)^T = (\pi/6, 0.8, 0.5)^T$ for the SSWC process and $\boldsymbol{\varphi}_{\mathrm{SSVM}} = (\mu, \kappa, \eta)^T = (-\pi/6, 0.6, -0.5)^T$ for the SSVM process. The error $\{\varepsilon_t\}$ was i.i.d. complex-valued normal noise with zero mean and variance $\sigma_\varepsilon^2 = 1 - \rho^2$ for the SSWC process and $\sigma_\varepsilon^2 = 1/I_0(2\kappa)$ for the SSVM process. The bias and root mean squared errors (RMSE) for the MM estimator $\hat{\boldsymbol{\varphi}}^{(\mathrm{MM})}$ and the Whittle estimator $\hat{\boldsymbol{\varphi}}^{(\mathrm{W})}$ were calculated with 1,000 replicated time series. The results are summarized in Tables 21.1 and 21.2 for SSWC and SSVM processes, respectively.

Table 21.2 Simulation results for the SSVM process. The true parameters were set at $\mu = -\pi/6$, $\kappa = 0.6$, and $\eta = -0.5$

	MM			Whittle		
	$\hat{\mu}^{(\mathrm{MM})}$	$\hat{\kappa}^{(\mathrm{MM})}$	$\hat{\eta}^{(\mathrm{MM})}$	$\hat{\mu}^{(\mathrm{W})}$	$\hat{\kappa}^{(\mathrm{W})}$	$\hat{\eta}^{(\mathrm{W})}$
$n = 250$						
Mean	−0.5447	0.6039	−0.4674	−0.5467	0.5844	−0.4668
RMSE	0.2883	0.0976	0.2815	0.2566	0.0886	0.2419
$n = 500$						
Mean	−0.5671	0.6044	−0.4449	−0.5523	0.5910	−0.4617
RMSE	0.2396	0.0730	0.2519	0.1994	0.0604	0.2018
$n = 1000$						
Mean	−0.5578	0.6025	−0.4544	−0.5552	0.5998	−0.4624
RMSE	0.1983	0.0545	0.2139	0.1610	0.0409	0.1695

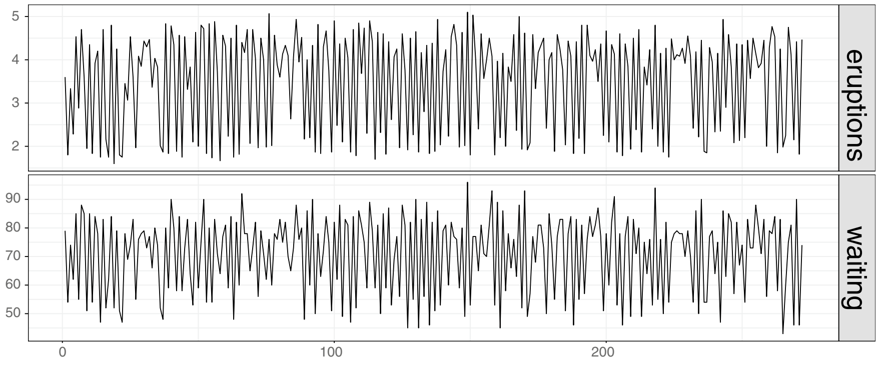


Fig. 21.6 Time series plots of the bivariate process of the Old Faithful geyser data

The following findings were made by the inspection of the tables. The RMSE decreases with an increase in the sample size, and this indicates the consistency of both estimators. The RMSE of the Whittle estimator is smaller than that of the MM estimator for all cases. Further, this fact indicates the asymptotic efficiency of the Whittle estimator. The bias and RMSE of the SSVM process are considerably higher than those of the SSWC process because the autocorrelation at lag 1 of the SSWC process is higher than that of the SSVM process; this results in smaller asymptotic variances of the estimator.

21.5 Data Analysis

Two datasets were considered to demonstrate the applicability of the proposed models, and the fitting performances were compared between models and estimation methods.

21.5.1 Old Faithful Geyser Data

The first example was the well-known Old Faithful geyser data; this dataset was considered by [3] and investigated by fitting hidden Markov models and finite mixture models by various researchers. The waiting time between eruptions and the eruption duration for the Old Faithful geyser in Yellowstone National Park were measured, and the sample size was set as $n = 272$. Figure 21.6 shows a time series plot of the bivariate process of the Old Faithful geyser data. The negative autocorrelations of both series are confirmed and the scatterplots and transitions in the two-dimensional plane are plotted in Fig. 21.7.

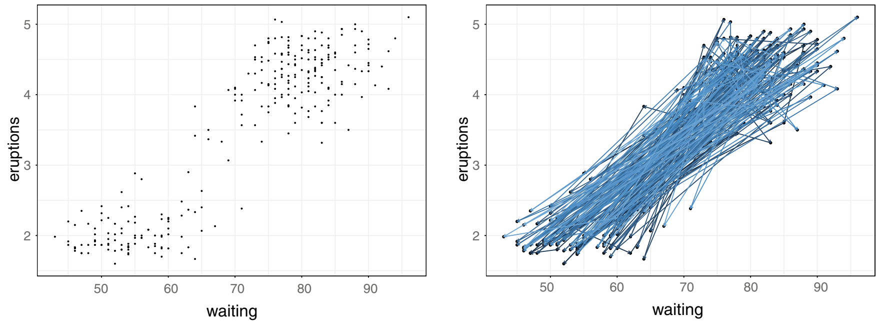


Fig. 21.7 Scatterplot of the waiting time and eruptions (left) and its time paths (right)

Table 21.3 Estimated parameters for different estimation methods

	Wrapped Cauchy process					
	$\hat{\mu}$	$\hat{\rho}$	$\hat{\eta}$	$-L(\hat{\boldsymbol{\varphi}})$	$\rho_1(\hat{\boldsymbol{\varphi}})$	$\mu_1(\hat{\boldsymbol{\varphi}})$
MM	3.1416	0.5450	0.1100	323.90	0.5463	−3.071
Whittle	3.1210	0.5439	0.1011	323.99	0.5450	−3.097
	von Mises process					
	$\hat{\mu}$	$\hat{\kappa}$	$\hat{\eta}$	$-L(\hat{\boldsymbol{\varphi}})$	$\rho_1(\hat{\boldsymbol{\varphi}})$	$\mu_1(\hat{\boldsymbol{\varphi}})$
MM	2.7786	0.6536	0.6282	301.88	0.6044	−3.057
Whittle	2.7371	0.4986	0.5172	318.46	0.5017	−3.068

The data were transformed by standardization with mean 0 and variance 1, and by taking these series as the complex plane such that $z_t = x_t + iy_t$, where x_t and y_t were the standardized waiting time and eruptions, respectively. Further, the process $\{z_t\}$ was normalized via a transformation $\tilde{z}_t = z_t/\sqrt{\hat{\gamma}_z(0)}$. We fitted the time series $\{\tilde{z}_t\}$ using the SSWC and SSVM processes. The estimated parameters were summarized in Table 21.3. In Table 21.3, the Whittle likelihood $-L(\hat{\boldsymbol{\varphi}})$ was shown to compare the fitted performances among the models. According to Table 21.3, we saw that the μ was around π, and it indicated that the transition was in the opposite direction from the observed point to the next. For the SSVM process, these characteristics were captured by the skewing parameter η, which was considerably larger than that of the SSWC process. The values of the autocorrelation at lag 1 calculated from the estimated parameters were also shown in Table 21.3, which indicated that the mean locations of the frequency were almost the same among the models at $-\pi$ and the persistence of the dependencies varied from 0.5 to 0.6.

The periodogram with the estimated parametric spectral densities was plotted in Fig. 21.8. The positively skewed structure of the periodogram was well captured by fitting the SSVM process.

Finally, we fitted the VAR(p) for the model comparison. The selected lag of the standardized time series was 1 for both AIC and BIC; the estimated model was

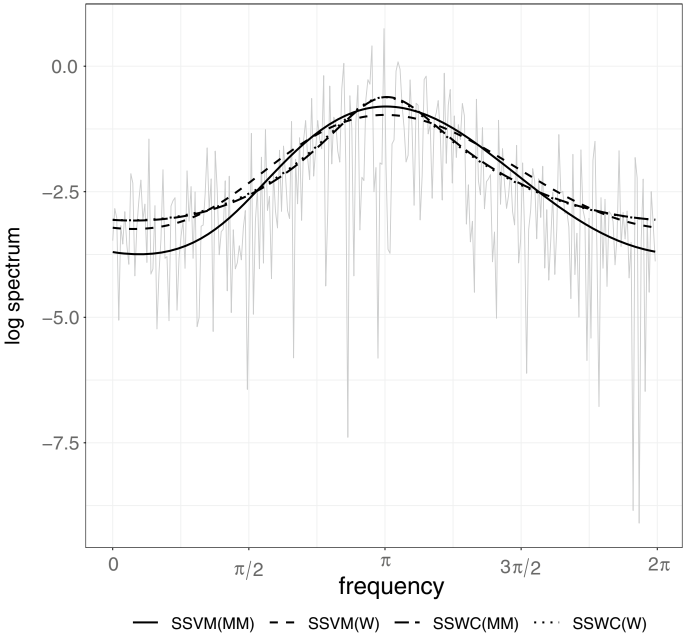


Fig. 21.8 Periodogram together with estimated spectral density functions for Old Faithful geyser dataset

$$\begin{pmatrix} Y_{1,t} \\ Y_{2,t} \end{pmatrix} = \begin{pmatrix} -0.5616 & 0.0246 \\ -0.3143 & -0.2670 \end{pmatrix} \begin{pmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix}, \quad \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.713 & 0.607 \\ 0.607 & 0.684 \end{pmatrix} \right).$$

Our proposed model has four parameters including the complex-valued white noise variance, whereas the simplest VAR(1) model has seven parameters. The skewness in the spectral density is attributed to the asymmetric effects of the cross-spectrum of the observed time series; this was confirmed by the off-diagonal elements of the VAR(1) coefficients.

21.5.2 Real Wage and Unemployment Rate in Canada

In the second example, the time series of the real wage and unemployment rate in Canada was investigated. This dataset was illustrated by [15] by fitting the vector error correction model. The datasets covered the period from the first quarter of 1980 to the fourth quarter of 2000, and the sample size was $n = 84$. From the time series plot shown in the left panel of Fig. 21.9, we observed that the process was nonstationary, and therefore, we detrended the data by taking the first difference of the series. The trajectory of the normalized detrended series was plotted in the right panel of Fig. 21.9, and it shows no clear patterns in the transition.

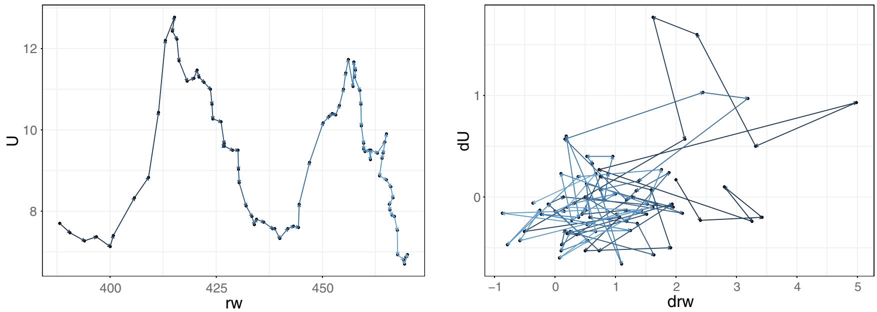


Fig. 21.9 Scatterplots of the real wage and unemployment rate data (left) and its first difference series (right)

Table 21.4 Estimated parameters for different estimation methods for the changes in real wage and unemployment rate in Canada

	Wrapped Cauchy process					
	$\hat{\mu}$	$\hat{\rho}$	$\hat{\eta}$	$-L(\hat{\varphi})$	$\rho_1(\hat{\varphi})$	$\mu_1(\hat{\varphi})$
MM	0.3752	0.5422	−0.4521	91.60	0.5152	0.0890
Whittle	0.2541	0.4603	−0.4869	94.00	0.4723	−0.1408
	von Mises process					
	$\hat{\mu}$	$\hat{\kappa}$	$\hat{\eta}$	$-L(\hat{\varphi})$	$\rho_1(\hat{\varphi})$	$\mu_1(\hat{\varphi})$
MM	0.7085	0.5733	−0.8258	73.53	0.6112	0.2582
Whittle	0.5052	0.3387	−0.6982	91.56	0.4605	−0.2345

Estimated parameters are summarized in Table 21.4. We observed that the SSWC process has a better fitting performance than that of the SSVM process in terms of the likelihood values. The fitted parameters for μ and λ are different between the models such that the parameters more skewed to the left spectral densities are fitted using the SSVM process. The fitted spectral densities with the periodogram were plotted in Fig. 21.10, and it indicated that the proposed skewed model can adequately capture the asymmetry of the data periodogram.

Finally, we fitted the VAR(p) for model comparison. The selected lag of the standardized time series was 1 for both the AIC and the BIC. The estimated model was given by

$$\begin{pmatrix} Y_{1,t} \\ Y_{2,t} \end{pmatrix} = \begin{pmatrix} 0.3653 & 0.1566 \\ 0.1535 & 0.5020 \end{pmatrix} \begin{pmatrix} Y_{1,t-1} \\ Y_{2,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix}, \quad \begin{pmatrix} \varepsilon_{1,t} \\ \varepsilon_{2,t} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0.806 & 0.194 \\ 0.194 & 0.676 \end{pmatrix} \right).$$

This result indicates that the source of the skewness in the spectral density is caused by the different effects on the autocorrelations at lag 1 of the series such that the larger autocorrelation in the unemployment series compared to the real wages yields a negatively skewed spectrum.

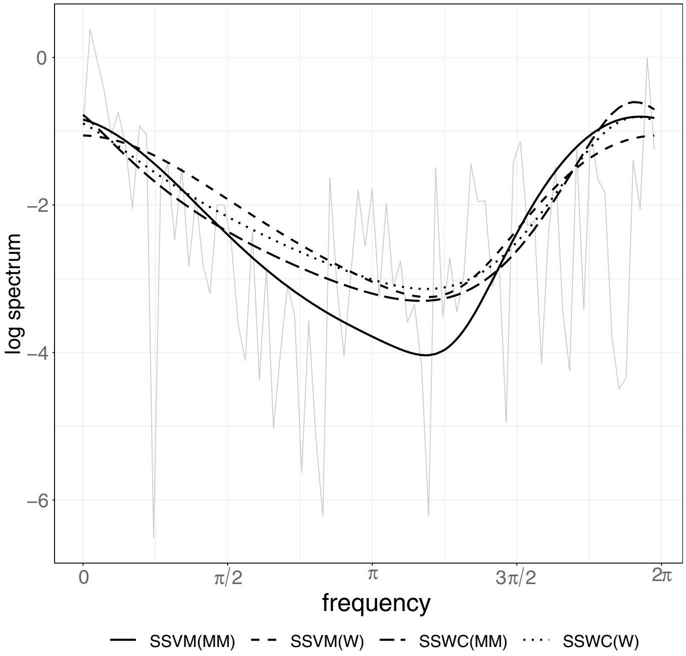


Fig. 21.10 Periodogram together with estimated spectral density functions for Canada data

21.6 Summary and Conclusions

We have considered the complex-valued time series models. The proposed models included the complex-valued AR(1) and MA(∞) models whose spectral densities corresponded to the wrapped Cauchy and von Mises distributions on a circle. Further, the SSWC and SSVM processes have been introduced, and two estimators (the MM and Whittle estimators) have been discussed, together with their asymptotic distributions. These estimation methods have been investigated via numerical simulations and two real datasets.

Acknowledgements The author would like to express deep thanks to the editorial team of this Festschrift for helpful comments and suggestions. This research was supported by JSPS KAKENHI Grant Numbers 18K01706 and 22K11944, and Nanzan University Pache Research I-A-2 for the 2022 and 2023 academic years.

References

1. Abe, T. and Pewsey, A. (2011). Sine-skewed circular distributions. *Statistical Paper* **52** 683–707.
2. Abe, T., Ogata, H., Shiohama, T. and Taniai, H. (2017). A circular autocorrelation of stationary circular Markov processes. *Statistical Inference for Stochastic Processes* **20** 275–290.

3. Azzalini, A. and Bowman, A. W. (1990). A look at some data on the Old Faithful geyser. *Journal of the Royal Statistical Society: Series C* **39** 3570–365.
4. Brockwell, P. J. and Davis, R. A. (1991). *Time Series: Theory and Methods.* Springer Science & Business Media.
5. Bloomfield, P. (1973). An exponential model for the spectrum of a scalar time series. *Biometrika* **60** 217–226.
6. Brillinger, D. R. (1981). *Time Series: Data Analysis and Theory*. SIAM
7. Gatto, R. (2022). Information theoretic results for stationary time series and the Gaussian-generalized von Mises time series. In *Directional Statistics for Innovative Applications, A Bicentennial Tribute to Florence Nightingale*, pp. 229–244.
8. Hannan, E. J. (1970). *Multiple Time Series*. Wiley.
9. Hochberg, K. J. and Orsingher, E. (1996). Composition of stochastic process governed by higher-order parabolic and hyperbolic equations. *Journal of Theoretical Probability* **9** 511–532.
10. Hosoya, Y. and Taniguchi, M. (1982). A central limit theorem for stationary processes and the parameter estimation of linear processes. *The Annals of Statistics* **10** 132-153.
11. Jumarie, G. (1999). Complex-valued Wiener measure: An approach via random walk in the complex plane. *Statistics & Probability Letters* **42** 61–67.
12. Kato, S. and Jones, M. C. (2013). An extended family of circular distributions related to wrapped Cauchy distributions via Brownian motion. *Bernoulli* **19** 154–171.
13. Kato, S. and Jones, M. C. (2015). A tractable and interpretable four-parameter family of unimodal distributions on the circle. *Biometrika* **102** 181–190.
14. Le Breton, A. (1988). A note on maximum likelihood estimation for the complex-valued first-order autoregressive process. *Statistics & Probability Letters* **7** 171–173.
15. Lütkepohl, H. and Krätzig, M. (Eds.). (2004). *Applied Time Series Econometrics*. Cambridge University Press.
16. Miller, K. S. (1974). *Complex Stochastic Processes: An Introduction to Theory and Application*. Addison Wesley Publishing Company.
17. Miyata, Y., Shiohama, T. and Abe, T. (2022). Identifiability of asymmetric circular and cylindrical distributions. *Sankyha A*. https://doi.org/10.1007/s13171-022-00294-3
18. Neeser, F. D. and Massey, J. L. (1993). Proper complex random processes with applications to information theory. *IEEE Transactions on Information Theory* **39** 1293–1302.
19. Ogata, H. and Shiohama. T. (2022). A mixture transition modeling for higher-order circular Markov processes. arXiv: 2304.00874.
20. Schreier, P. J. and Scharf, L. L. (2003). Second-order analysis of improper complex random vectors and processes. *IEEE Transactions on Signal Processing* **51** 714–725.
21. Schreier, P. J. and Scharf, L. L. (2010). *Statistical Signal Processing of Complex-Valued Data: The Theory of Improper and Noncircular Signals.* Cambridge University Press.
22. Sykulski, A. M., Olhede, S. C. and Lilly, J. M. (2016). A widely linear complex autoregressive process of order one. IEEE Transactions on Signal Processing **64** 6200–6210.
23. Taniguchi, M. and Kakizawa, Y. (2000). *Asymptotic Theory of Statistical Inference for Time Series*. Springer Science & Business Media.
24. Taniguchi, M., Kato, S., Ogata, H. and Pewsey, A. (2020). Models for circular data form time series spectra. *Journal of Time Series Analysis* **41** 808–829.

Chapter 22
Semiparametric Estimation of Optimal Dividend Barrier for Spectrally Negative Lévy Process

Yasutaka Shimizu and Hiroshi Shiraishi

Abstract We discuss a statistical estimation problem of an optimal dividend barrier when the surplus process follows a Lévy insurance risk process. The optimal dividend barrier is defined as the level of the barrier that maximizes the expectation of the present value of all dividend payments until ruin. In this paper, an estimator of the expected present value of all dividend payments is defined based on "quasi-process" in which sample paths are generated by shuffling increments of a sample path of the Lévy insurance risk process. The consistency of the optimal dividend barrier estimator is shown. Moreover, our approach is examined numerically in the case of the compound Poisson risk model perturbed by diffusion.

22.1 Introduction

In risk theory, surplus process is a very important model for understanding how the capital or surplus of an insurance company evolves over time. The classical model for the surplus process is the so-called "Cramér–Lundberg insurance risk model". In this model, the insurance company collects premiums at a fixed rate $c > 0$ from its customers. On the other hand, a customer can make a claim causing the surplus to jump downwards. The claim frequency follows a Poisson process, and the claim sizes are assumed to be independent and identically distributed (i.i.d.). A natural generalization of the Cramér–Lundberg model is a spectrally negative Lévy process also called "Lévy insurance risk model", which has been studied in many actuarial literature, such as [5, 6, 12] and so on. Thanks to the Lévy insurance risk model, we can grasp many realistic social phenomena in the surplus process such as the fluctuation of premium income, the effect of investment result, the effect of the small

Y. Shimizu
Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
e-mail: shimizu@waseda.jp

H. Shiraishi (✉)
Keio University, 3-14-1 Hiyoshi, Kohoku-ku, Yokohama Kanagawa, 223-8522, Japan
e-mail: shiraishi@math.keio.ac.jp

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_22

claim, and so on. In this paper, we also suppose that the surplus of an insurance company follows the Lévy insurance risk process.

In risk theory, the central topics are the ruin time or ruin probability, but there is a dividend problem as one of the applications. In the dividend problem introduced by De Finetti [4] (especially, the so-called "constant barrier strategy"), assuming that there is a horizontal barrier of level ϑ, such that when an insurance company's surplus reaches level ϑ, dividends are paid continuously such that the surplus stays at level ϑ until it becomes less than ϑ. The optimal strategy is to maximize the expectation of the present value of all dividend payments and the "optimal dividend barrier" is defined as a barrier of level ϑ where the maximization can be achieved. Bühlmann [1], Lin [8], Gerber et al. [7], Li [9] and Loeffen [10] derived the optimal dividend barrier explicitly in some special models such as the Cramér–Lundberg model with exponential claim amount distribution. On the other hand, Kyprianou [12] discussed a stochastic control problem for the optimal dividend strategy when the surplus process follows the Lévy insurance risk process. In these papers, the main concern is the property from a probabilistic point of view, but there is a limited contribution in the statistical point of view. From the statistical point of view in ruin theory, a ruin probability by Croux and Veraverbeke [3] and Shimizu [14], a Gurber–Shiu function by Feng and Shimizu [6] and an optimal dividend problem in the Cramér–Lundberg model by Shiraishi and Lu [16] were discussed, respectively. In this paper, we discuss the statistical estimation problem in the Lévy insurance risk process.

Considering the optimal dividend problem in a statistical estimation framework, it can be reduced to an M-estimation problem if we can define our optimal dividend barrier estimator as a maximizer of an objective function that corresponds to an estimator of the expectation of the present value of all dividend payments. Note that, for the usual M-estimator, the objective function sometimes called a contrast estimator is defined by a sample mean of i.i.d. random variables. In the same way, our contrast estimator would be defined by a sample mean of the present value of all dividend payments. However, since the present value of all dividend payments is path dependent, in order to construct the contrast estimator, we need to provide a number of independent copies of sample paths; consequently, it is impossible to observe multiple sample paths. In addition, in practical point of view, it is reasonable to assume that the surplus of an insurance portfolio is observable discretely not continuously, such as hourly, daily, monthly, and so on. To overcome these problems, we introduce "quasi-process", that is, an approximation of the true Lévy insurance risk process. The quasi-process is composed by rearranging the increments of discretely observed data, thus, it is possible to generate multiple sample paths by changing the permutation. Essentially, it takes advantage of the exchangeability of the increments in the Lévy insurance risk process. Generating multiple quasi-process, it is possible to provide a number of (approximated) present values of all dividend payments, which implies that a contrast estimator can be defined. In our estimation procedure, the complexity of an estimator is characterized by the class of functions. In other words, our procedure is applicable not only to the optimal dividend problem but also to many statistical inference problems defined as an M-estimation problem.

The rest of the paper is organized as follows. Section 22.2 defines the surplus process following the Lévy insurance risk process and the true optimal dividend barrier as a maximizer of the expectation of the present value of all dividend payments. We define the quasi-process from discretely observed data and show its weak convergence in Sect. 22.3. Then, the optimal dividend barrier estimator is also defined. Section 22.4 shows the consistency of the optimal dividend barrier estimator. To do so, the uniform consistency for the contrast estimator in the function set is shown based on the empirical process theory. In Sect. 22.5, we examine our approach numerically. When the surplus process follows the compound Poisson risk model perturbed by diffusion discussed in [9], it is numerically confirmed that our proposed estimator converges in probability to the true optimal dividend barrier as observed interval goes to 0 and the size of the permutation set goes to infinity. We place all the proofs of the theorems and lemmas in Sect. 22.6.

22.2 Optimal Dividend Barrier

Given a stochastic basis $(\Omega, \mathcal{F}, \mathbb{P}; \mathbb{F})$ with a filtration $\mathbb{F} = (\mathcal{F}_t)_{t \geq 0}$, we consider a $\mathbb{F}-$Lévy process $X = (X_t)_{t \geq 0}$ starting at $X_0 = u$ of the form

$$X_t = u + ct + \sigma W_t - S_t, \tag{22.1}$$

where $u, \sigma \geq 0$, $c > 0$, $W = (W_t)_{t \geq 0}$ is a Wiener process and $S = (S_t)_{t \geq 0}$ is a pure-jump Lévy process, independent of W, with the characteristic exponent

$$\psi_S(\lambda) = \log \mathbb{E}[e^{i\lambda S_1}] = \int_{\mathbb{R}} \left(e^{i\lambda z} - 1 - i\lambda z \mathbf{1}_{\{|z| \leq 1\}}\right) \nu(dz).$$

When $\nu((-\infty, 0)) = 0$ and $\int_{(0,\infty)} (1 \wedge x)\nu(dx) < \infty$, S is called a subordinator, that is, a special class of Lévy processes taking values in $[0, \infty)$ and having non-decreasing paths. Let $\mathbb{D}_\infty := D[0, \infty)$ be a space of càdlàg functions on $[0, \infty)$, and the subset $\tilde{\mathbb{D}}_\infty (\subset \mathbb{D}_\infty)$ be also a space of càdlàg functions on $[0, \infty)$, restricted as follows:

For all $X \in \tilde{\mathbb{D}}_\infty$, X has the form (22.1), where $c > \mathbb{E}[S_1]$ and S is a subordinator with $\int_{(0,\infty)} x^2 \nu(dx) < \infty$.

Then, $\tilde{\mathbb{D}}_\infty$ belongs to the class of spectrally negative Lévy processes with $\mathbb{E}[S_t^2] < \infty$ for all $t > 0$ (see, e.g., [13]). In this paper, we suppose that $X \in \tilde{\mathbb{D}}_\infty$ is an insurance risk process where $u (= X_0)$ is the insurer's initial surplus, c is a given premium rate per unit time, with the net profit condition $c > \mathbb{E}[S_1]$, and $S = (S_t)_{t \geq 0}$ is the aggregate claims process.

Let $\Theta = (u, \bar{\vartheta}) \subset \mathbb{R}$, where $\bar{\vartheta}$ is a known positive value. For an insurance risk process $X \in \tilde{\mathbb{D}}_\infty$ and $\vartheta \in \Theta$, we introduce a process $\xi^\vartheta = (\xi_t^\vartheta)_{t \geq 0}$ by $\xi_0^\vartheta = 0$ and

$$\xi_t^\vartheta = \vartheta \vee \bar{X}_t - \vartheta = \left(\bar{X}_t - \vartheta\right) \vee 0 \quad \text{for } t > 0,$$

where $\bar{X}_t = \sup_{0 \le s < t} X_s$. Note that $\xi^\vartheta \equiv \xi^\vartheta(X)$ is called the dividend strategy consisting of a process with initial value zero, which has paths that are left-continuous, non-negative, non-decreasing and adapted to the filtration of insurance risk process X defined by (22.1). Let $\Xi \equiv \Xi(X) = \left\{\xi^\vartheta(X) | \vartheta \in \Theta\right\}$ be the family of dividend strategies, and for each $\xi^\vartheta \in \Xi$, write $\tau^\vartheta \equiv \tau^\vartheta(X) = \inf\{t > 0 | U_t^\vartheta := X_t - \xi_t^\vartheta < 0\}$ for the time of ruin under the dividend strategy ξ^ϑ. Here we call $U^\vartheta = (U_t^\vartheta)_{t \ge 0}$ the controlled risk process and τ^ϑ the time of ruin for the controlled risk process (see, e.g., Kyprianou [12]); ξ_t^ϑ represents the cumulative dividends that the insurer has paid out until the time t under a dividend strategy in which the dividend payments are continued while the controlled risk process attains ϑ up to the time of ruin τ^ϑ (see, e.g., Loeffen [10]). The expected present value of all dividend payments, with discounting at rate $r > 0$, associated with the dividend strategy ξ^ϑ is given by

$$v(\xi^\vartheta) = \mathbb{E}\left[h^\vartheta(X)\right], \quad h^\vartheta(X) = \int_0^{\tau^\vartheta(X)} e^{-rt} d\xi_t^\vartheta(X). \tag{22.2}$$

Loeffen [10] and Yin, et al. [18] discussed the concavity for $v(\xi^\vartheta)$ under some conditions. We suppose that $v(\xi^\vartheta)$ is a bounded, infinitely differentiable, and strictly concave function with respect to $\vartheta \in \Theta$. Then, for any $\epsilon > 0$, there exists $\vartheta_0 \in \Theta$ such that for all $\vartheta \in \Theta$ satisfying $|\vartheta - \vartheta_0| > \epsilon$, it follows $v(\xi^\vartheta) < v(\xi_0^\vartheta)$. We assume the suitable conditions for $v(\xi^\vartheta)$ in our main theorem (Theorem 22.2). In insurance risk theory, the expected present value of a *ruin-related "loss" up to time of ruin* was often discussed (e.g., Feng [5] and Feng and Shimizu [6]). Among them, the dividend problem discussed in [4] consists of solving the stochastic control problem $v(\xi_*) := \sup_{\xi^\vartheta \in \Xi} \mathbb{E}\left[h^\vartheta(X)\right]$ which corresponds to a optimization problem:

$$\vartheta_0 := \arg\max_{\vartheta \in \bar{\Theta}} \mathbb{E}\left[h^\vartheta(X)\right]. \tag{22.3}$$

In this paper, we consider a statistical estimation problem for ϑ_0 when we observe an insurance risk process $X \in \tilde{\mathbb{D}}_\infty$ discretely.

22.3 Estimation of Optimal Dividend Barrier

Let $\mathcal{D}_\infty$ be the Borel field on $\mathbb{D}_\infty$ generated by the Skorokhod topology. We denote a distribution of X on $\mathcal{D}_\infty$ by $P := \mathbb{P} \circ X^{-1}$ and write

$$Pf := \int_{\mathbb{D}_\infty} f(x) P(dx) = \mathbb{E}[f(X)],$$

for a measurable function $f : \mathbb{D}_\infty \to \mathbb{R}$. Suppose that for a $B \in \mathbb{N}$, random elements $X^{(1)}, X^{(2)}, \dots, X^{(B)}$ are independent copies of process $X \in \tilde{\mathbb{D}}_\infty (\subset \mathbb{D}_\infty)$, and denote its empirical measure as

$$\mathbb{P}_B^* := \frac{1}{B} \sum_{\beta=1}^{B} \delta_{X^{(\beta)}},$$

where δ_x is the delta measure concentrated on $x \in \tilde{\mathbb{D}}_\infty$. In practice, it is often impossible to observe the independent copies of X and to observe the sample path continuously. To overcome these problems, we consider a construction of "multiple quasi-processes" from a discrete sample path. Suppose that we observe a discrete sample path from an insurance risk process $X = (X_t)_{t \geq 0} \in \tilde{\mathbb{D}}_\infty$, where the discrete sample path consists of $\{X_{t_k}\}_{k=0,1,\dots,n}$ with

$$0 = t_0 < t_1 < \cdots < t_n = T, \quad h_n \equiv t_k - t_{k-1}.$$

Let $\mathbb{X} = (\Delta_1 X, \Delta_2 X, \dots, \Delta_n X)$ be a vector of increments with $\Delta_k X := X_{t_k} - X_{t_{k-1}}$, and let

$$\Lambda_n := \left\{ i_m = \begin{pmatrix} 1 & 2 & \cdots & n \\ i_m(1) & i_m(2) & \cdots & i_m(n) \end{pmatrix} \middle| m = 1, 2, \dots, n! \right\}$$

be a family of all the permutations of $(1, 2, \dots, n)$. Since $\Delta_k X$, $1 \leq k \leq n$, are i.i.d. for each n, $\mathbb{X}$ is *exchangeable*, i.e., for any permutation $i \in \Lambda_n$,

$$i(\mathbb{X}) := (\Delta_{i(1)} X, \dots, \Delta_{i(n)} X)$$

has the same distribution as $\mathbb{X}$.

Definition 22.1 For given $\mathbb{X}$ and $i \in \Lambda_n$, a stochastic process $\hat{X}^{i,n} = (\hat{X}_t^{i,n})_{t \geq 0}$ given by

$$\hat{X}_t^{i,n} = u + \sum_{k=1}^{n} \Delta_{i(k)} X \cdot \mathbf{1}_{[t_k, \infty)}(t)$$

is said to be a *quasi-process* of X for a permutation $i \in \Lambda_n$.

Note that a path of the quasi-process $\hat{X}^{i,n} = (\hat{X}_t^{i,n})_{t \geq 0}$ belongs to $\mathbb{D}_\infty$ (but not to $\tilde{\mathbb{D}}_\infty$), a right continuous step function that has a jump at $t = t_k$ $(k = 1, 2, \dots, n)$ with the amplitude $\Delta_{i(k)} X$. For the discrete sampling scheme, we impose the following assumption.

Assumption 22.1 (High-Frequency sampling in the Long Term; HFLT) $h_n \to 0$ and $T = n h_n \to \infty$ as $n \to \infty$.

Shimizu and Shiraishi [15] showed the followings under HFLT.

Theorem 22.1 *Under Assumption 22.1, we have, for any sequence of permutations* $\{i^n\} \subset \Lambda_n$,

$$\hat{X}^{i^n,n} \rightsquigarrow X \quad \text{in } \mathbb{D}_\infty \quad \text{as } n \to \infty.$$

For a given size $\alpha_n (\le n!)$, let $A_n := \{i_{(1)}, \dots, i_{(\alpha_n)}\}$ be a set of i.i.d. samples drawn uniformly from Λ_n, i.e., for a given $m = 1, 2, \dots, n!$,

$$\mathbb{P}(i_{(k)} = i_m) = \frac{1}{n!} \quad \text{for every } k = 1, 2, \dots, n!.$$

Based on A_n, we introduce two empirical measures $\mathbb{P}^*_{\alpha_n}$ and $\mathbb{P}_{\alpha_n}$ by

$$\mathbb{P}^*_{\alpha_n} := \frac{1}{\alpha_n} \sum_{k=1}^{\alpha_n} \delta_{X^{(k)}}, \quad \mathbb{P}_{\alpha_n} := \frac{1}{\alpha_n} \sum_{k=1}^{\alpha_n} \delta_{\hat{X}^{i_{(k)},n}}.$$

Next, we propose an estimator of ϑ_0 defined by (22.3) based on the empirical measure of the quasi-process.

Definition 22.2 Given a vector of increments $\mathbb{X}$ and permutation sets $\{A_n\}_{n\in\mathbb{N}}$, we denote a maximum contrast estimator of ϑ_0 defined by (22.3) as

$$\hat{\vartheta}_n = \arg\max_{\vartheta \in \bar{\Theta}} \mathbb{P}_{\alpha_n} h^\vartheta,$$

where

$$\mathbb{P}_{\alpha_n} h^\vartheta = \frac{1}{\alpha_n} \sum_{i \in A_n} h^\vartheta(\hat{X}^{i,n}) = \frac{1}{\alpha_n} \sum_{i \in A_n} \int_0^{\tau^\vartheta(\hat{X}^{i,n})} e^{-rt} d\xi_t^\vartheta(\hat{X}^{i,n}).$$

For the moments of $h^\vartheta(\hat{X}^{i,n})$, we have following result.

Lemma 22.1 *Under Assumption 22.1, we have, for any* $\vartheta \in \Theta$ *and* $i \in \Lambda_n$,

$$\mathbb{E}\left[h^\vartheta(\hat{X}^{i,n})^m\right] = O(1), \quad m = 1, 2.$$

22.4 Asymptotic Results

Our main result in this paper is to provide the consistency for $\hat{\vartheta}_n$ defined in Definition 22.2. To do so, we assume that a size of permutation sets $\alpha_n := \sharp A_n$ satisfies followings.

Assumption 22.2 (Size of permutation sets) $\frac{n}{\alpha_n} \to 0$ as $n \to \infty$.

We recall that the empirical measure of the quasi-process $\mathbb{P}_{\alpha_n}$ is asymptotically equivalent in law with $\mathbb{P}^*_{\alpha_n}$ based on the independent copy $X^{(1)}, \dots, X^{(\alpha_n)}$ of the process X. Moreover, we introduce a sequence of a family of measurable functions $\mathcal{H} = \{\mathcal{H}_n\}_{n\in\mathbb{N}}$ on $\tilde{\mathbb{D}}_\infty$, where $\mathcal{H}_n$ is a family of measurable functions $h_n^\vartheta : \tilde{\mathbb{D}}_\infty \to \mathbb{R}$ for each $\vartheta \in \Theta$, given by

$$h_n^\vartheta(X) = \int_0^{\tau_n^\vartheta(X)} e^{-rt} d\xi_{n,t}^\vartheta(X), \tag{22.4}$$

where $\tau_n^\vartheta(X) = \tau^\vartheta(\hat{X}^{i_{\rm id},n})$ and $\xi_{n,t}^\vartheta(X) = \xi_t^\vartheta(\hat{X}^{i_{\rm id},n})$ for all $X \in \tilde{\mathbb{D}}_\infty$, $\vartheta \in \Theta$ and $n \in \mathbb{N}$. Here $i_{\rm id} \in \Lambda_n$ is an identical permutation, i.e., $i_{\rm id} = \begin{pmatrix} 1\ 2 \cdots n \\ 1\ 2 \cdots n \end{pmatrix}$. For the class $\mathcal{H}_n = \{h_n^\vartheta : \tilde{\mathbb{D}}_\infty \to \mathbb{R} | \vartheta \in \Theta\}$, we denote by $N(\epsilon, \mathcal{H}_n, L^1(\mathbb{P}_{\alpha_n}))$ the covering number of $L^1(\mathbb{P}_{\alpha_n})$ which is the minimum number of ϵ-balls needed to cover $\mathcal{H}_n$, where an ϵ-ball around a function $g \in L^1(\mathbb{P}_{\alpha_n})$ being the set $\{h_n^\vartheta \in L^1(\mathbb{P}_{\alpha_n}) |\ \|h_n^\vartheta - g\|_{\mathbb{P}_{\alpha_n},1} = \mathbb{P}_{\alpha_n}(|h_n^\vartheta - g|) < \epsilon\}$, with $\|\cdot\|_{\mathbb{P}_{\alpha_n},1}$ being the $L^1(\mathbb{P}_{\alpha_n})$-norm. In addition, we denote by $N_{[]}(\epsilon, \mathcal{H}_n, L^1(\mathbb{P}_{\alpha_n}))$ the bracketing number which is the minimum number of ϵ-brackets in $L^1(\mathbb{P}_{\alpha_n})$ needed to ensure that every $h_n^\vartheta \in \mathcal{H}_n$ lines in at least one bracket, where an ϵ-bracket in $L^1(\mathbb{P}_{\alpha_n})$ is a pair of functions $l, u \in L^1(\mathbb{P}_{\alpha_n})$ with $\mathbb{P}_{\alpha_n}\left(\mathbf{1}_{\{l(X)\le u(X)\}}\right) = 1$ and $\|l - u\|_{\mathbb{P}_{\alpha_n},1} \le \epsilon$. For the covering number and bracketing number, we have the following result.

Lemma 22.2 *Under Assumptions 22.1 and 22.2, we have, for any* $\epsilon > 0$

$$\mathbb{E}\left[N(\epsilon, \mathcal{H}_n, L^1(\mathbb{P}_{\alpha_n}))\right] = o(\alpha_n) \quad \text{and} \quad \mathbb{E}\left[N_{[]}(\epsilon, \mathcal{H}_n, L^1(\mathbb{P}_{\alpha_n}))\right] = o(\alpha_n).$$

This lemma implies that both of the covering number and bracketing number diverge slower than α_n. This result is applied for the proof of uniformly consistency below. The following result is due to a slight modification by Kosorok [11, Theorem 8.15].

Lemma 22.3 *Under Assumptions 22.1 and 22.2, we have*

$$\sup_{\vartheta\in\bar{\Theta}} |(\mathbb{P}_{\alpha_n} - P)h_n^\vartheta| =: \|\mathbb{P}_{\alpha_n} - P\|_{\mathcal{H}_n} \xrightarrow{p} 0.$$

This lemma shows that two measure $\mathbb{P}_{\alpha_n}$ and P are asymptotically equivalent on the function space $\mathcal{H}_n$. On the other hand, the true optimal dividend barrier ϑ_0 defined by (22.3) is a maximizer of $\mathbb{E}\left[h^\vartheta(X)\right] = Ph^\vartheta$, where $h^\vartheta \notin \mathcal{H}_n$. Hence, we have to evaluate the difference between h^ϑ and $h_n^\vartheta \in \mathcal{H}_n$ defined by (22.2) and (22.4) based on the measure P. The following lemma is also applied for the proof of our main result.

Lemma 22.4 *Under Assumptions 22.1 and 22.2, we have*

$$\sup_{\vartheta \in \bar{\Theta}} |P(h_n^\vartheta - h^\vartheta)| \to 0.$$

By Lemmas 22.3 and 22.4, we can show the consistency for $\hat{\vartheta}_n$, as follows.

Theorem 22.2 *Suppose that Assumptions 22.1 and 22.2 are hold, and that there exists $\vartheta_0 \in \Theta$ such that, for any $\epsilon > 0$,*

$$\sup_{\vartheta \in \bar{\Theta}: |\vartheta - \vartheta_0| > \epsilon} Ph^\vartheta < Ph^{\vartheta_0}. \tag{22.5}$$

Then, $\hat{\vartheta}_n$ is weakly consistent to ϑ_0, i.e.,

$$\hat{\vartheta}_n \xrightarrow{p} \vartheta_0, \quad n \to \infty.$$

22.5 Numerical Results

In this section, we present simulation results to evaluate the finite-sample performance of the proposed estimator of the optimal dividend barrier based on the discrete sample from spectrally negative Lévy processes. We consider the following data generating process (DGP), sampling scheme, permutation set, and discount rate.

- **DGP** (**Brownian motion + compound Poisson process**): Let $X_t = u + ct + \sigma W_t - S_t$, where $u = 10,\ c = 15,\ \sigma = 2$, $W = (W_t)_{t \geq 0}$ is a standard Brownian motion, and $S_t = \sum_{r=1}^{N_t} \xi_r$ is a compound Poisson process. The Poisson process $N = (N_t)_{t \geq 0}$ has intensity $\lambda > 0$. We set $\lambda = 5$. The jump size $\{\xi_r\}$ is a sequence of i.i.d. random variables having exponential distribution with parameter $1/2$, that is, $\mathbb{E}[\xi_r] = 2$.
- **Sampling scheme**: We consider the sampling interval $h_n = 1, 0.1, 0.01, 0.001$ and the terminal $T = 100$ (fixed), which implies that sample size n is $n = T/h_n = 100, 1000, 10000$, respectively.
- **Permutation set**: We consider the subset of permutation set $A_n = \{i_{m_j} | j = 1, \ldots, \alpha_n\} \subset \Lambda_n$ with $\alpha_n = 10, 100, 1000$, where the suffix m_j is independently selected with same probability from $\{1, 2, \ldots, n!\}$.
- **Discount rate**: We set $r = 0.2$.

22.5.1 Quasi-process

We first examine the finite-sample performance of the quasi-process $\hat{X}^{i,n} = (\hat{X}^{i,n}_t)_{t \geq 0}$ for each h_n and α_n. Figure 22.1 shows 100 sample paths for the risk process of an insurance business $X = (X_t)_{t \geq 0}$ defined above. It looks that we cannot know the distribution of X only from one sample path without any additional assumption. In this study, we consider such a situation. When we suppose that only one sample path is observed discretely, we would like to know its distribution. In Fig. 22.1, the blue line is observed discretely. Under the sampling scheme defined above, we can construct a number of sample paths of the quasi-process $\hat{X}^{i,n}$ from one sample path. Then, we can approximate the distribution of X based on these sample paths. In Fig. 22.2, the blue line is an observed (but discretely) sample path from the stochastic process X (this is the same as Fig. 22.1). From this sample path, we construct α_n sample paths of the quasi-process $\hat{X}^{i,n}$ based on Definition 22.1. Each sample path depends on the observed sample path and the permutation $i \in A_n \subset \Lambda_n$. The top figure shows the case of $h_n = 1$ and $\alpha_n = 100$, and the bottom figure shows the case of $h_n = 0.001$ and $\alpha_n = 100$. It looks that the top figure is not, but the bottom figure is well approximated by the distribution of X. This phenomenon comes from the exchangeability of the increments of the Lévy processes and if the sampling interval h_n is sufficiently small and the size of the permutation set α_n is sufficiently large, we can well approximate the distribution of X even from only one sample path.

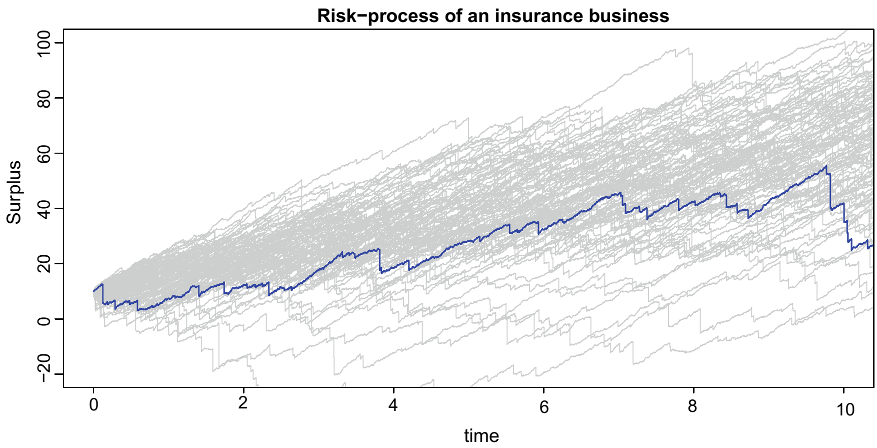


Fig. 22.1 100 sample paths for the risk process of an insurance business X

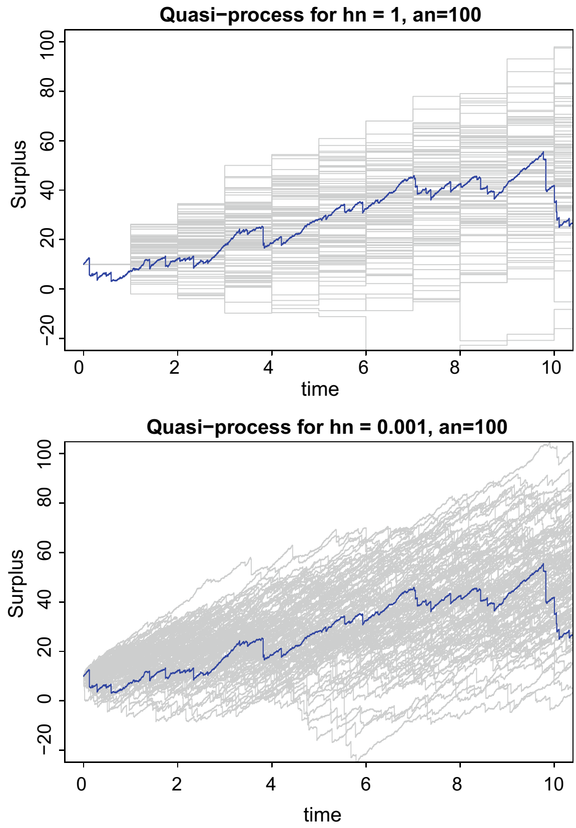


Fig. 22.2 (Discretely) observed sample path (blue line) and $\alpha_n (= 100)$ sample paths for the quasi-process (top ($h_n = 1$) and bottom ($h_n = 0.001$))

22.5.2 Maximum Contrast Estimator

Next, we examine the behavior of the objective function $h_n^\vartheta(\hat{X}^{i,n})$. Given a sample path of the quasi-process $\hat{X}^{i,n}$, we can construct $h_n^\vartheta(\hat{X}^{i,n})$ as a function of the parameter ϑ. Since each sample path of the quasi-process $\hat{X}^{i,n} = (\hat{X}^{i,n})_{t\geq 0}$ is locally constant on time t, we can write

$$h_n^\vartheta(\hat{X}^{i,n}) = \int_0^{\tau_n^\vartheta(\hat{X}^{i,n})} e^{-rt} d\xi_{n,t}^\vartheta(\hat{X}^{i,n}) = \sum_{k=1}^n \mathbf{1}_{\{\tau_n^\vartheta(\hat{X}^{i,n})>t_k\}} e^{-rt_k} \Delta_k \xi_n^\vartheta(\hat{X}^{i,n}),$$

where $\Delta_k \xi_n^\vartheta(\hat{X}^{i,n}) = \xi_{n,t_k}^\vartheta(\hat{X}^{i,n}) - \xi_{n,t_{k-1}}^\vartheta(\hat{X}^{i,n})$ with $\xi_{n,t_0}^\vartheta(\hat{X}^{i,n}) = 0$. Figure 22.3 shows the plots of $h_n^\vartheta(\hat{X}^{i,n})$ for five sample paths of the quasi-process $\hat{X}^{i,n}$. In this figure, the horizontal axis represents the magnitude of ϑ and the vertical axis represents the magnitude of h_n^ϑ. Under a fixed sample path, it can be seen that h_n^ϑ is a locally decreasing function with some positive jumps. The locally decreasing prop-

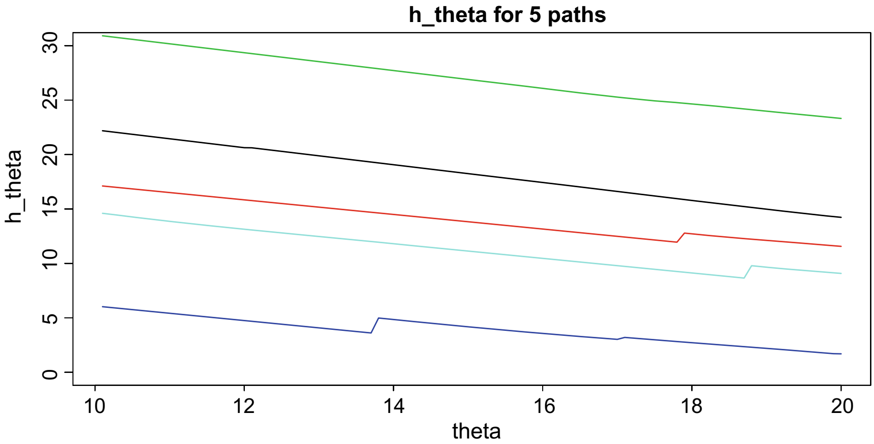


Fig. 22.3 Plots of h_n^{ϑ} for 5 sample paths of the quasi-process

erty is that the total dividend amount tends to decrease as the dividend barrier ϑ increases while the ruin time τ_n^{ϑ} is fixed. On the other hand, the existence of a positive jump shows that the total dividend amount increases discontinuously since the ruin time is extended at some ϑ.

Figure 22.4 shows the behavior of the contrast function $\mathbb{P}_{\alpha_n} h_n^{\vartheta} = \frac{1}{\alpha_n}\sum_{i\in A_n} h_n^{\vartheta}(\hat{X}^{i,n})$ for some h_n and α_n. The left figure shows plots of the contrast function for the size of the permutation set $\alpha_n = 5, 20, 100$ under fixed sampling interval $h_n = 1$, and the dotted line shows its maximization point. It can be seen that the function approaches the true function as α_n increases which implies that the maximization point tends to the true maximation point, that is, our proposed estimator $\hat{\vartheta}_n$ converges to the true optimal dividend barrier ϑ_0. On the other hand, the right figure shows plots of the contrast function for $h_n = 1, 0.1, 0.01$ under fixed $\alpha_n = 100$. It can be seen that the function approaches the true function as h_n decreases which implies that our estimator converges to the true optimal dividend barrier. In both figures, the black line represents the true objective function $\mathbb{E}[h^{\vartheta}(X)]$ (see, e.g., [9]). These figures confirm the validity of the theoretical result in Theorem 22.2.

22.5.3 Simulation Result

Now, we examine mean, standard deviation (std), bias, and MSE for $h_n = 1, 0.1, 0.01$ and $\alpha_n = 10, 100, 1000$. We generate 100 replications for each run of the simulations. Figure 22.5 shows the box-plot for estimated values $\hat{\vartheta}_n^{(j)}$ for each α_n and h_n. The left figure is the case of $\alpha_n = 10$, the middle figure is the case of $\alpha_n = 100$, and the right figure is the case of $\alpha_n = 1000$. In each figure, the left box is the case of $h_n = 1$, the middle box is the case of $h_n = 0.1$, and the right box is the case of

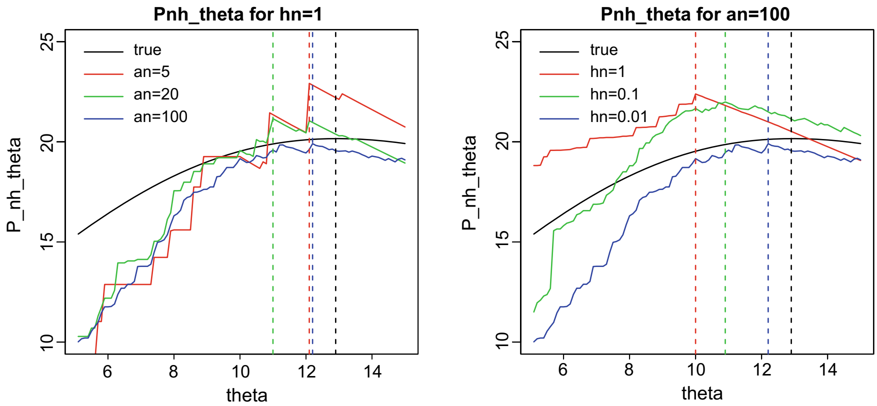


Fig. 22.4 Plots of the contrast function for the size of the permutation set $\alpha_n = 5, 20, 100$ under fixed sampling interval $h_n = 1$ (left), and for $h_n = 1, 0.1, 0.01$ under fixed $\alpha_n = 100$ (right). The dotted line shows these maximization points, and the black line represents the true objective function $\mathbb{E}[h^\vartheta(X)]$

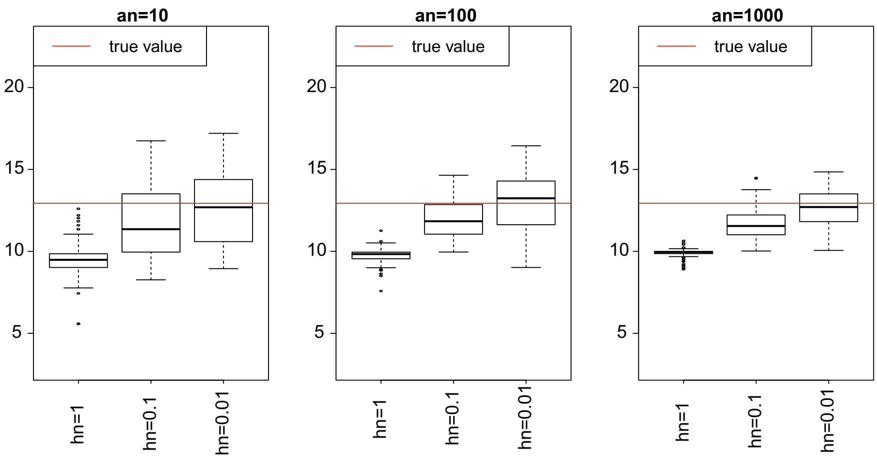


Fig. 22.5 Box plots for the estimated values of the optimal dividend barrier for $\alpha_n = 10$ (left), $\alpha_n = 100$ (middle), and $\alpha_n = 1000$ (right). In each figure, $h_n = 1$ (left), $h_n = 0.1$ (middle), and $h_n = 0.01$ (right). The red line shows the true value

$h_n = 0.01$. The red line shows the true value $\vartheta_0 = 12.93958$. In view of the median (and mean) of the estimated values, it can be seen that the value converges to the true value as α_n increases and h_n decreases. On the other hand, in view of the dispersion, it looks that the estimated values shrink as α_n increases. However, when h_n is not sufficiently small, it seems that the estimated values converge to a value different from the true value. This phenomenon indicates that it is a warning that an asymptotic bias will occur unless the sampling interval h_n is sufficiently small.

Table 22.1 Mean, std, bias, and MSE for the estimated values in 100 replications where the true optimal dividend barrier is $\vartheta_0 = 12.93958$

α_n	h_n	Mean	std	bias	MSE
10	1	9.482	1.039	–3.45	13.03
	0.1	11.778	2.106	–1.16	5.78
	0.01	12.560	2.126	–0.37	4.66
100	1	9.681	0.489	–3.25	10.85
	0.1	12.009	1.290	–0.92	2.52
	0.01	12.964	1.558	0.02	2.42
1000	1	9.895	0.260	–3.04	9.33
	0.1	11.680	0.998	–1.25	2.58
	0.01	12.680	1.183	–0.25	1.46

Table 22.1 shows mean $\left(\mu_n := \frac{1}{B}\sum_{j=1}^{B}\hat{\vartheta}_n^{(j)}\right)$, std $\left(\sigma_n := \sqrt{\frac{1}{B}\sum_{j=1}^{B}(\hat{\vartheta}_n^{(j)} - \vartheta_0)^2}\right)$, bias $\left(\frac{1}{B}\sum_{j=1}^{B}(\hat{\vartheta}_n^{(j)} - \vartheta_0)\right)$, and MSE $\left(\mu_n^2 + \sigma_n^2\right)$ of the estimated values $\hat{\vartheta}_n^{(j)}$ for $\alpha_n = 10, 100, 1000$, $h_n = 1, 0.1, 0.01$ and $B = 100$. It can be seen that the mean converges to the true value, and the MSE converges to 0 as α_n increases and h_n decreases. Note that the bias tends to have a negative value which implies that the distribution of the estimator tends to be asymmetric. From the insurer's point of view, this phenomenon is a warning because setting a lower dividend barrier poses a risk to insurers.

22.6 Proofs

This section gives proofs of lemmas and theorems.

22.6.1 Proof of Lemma 22.1

From the definition, we can write

$$h^{\vartheta}(\hat{X}^{i,n}) = \int_0^{\tau^{\vartheta}(\hat{X}^{i,n})} e^{-rt} d\xi_t^{\vartheta}(\hat{X}^{i,n}) \le \int_0^{\infty} e^{-rt} d\overline{\hat{X}}_t^{i,n}$$
$$= \sum_{k=1}^{n} e^{-rt_k}(\Delta_{i(k)}X \vee 0) \le \sum_{k=1}^{n} e^{-rt_k}|\Delta_{i(k)}X|,$$

where $\overline{\hat{X}}_t^{i,n} = \sup_{0\le s<t} \hat{X}_s^{i,n}$ and $\{|\Delta_{i(k)}X|\}_{k=1,\dots,n}$ is a sequence of i.i.d. random variables. From (22.1), we have

$$|\Delta_{i(k)}X| \stackrel{d}{=} |\Delta_k X| = |X_k - X_{k-1}| \stackrel{d}{=} |ch_n + \sigma W_{h_n} - S_{h_n}| \le ch_n + \sigma |W_{h_n}| + S_{h_n}.$$

It is easy to see $\mathbb{E}\left[|W_{h_n}|^k\right] \lesssim h_n$. Since $\varphi_n(\lambda) := \mathbb{E}\left[e^{i\lambda S_{h_n}}\right] = e^{h_n \psi_S(\lambda)}$, we have

$$\mathbb{E}\left[S_{h_n}\right] = i^{-1}\frac{d\varphi_n(\lambda)}{d\lambda}\Big|_{\lambda=0} = h_n \mathbb{E}[S_1] \lesssim h_n,$$
$$\mathbb{E}\left[S_{h_n}^2\right] = -\frac{d^2\varphi_n(\lambda)}{(d\lambda)^2}\Big|_{\lambda=0} = h_n \mathbb{E}[S_1^2] - h_n^2 \mathbb{E}[S_1]^2 \lesssim h_n,$$

which imply that $\mathbb{E}\left[|\Delta_{i(k)}X|^m\right] \lesssim h_n$ for $m = 1, 2$. By the Taylor expansion, we have

$$\sum_{k=1}^n e^{-rt_k} = \frac{e^{-rh_n}(1-e^{-rnh_n})}{1-e^{-rh_n}} \lesssim \frac{1}{1-e^{-rh_n}} \lesssim h_n^{-1}.$$

Therefore, we have

$$\mathbb{E}\left[h^\vartheta(\hat{X}^{i,n})\right] \le \sum_{k=1}^n e^{-rt_k}\mathbb{E}\left[|\Delta_{i(k)}X|\right] = O(h_n^{-1})O(h_n) = O(1),$$
$$\mathbb{E}\left[h^\vartheta(\hat{X}^{i,n})^2\right] \le \sum_{k=1}^n e^{-2rt_k}\mathbb{E}\left[|\Delta_{i(k)}X|^2\right] = O(h_n^{-1})O(h_n) = O(1).$$

22.6.2 *Proof of Lemma 22.2*

By definition, $\tau_n^\vartheta \in \{t_1, t_2, \dots, t_n\}$ for any $X \in \tilde{D}_\infty$ and $\vartheta \in \Theta$. Note that if $U_{t_n}^\vartheta(\hat{X}^{i_{\mathrm{id}},n}) := \hat{X}_{t_n}^{i_{\mathrm{id}},n} - \xi_{t_n}^\vartheta(\hat{X}^{i_{\mathrm{id}},n}) \ge 0$, we define $\tau^\vartheta(\hat{X}^{i_{\mathrm{id}},n}) = \tau_n^\vartheta(X) = t_n$. This implies that we can divide $\mathcal{H}_n$ into $\mathcal{H}_{n,k}$ $(k = 1, \dots, n)$, where $\mathcal{H}_{n,k} = \{h_{n,k}^\vartheta, \vartheta \in \Theta\}$ with $h_{n,k}^\vartheta(X) = \int_0^{t_k} e^{-rt} d\xi_{n,t}^\vartheta(X)$. For a fixed $X \in \tilde{\mathbb{D}}_\infty$, it can be seen that $\xi_{n,t}^\vartheta(X) \le \xi_{n,t}^{\vartheta'}(X)$ for any $n \in \mathbb{N}$, $t > 0$ if $\vartheta \ge \vartheta'$, which implies that

$$\mathbb{P}_{\alpha_n}\left(\mathbf{1}_{\left\{h_{n,k}^{\bar{\vartheta}}(X) \le h_{n,k}^\vartheta(X) \le h_{n,k}^u(X)\right\}}\right) = 1, \quad \forall \vartheta \in \Theta = (u, \bar{\vartheta}).$$

In addition, since $u \le \overline{\hat{X}}_{t_{l-1}}^{i_{\mathrm{id}},n} \le \overline{\hat{X}}_{t_l}^{i_{\mathrm{id}},n}$ $(l = 1, \dots, n)$, where $\overline{\hat{X}}_t^{i_{\mathrm{id}},n} = \sup_{0\le s<t} \hat{X}_s^{i_{\mathrm{id}},n}$, we can write

$$
\begin{aligned}
&\left|\left\{\xi^{u}_{n,t_l}(\hat{X}^{i_{\mathrm{id}},n})-\xi^{u}_{n,t_{l-1}}(\hat{X}^{i_{\mathrm{id}},n})\right\}-\left\{\xi^{\bar\vartheta}_{n,t_l}(\hat{X}^{i_{\mathrm{id}},n})-\xi^{\bar\vartheta}_{n,t_{l-1}}(\hat{X}^{i_{\mathrm{id}},n})\right\}\right|\\
&=\left|\left(u\vee\overline{\hat{X}}^{i_{\mathrm{id}},n}_{t_{l-1}}\right)-\left(u\vee\overline{\hat{X}}^{i_{\mathrm{id}},n}_{t_l}\right)-\left(\bar\vartheta\vee\overline{\hat{X}}^{i_{\mathrm{id}},n}_{t_{l-1}}\right)+\left(\bar\vartheta\vee\overline{\hat{X}}^{i_{\mathrm{id}},n}_{t_l}\right)\right|\\
&=\mathbf{1}_{\{\overline{\hat{X}}^{i_{\mathrm{id}},n}_{t_{l-1}}\le\bar\vartheta\}}\left|\left(\overline{\hat{X}}^{i_{\mathrm{id}},n}_{t_l}-\overline{\hat{X}}^{i_{\mathrm{id}},n}_{t_{l-1}}\right)\wedge\left(\bar\vartheta-\overline{\hat{X}}^{i_{\mathrm{id}},n}_{t_{l-1}}\right)\right|\\
&\le|\Delta_l X|,
\end{aligned}
$$

which implies that

$$
\begin{aligned}
\|h^u_{n,k}-h^{\bar\vartheta}_{n,k}\|_{\mathbb{P}_{\alpha_n},1}&=\mathbb{P}_{\alpha_n}\left(|h^u_{n,k}(X)-h^{\bar\vartheta}_{n,k}(X)|\right)\\
&=\mathbb{P}_{\alpha_n}\left(\left|\int_0^{t_k}e^{-rt}\left\{d\xi^u_{n,t}(X)-d\xi^{\bar\vartheta}_{n,t}(X)\right\}\right|\right)\\
&\le\sum_{l=1}^{k}e^{-rt_l}\mathbb{P}_{\alpha_n}\Bigg(\bigg|\left\{\xi^{u}_{n,t_l}(\hat{X}^{i_{\mathrm{id}},n})-\xi^{u}_{n,t_l}(\hat{X}^{i_{\mathrm{id}},n})\right\}\\
&\qquad-\left\{\xi^{\bar\vartheta}_{n,t_l}(\hat{X}^{i_{\mathrm{id}},n})-\xi^{\bar\vartheta}_{n,t_l}(\hat{X}^{i_{\mathrm{id}},n})\right\}\bigg|\Bigg)\\
&\le\sum_{l=1}^{k}e^{-rt_l}\mathbb{P}_{\alpha_n}\left(|\Delta_l X|\right)\\
&\overset{d}{=}\left(\sum_{l=1}^{k}e^{-rt_l}\right)|\Delta_1 X|.
\end{aligned}
$$

Since $\sum_{l=1}^{k}e^{-rt_l}=O(h_n^{-1})$ and $\mathbb{E}\left[|\Delta_1 X|\right]=O(h_n)$, we have for any $\epsilon>0$

$$
\begin{aligned}
\sum_{k=1}^{n}\mathbb{E}\left[\frac{\|h^u_{n,k}-h^{\bar\vartheta}_{n,k}\|_{\mathbb{P}_{\alpha_n},1}}{\epsilon}\right]&\le\frac{1}{\epsilon}\sum_{k=1}^{n}\left(\sum_{l=1}^{k}e^{-rt_l}\right)\mathbb{E}\left[|\Delta_1 X|\right]\\
&=\frac{n}{\epsilon}O(h_n^{-1})O(h_n)=O(n).
\end{aligned}
$$

Note that the $L^1(\mathbb{P}_{\alpha_n})$-size of the brackets is bounded by ϵ, which implies that

$$
\begin{aligned}
\mathbb{E}\left[N_{[]}(\epsilon,\mathcal{H}_n,L^1(\mathbb{P}_{\alpha_n}))\right]&=\sum_{k=1}^{n}\mathbb{E}\left[N_{[]}(\epsilon,\mathcal{H}_{n,k},L^1(\mathbb{P}_{\alpha_n}))\right]\\
&\le\sum_{k=1}^{n}\mathbb{E}\left[\frac{\|h^u_{n,k}-h^{\bar\vartheta}_{n,k}\|_{\mathbb{P}_{\alpha_n},1}}{\epsilon}+1\right].
\end{aligned}
$$

Therefore, from Assumption 22.2, it follows that

$$\mathbb{E}\left[N_{[]}(\epsilon, \mathcal{H}_n, L^1(\mathbb{P}_{\alpha_n}))\right] = o(\alpha_n).$$

From the relationship between bracketing number and covering number (cf., Kosorok [11, Lemma 9.18]), we have

$$\mathbb{E}\left[N(\epsilon, \mathcal{H}_n, L^1(\mathbb{P}_{\alpha_n}))\right] \leq \mathbb{E}\left[N_{[]}(\epsilon, \mathcal{H}_n, L^1(\mathbb{P}_{\alpha_n}))\right] = o(\alpha_n).\square$$

22.6.3 Proof of Lemma 22.3

By the symmetrization result (cf., Kosorok [11, Theorem 8.8]), we can write

$$\begin{aligned}\mathbb{E}\left[\|\mathbb{P}_{\alpha_n} - P\|_{\mathcal{H}_n}\right] &\leq 2\mathbb{E}_X\left[\mathbb{E}_\epsilon\left[\sup_{\vartheta\in\Theta}\left|\mathbb{P}_{\alpha_n}\left(\epsilon h_n^\vartheta(X)\right)\right|\Big| X\right]\right] \\ &= 2\mathbb{E}_X\left[\mathbb{E}_\epsilon\left[\sup_{\vartheta\in\Theta}\left|\frac{1}{\alpha_n}\sum_{i\in A_n}\epsilon^{(i)}h_n^\vartheta(X^{(i)})\right|\Bigg| X^{(i)}, i\in A_n\right]\right],\end{aligned}$$

where $\{\epsilon^{(i)}\}$ is a sequence of independent Rademacher random variables which are independent of $\{X^{(i)}\}$ and satisfy $\mathbb{P}(\epsilon^{(i)} = -1) = \mathbb{P}(\epsilon^{(i)} = 1) = 1/2$, and $\mathbb{E}_X, \mathbb{E}_\epsilon$ are the expectations with respect to $X^{(i)}, \epsilon^{(i)}$, respectively. For any fixed $n \in \mathbb{N}$, $\delta > 0$ and $\{X^{(i)}\}_{i\in A_n}$, let $\mathcal{H}_{n,j}$ $(j = 1, \ldots, N(\delta, \mathcal{H}_n, L^1(\mathbb{P}_{\alpha_n})))$ be a sequence of finite δ-balls in $L^1(\mathbb{P}_{\alpha_n})$ over $\mathcal{H}_n$ (i.e., $\mathcal{H}_{n,j}$ is a subset of $\mathcal{H}_n$ and for any $h_n^\vartheta, h_n^{\vartheta'} \in \mathcal{H}_{n,j}$, $\|h_n^\vartheta - h_n^{\vartheta'}\|_{\mathbb{P}_{\alpha_n},1} < \delta$ and $\cup_j\mathcal{H}_{n,j} \supset \mathcal{H}_n$). For each $\mathcal{H}_{n,j}$, we fix ϑ_j (satisfying $\vartheta_j \neq \vartheta_{j'}$ if $j \neq j'$) which is a representative $h_n^{\vartheta_j}$ such that for any $h_n^\vartheta \in \mathcal{H}_{n,j}$

$$\begin{aligned}&\mathbb{E}_\epsilon\left[\left|\frac{1}{\alpha_n}\sum_{i\in A_n}\epsilon^{(i)}h_n^\vartheta(X^{(i)})\right|\Bigg| X^{(i)}, i\in A_n\right] \\ &\leq \mathbb{E}_\epsilon\left[\left|\frac{1}{\alpha_n}\sum_{i\in A_n}\epsilon^{(i)}h_n^{\vartheta_j}(X^{(i)})\right|\Bigg| X^{(i)}, i\in A_n\right] + \delta,\end{aligned}$$

which implies that

$$\begin{aligned}&\mathbb{E}_\epsilon\left[\sup_{\vartheta\in\Theta}\left|\frac{1}{\alpha_n}\sum_{i\in A_n}\epsilon^{(i)}h_n^\vartheta(X^{(i)})\right|\Bigg| X^{(i)}, i\in A_n\right] \\ &\leq \mathbb{E}_\epsilon\left[\max_j\left|\frac{1}{\alpha_n}\sum_{i\in A_n}\epsilon^{(i)}h_n^{\vartheta_j}(X^{(i)})\right|\Bigg| X^{(i)}, i\in A_n\right] + \delta. \qquad (22.6)\end{aligned}$$

Let $Z_j = \frac{1}{\alpha_n}\sum_{i\in A_n} \epsilon^{(i)} h_n^{\vartheta_j}(X^{(i)})$ and $Z = \max_j |Z_j|$. Then,

$$\|Z\|_{1|X} := \mathbb{E}_\epsilon[|Z|\ |X^{(i)}, i \in A_n] \le \mathbb{E}_\epsilon[Z^2|X^{(i)}, i \in A_n]^{1/2} =: \|Z\|_{2|X},$$

from Jensen's inequality. On the other hand, based on the nondecreasing, nonzero convex function $\psi_2(x) = \exp(x^2) - 1$, we introduce the Orlicz-norm

$$\|Z\|_{\psi_2|X} := \inf\left\{c > 0 \middle| \mathbb{E}_\epsilon\left[\frac{\psi_2(|Z|)}{c}\middle| X^{(i)}, i \in A_n\right] \le 1\right\},$$

for which $\|Z\|_{2|X} \le \|Z\|_{\psi_2|X}$. Applying the maximal inequality (cf., Kosorok [11, Lemma 8.2]), we have

$$\|Z\|_{\psi_2|X} = \|\max_j |Z_j|\|_{\psi_2|X} \le K\psi_2^{-1}(N(\delta, \mathcal{H}_n, L^1(\mathbb{P}_{\alpha_n})))\max_j \|Z_j\|_{\psi_2|X},$$

where the constant K depends only on ψ_2, which implies that the left-hand-side of (22.6) is bounded by

$$\sqrt{\log\left\{1 + N(\delta, \mathcal{H}_n, L^1(\mathbb{P}_{\alpha_n}))\right\}}\max_j \left\|Z_j\right\|_{\psi_2|X} + \delta,$$

up to a constant. By Hoeffding's inequality (cf., Kosorok [11, Lemma 8.7]), we have

$$\mathbb{E}_\epsilon\left[\mathbf{1}_{\{|Z_j|>x\}}\middle| X^{(i)}, i \in A_n\right] = \mathbb{P}_\epsilon\left(|Z_j| > x\middle| X^{(i)}, i \in A_n\right) \le 2\exp\left(-\frac{1}{2}x^2/\|Z_j\|_\epsilon^2\right),$$

for any $x > 0$ and each j, where $\|Z_j\|_\epsilon = \mathbb{E}_\epsilon[|Z_j|\ |X^{(i)}, i \in A_n]$. Hence, from Kosorok ([11, Lemma 8.1]) and Jensen's inequality,

$$\begin{aligned}
\left\|Z_j\right\|_{\psi_2|X} &\le \left(\frac{1+2}{1/(2\|Z_j\|_\epsilon^2)}\right)^{1/2} \\
&= \sqrt{6}\|Z_j\|_\epsilon \\
&\le \frac{\sqrt{6}}{\alpha_n}\mathbb{E}_\epsilon\left[\left|\sum_{i\in A_n}\epsilon^{(i)}h_n^{\vartheta_j}(X^{(i)})\right|\middle| X^{(i)}, i \in A_n\right] \\
&\le \frac{\sqrt{6}}{\alpha_n}\left\{\mathbb{E}_\epsilon\left[\left|\sum_{i\in A_n}\epsilon^{(i)}h_n^{\vartheta_j}(X^{(i)})\right|^2\middle| X^{(i)}, i \in A_n\right]\right\}^{1/2} \\
&= \sqrt{\frac{6}{\alpha_n}}\left\{\frac{1}{\alpha_n}\sum_{i\in A_n}\left|h_n^{\vartheta_j}(X^{(i)})\right|^2\right\}^{1/2}
\end{aligned}$$

$$= \sqrt{\frac{6}{\alpha_n}}\sqrt{\mathbb{P}_{\alpha_n}\left((h_n^{\vartheta_j})^2\right)},$$

which, together with Lemma 22.1, implies that

$$\mathbb{E}_X\left[\left\|Z_j\right\|_{\psi_2|X}\right] \leq \sqrt{\frac{6}{\alpha_n}}\mathbb{E}_X\left[\sqrt{\mathbb{P}_{\alpha_n}\left((h_n^{\vartheta_j})^2\right)}\right]$$
$$\leq \sqrt{\frac{6}{\alpha_n}}\left\{\mathbb{E}_X\left[\mathbb{P}_{\alpha_n}\left((h_n^{\vartheta_j})^2\right)\right]\right\}^{1/2} = O\left(\frac{1}{\sqrt{\alpha_n}}\right),$$

uniformly in $\vartheta_j \in \Theta$. From this and Lemma 22.2, $\mathbb{E}\|\mathbb{P}_{\alpha_n} - P\|_{\mathcal{H}_n}$ converges to 0 as $n \to \infty$ and $\delta \to 0$.

22.6.4 Proof of Lemma 22.4

From the definition, we can write for any $\vartheta \in \Theta$

$$\begin{aligned}
P(h_n^\vartheta - h^\vartheta) &= \mathbb{E}\left[\int_0^{\tau_n^\vartheta(X)} e^{-rt} d\xi_{n,t}^\vartheta(X)\right] - \mathbb{E}\left[\int_0^{\tau^\vartheta(X)} e^{-rt} d\xi_t^\vartheta(X)\right] \\
&= \mathbb{E}\left[\int_0^\infty \left\{\mathbf{1}_{\{\tau_n^\vartheta(X)>t\}} - \mathbf{1}_{\{\tau^\vartheta(X)>t\}}\right\} e^{-rt} d\xi_{n,t}^\vartheta(X)\right] \\
&+ \mathbb{E}\left[\int_0^\infty \mathbf{1}_{\{\tau^\vartheta(X)>t\}}\left\{\mathbf{1}_{\{\overline{\hat{X}}_t^{i,n}>\vartheta\}} - \mathbf{1}_{\{\bar{X}_t>\vartheta\}}\right\} e^{-rt} d\overline{\hat{X}}_t^{i,n}\right] \\
&+ \mathbb{E}\left[\int_0^\infty \mathbf{1}_{\{\tau^\vartheta(X)>t\}}\mathbf{1}_{\{\bar{X}_t>\vartheta\}} e^{-rt} d\left\{\overline{\hat{X}}_t^{i,n} - \bar{X}_t\right\}\right] \\
&=: I_1 + I_2 + I_3 \quad \text{(say)},
\end{aligned}$$

where $\overline{\hat{X}}_t^{i,n} = \sup_{0\leq s<t} \hat{X}_s^{i,n}$. For the term I_1, Lemma 22.1 yields

$$|I_1| \leq \left|\mathbb{E}\left[\int_0^T \left\{\mathbf{1}_{\{\tau_n^\vartheta(X)>t\}} - \mathbf{1}_{\{\tau^\vartheta(X)>t\}}\right\} e^{-rt} d\xi_{n,t}^\vartheta\right]\right| + O\left(e^{-rT}\right), \tag{22.7}$$

for any fixed $T > 0$. Denoting $F_{n,\tau}^\vartheta(t) = \mathbb{P}\left(\tau_n^\vartheta(X) \leq t\right)$ and $F_\tau^\vartheta(t) = \mathbb{P}\left(\tau^\vartheta(X) \leq t\right)$, the first term of the right hand side of (22.7) is bounded by

$$\mathbb{E}\left[\int_0^\infty e^{-rt} d\xi_{n,t}^\vartheta\right] \left|F_{n,\tau}^\vartheta(T) - F_\tau^\vartheta(T)\right|.$$

Let $\tilde{\mathbb{D}}_T \subset \tilde{\mathbb{D}}_\infty$ be a space of càdlàg functions on $[0, T]$. We now consider the Skorokhod topology $(\tilde{\mathbb{D}}_T, d_T)$, where d_T is the Skorokhod metric defined by

$$d_T(x, y) = \inf_{\lambda \in \Lambda_T} \left(\max \{ \|x \circ \lambda - y\|_T, \|\lambda - I\|_T \} \right),$$

for any $x = (x_t), y = (y_t) \in \tilde{\mathbb{D}}_T$ (cf., Billingsley [2]). Note that Λ_T is the class of strictly increasing, continuous mappings of $[0, T]$ onto itself, $x \circ \lambda = (x_{\lambda_t})$ for any $\lambda = (\lambda_t) \in \Lambda_T$, I is the identity map on $[0, T]$ and $\|z\|_T = \sup_{0<t\le T} z_t$. On this topology, we define a map $g^\vartheta : (\tilde{\mathbb{D}}_T, d_T) \to (\bar{\mathbb{R}}, |\cdot|)$ by

$$g^\vartheta(x) = \inf_{0<t\le T} \left\{ x_t - \left(\sup_{0<s<t} x_s - \vartheta \right) \vee 0 \right\}.$$

Then, it is easy to see that

$$\begin{aligned}
&|g^\vartheta(x) - g^\vartheta(y)| \\
&= \left| \inf_{0<t\le T} \left\{ x_t + \left(\inf_{0<s<t} (-x_s) + \vartheta \right) \wedge 0 \right\} - \inf_{0<t\le T} \left\{ y_t + \left(\inf_{0<s<t} (-y_s) + \vartheta \right) \wedge 0 \right\} \right| \\
&\le \sup_{0<t\le T} | \inf_{0<s<t} x_s - \inf_{0<s<t} y_s | \\
&\qquad + \sup_{0<t\le T} \left| \left\{ \inf_{0<s<t} (-x_s) + \vartheta \right\} \wedge 0 - \left\{ \inf_{0<s<t} (-y_s) + \vartheta \right\} \wedge 0 \right| \\
&\le \sup_{0<t\le T} |x_t - y_t| + \sup_{0<t\le T} \left| \sup_{0<s<t} x_s - \sup_{0<s<t} y_s \right| \\
&\le 2 \sup_{0<t\le T} |x_t - y_t| \\
&\lesssim d_T(x, y),
\end{aligned}$$

which implies that g is continuous on $(\tilde{\mathbb{D}}_T, d_T)$. Therefore, by using the continuous mapping theorem, Theorem 22.1 implies

$$\inf_{0<t\le T} U_t^\vartheta(\hat{X}^{i_{\mathrm{id}},n}) = g^\vartheta(\hat{X}^{i_{\mathrm{id}},n}) \rightsquigarrow g^\vartheta(X) = \inf_{0<t\le T} U_t^\vartheta(X).$$

From the definition of the weak convergence, it follows that

$$\begin{aligned}
|F_{n,\tau}^\vartheta(T) - F_\tau^\vartheta(T)| &= \left| \mathbb{P}\left(\tau_n^\vartheta \le T\right) - \mathbb{P}\left(\tau^\vartheta \le T\right) \right| \\
&= \left| \mathbb{P}\left(\inf_{0<t\le T} U_t^\vartheta(\hat{X}^{i_{\mathrm{id}},n}) < 0 \right) - \mathbb{P}\left(\inf_{0<t\le T} U_t^\vartheta(X) < 0 \right) \right| \to 0,
\end{aligned}$$

as $n \to \infty$ for any $\vartheta \in \Theta$. Since $\mathbb{E}\left[\int_0^\infty e^{-rt} d\xi_{n,t}^\vartheta\right] = O(1)$ from Lemma 22.1, we have $|I_1| \to 0$ as $n \to \infty$ and $T \to \infty$. In the same way, we have $|I_2| \to 0$ as $n \to \infty$

and $T \to \infty$. For the term I_3, Lemma 22.1 yields

$$|I_3| \le \left| \mathbb{E}\left[\int_0^T \mathbf{1}_{\{\tau^\vartheta(X)>t\}} \mathbf{1}_{\{\bar{X}_t > \vartheta\}} e^{-rt} d\left\{ \overline{\hat{X}}_t^{i,n} - \bar{X}_t \right\} \right] \right| + O(e^{-rT}),$$

for any fixed $T > 0$. Then, there exists a constant $M > 0$ such that

$$\begin{aligned}
&\left| \mathbb{E}\left[\int_0^T \mathbf{1}_{\{\tau^\vartheta(X)>t\}} \mathbf{1}_{\{\bar{X}_t > \vartheta\}} e^{-rt} d\left\{ \overline{\hat{X}}_t^{i,n} - \bar{X}_t \right\} \right] \right| \\
&\le \mathbb{E}\left[\sup_{t\in[0,T]} \frac{\left| \overline{\hat{X}}_t^{i,n} - \bar{X}_t \right|}{\bar{X}_t} \int_0^T e^{-rt} d\bar{X}_t \right] \\
&\le M\mathbb{E}\left[\sup_{t\in[0,T]} \left| \hat{X}_t^{i,n} - X_t \right| \int_0^T e^{-rt} d\bar{X}_t \right] \qquad (22.8) \\
&\le M\mathbb{E}\left[\left(\sup_{t\in[0,T]} \left| \hat{X}_t^{i,n} - X_t \right| \right)^2 \right]^{1/2} \mathbb{E}\left[\left(\int_0^T e^{-rt} d\bar{X}_t \right)^2 \right]^{1/2} \to 0, \qquad (22.9)
\end{aligned}$$

as $n \to \infty$. Note that (22.8) is shown by $\sup_t |\sup_{s<t} x_s - \sup_{s<t} y_s| \le \sup_t |x_s - y_s|$ and $\bar{X}_t \ge u$, and (22.9) is shown by $\mathbb{E}\left[\left(\sup_{t\in[0,T]} \left| \hat{X}_t^{i,n} - X_t \right| \right)^2 \right] = o(1)$ from Theorem 22.1 and $\mathbb{E}\left[\left(\int_0^T e^{-rt} d\bar{X}_t \right)^2 \right] = O(1)$. Hence, $|I_3| \to 0$ as $n \to \infty$ and $T \to \infty$. Therefore, we have

$$|P(h_n^\vartheta - h^\vartheta)| \le |I_1| + |I_2| + |I_3| \to 0,$$

as $n \to \infty$ and $T \to \infty$, uniformly in $\vartheta \in \Theta$.

22.6.5 Proof of Theorem 22.2

Lemmas 22.3 and 22.4 imply that

$$\sup_{\vartheta\in\bar{\Theta}} |\mathbb{P}_{\alpha_n} h_n^\vartheta - Ph^\vartheta| \le \sup_{\vartheta\in\bar{\Theta}} |(\mathbb{P}_{\alpha_n} - P)h_n^\vartheta| + \sup_{\vartheta\in\bar{\Theta}} |P(h_n^\vartheta - h^\vartheta)| \xrightarrow{p} 0,$$

as $n \to \infty$. Combining this and (22.5), we immediately have the conclusion from van der Vaart ([17, Theorem 5.7]).

Acknowledgements The authors would like to thank the Editors for their constructive comments. The first author was partially supported by JSPS KAKENHI Grant Number JP21K03358 and JST CREST JPMJCR14D7, Japan. The second author was supported by JSPS KAKENHI Grant Number JP21K11793.

References

1. BÜHLMANN, H. (1970). *Mathematical Methods in Risk Theory*. Springer-Verlag, Berlin; Heidelberg; New York.
2. BILLINGSLEY, P. (1999). *Convergence of Probability Measures*. 2nd ed. John Wiley & Sons, New York.
3. CROUX, K. AND VERAVERBEKE, N. (1990). Nonparametric estimators for the probability of ruin. *Insurance: Mathematics and Economics* **9** 127–130.
4. DE FINETTI, B. (1957). Su un'impostazione alternativa della teoria collettiva del rischio. *Transactions of the XVth international congress of Actuaries* **2** 433–443.
5. FENG, R. (2011). An operator-based approach to the analysis of ruin-related quantities in jump diffusion risk models. *Insurance: Mathematics and Economics* **48** 304–313.
6. FENG, R. AND SHIMIZU, Y. (2013). On a generalization from ruin to default in a Lévy insurance risk model. *Methodol. Comput. Appl. Probab.* **15** 773–802.
7. GERBER, H.U., SHIU, E.S.W., AND SMITH, N. (2006). Maximizing dividends without bankruptcy. *Astin Bullitain* **36** 5–23.
8. LIN, X.S., WILLMOT, G.E., AND DREKIC, S. (2003). The classical risk model with a constant dividend barrier: analysis of the Gerber–Shiu discounted penalty function. *Insurance: Mathematics and Economics* **33** 551–566.
9. LI, S. (2006). The distribution of the dividend payment in the compound Poisson risk model perturbed by diffusion. *Scandinavian Actuarial Journal* **2006** 73–85.
10. LOEFFEN, R. L. (2008). On optimality of the barrier strategy in De Finetti's dividend problem for spectrally negative Lévy processes. *The Annals of Applied Probability* **18** 1669–1680.
11. KOSOROK, M. (2008). *Introduction to Empirical Processes and Semiparametric Inference*. Springer Science & Business Media.
12. KYPRIANOU, A.E. (2014). *Fluctuations of Lévy Processes with Applications*. 2nd ed. Springer, Heidelberg.
13. SATO, K. (1999). *Lévy Processes and Infinitely Divisible Distributions*. Cambridge University Press, Cambridge.
14. SHIMIZU, Y. (2009). A new aspect of a risk process and its statistical inference. *Insurance: Mathematics and Economics* **44** 70–77.
15. SHIMIZU, Y. AND SHIRAISHI, H. (2022). M-Estimation based on quasi-processes from discrete samples of Lévy process. arXiv:2112.08199.
16. SHIRAISHI, H. AND LU, Z. (2018). Semiparametric estimation in the optimal dividend barrier for the classical risk model. *Scandinavian Actuarial Journal* **9** 845–862.
17. VAN DER VAART, A.W. (1998). *Asymptotic Statistics*. Cambridge University Press, Cambridge.
18. YIN, C., YUEN, K.C. AND SHEN, Y. (2015). Convexity of ruin probability and optimal dividend strategies for a general Lévy process. *The Scientific World Journal* **2015**

Chapter 23
Local Signal Detection for Categorical Time Series

David S. Stoffer

Abstract Frequency domain signal detection for qualitative-valued time series was developed under the assumption of homogeneity using the concept of the spectral envelope. The technique was developed in relation to the optimal scaling of qualitative data. After reviewing some established results, we present a method for fitting a local spectral envelope to heterogeneous sequences based on a minimum description length criterion for choosing the best fitting model based on parsimony. In particular, we focus on the detection of breakpoints in long sequences. Because of the enormous size of the search space, optimization is accomplished using a genetic algorithm to effectively tackle the problem.

Keywords Breakpoint detection · Genetic algorithm · Minimum description length · Nonhomogeneous processes · Qualitative time series · Scaling categorical data · Spectral envelope

23.1 Introduction

Categorical-valued time series are frequently encountered in diverse applications such as economics, medicine, psychology, geophysics, and genomics, to mention a few. The fact that the data are qualitative does not preclude the need to detect signals in the same way that is done with quantitative-valued time series. Here, we explore an approach based on scaling and the spectral envelope introduced in [17].

First we discuss the concept of scaling categorical random variables, and then we use the idea to develop frequency domain analysis of qualitative-valued processes. In doing so, the concept of the spectral envelope and optimal scaling is introduced. While the spectral envelope and the corresponding optimal scaling are a population

D. S. Stoffer (✉)
University of Pittsburgh, Pittsburgh, Pennsylvania, USA
e-mail: stoffer@pitt.edu

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_23

Table 23.1 Per minute infant EEG sleep states (read across and down)

REM	REM	REM	REM	REM	REM	REM	REM	REM	REM	REM	REM
REM	REM	REM	REM	REM	REM	REM	NR2	NR2	NR2	NR3	NR2
NR3	NR4	NR4	NR4	NR4	NR4	NR2	NR4	NR4	NR4	NR4	NR4
NR4	NR4	NR4	NR4	NR4	NR4	NR1	NR1	NR1	NR1	NR1	NR1
NR1	NR1	REM	REM	REM	REM	NR1	REM	REM	REM	REM	REM
REM	REM	REM	REM	REM	REM	REM	REM	REM	REM	REM	REM
REM	NR2	NR2	NR3	NR2	NR4	NR4	NR4	NR4	NR4	NR4	NR4
NR4	NR4	NR4	NR4	NR4	NR4	NR4	NR2	NR1	NR1	NR1	NR1
NR1	NR1	NR1	REM	REM	REM	REM	REM	REM	REM	REM	REM
REM	REM	REM	REM	REM	REM	REM	REM	REM	REM	REM	REM

concept, we also review efficient estimation in the homogeneous case. Pertinent theoretical results are also summarized. Examples of using the methodology on sleep state and DNA sequences, which are typically heterogeneous, are given. The main contribution here is the development of a local procedure using minimum description length (MDL) coupled with optimization via a genetic algorithm (GA).

Our work on the spectral envelope was motivated by collaborations with neurologists who performed sleep studies on neonates with an interest in sleep cycles. For example, Table 23.1 shows the per minute sleep state of an infant taken from a study on the effects of prenatal exposure to alcohol. Details can be found in [16], but briefly, an electroencephalographic (EEG) sleep recording of approximately 2 h was obtained on a full term infant 24–36 h after birth, and the recording was scored by a pediatric neurologist for sleep state. The classification of sleep state is accomplished using a protocol defined by the American Academy of Sleep Medicine. There are two main types of sleep, Non-Rapid Eye Movement (Non-REM), also known as *quiet sleep* and Rapid Eye Movement (REM), also known as *active sleep*. In addition, there are four stages of Non-REM (NR1 – NR4), with NR1 being the "most active" of the four states, and finally awake (AW), which naturally occurs briefly through the night. Each stage of sleep has its own unique characteristics including variations in brain wave patterns, eye movements, and muscle tone. This particular infant was never awake during the study.

Somnologists usually order sleep states by brain activity and movement using the values 1, 2, 3, . . . , however, the idea of ordering sleep states is somewhat tenuous. For example, for a typical normal healthy adult, sleep begins in stage NR1 and progresses into stages NR2, NR3, and NR4. Sleep moves through these stages repeatedly before entering REM sleep. But sleep does not progress through these stages in sequence. Typically, sleep transitions between REM and stage NR2 so that one can move between the states without passing through other sleep states. Moreover, there are no reasons to believe—and no one has suggested—that the distance between, say, NR4 and NR3 is the same as between NR2 and NR1.

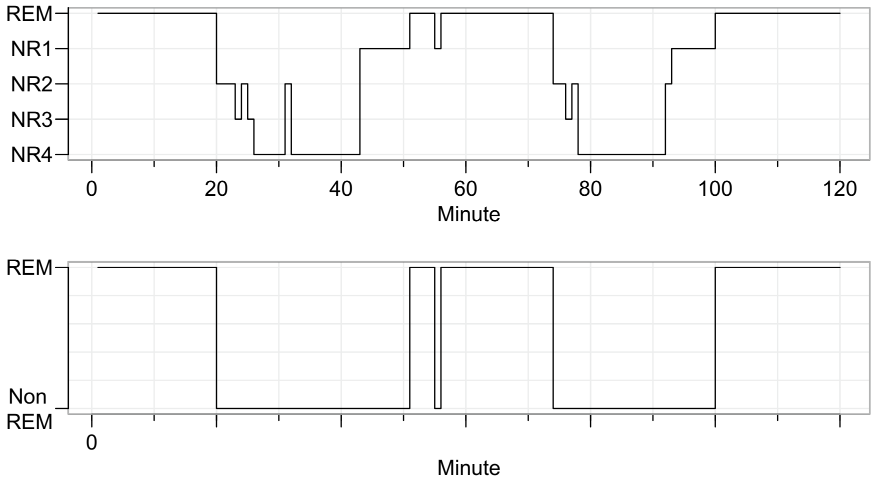


Fig. 23.1 Time plot of the EEG sleep state data in Table 23.1 using the scaling in (23.1) [TOP] and using the scaling in (23.2) [BOTTOM]

It is not too difficult to notice a pattern in the data if one concentrates on REM versus Non-REM sleep states. But, it would be difficult to try to assess patterns in a longer sequence—or if there were more categories—without some graphical aid. One simple method would be to *scale* the data, that is, *assign numerical values to the categories* and then draw a time plot of the scales. As previously mentioned, an obvious scaling that is frequently used by somnologists is

$$\texttt{NR4} = 1, \quad \texttt{NR3} = 2, \quad \texttt{NR2} = 3, \quad \texttt{NR1} = 4, \quad \texttt{REM} = 5, \quad \texttt{AW} = 6, \qquad (23.1)$$

and the top of Fig. 23.1 (often referred to as a *hypnogram*) shows the time plot using this scaling. This scaling, of course, assumes equidistance between the sleep states. Another interesting scaling might be to combine the quiet states and the active states:

$$\texttt{NR4} = \texttt{NR3} = \texttt{NR2} = \texttt{NR1} = 0, \quad \texttt{REM} = \texttt{AW} = 1. \qquad (23.2)$$

The time plot using scalings (23.1) and (23.2) shown in Fig. 23.1 are similar and we notice the general cyclic (in and out of REM sleep) behavior of this infant's sleep pattern. Figure 23.2 shows the estimated spectrum of the sleep data using the scalings in both (23.1) and (23.2). Note that there is a large peak at the frequency corresponding to 1 cycle every 60 min using either scaling. Most of us would feel comfortable with this analysis even though we made arbitrary and ad hoc choices about the particular scaling. It is evident from the data (without any scaling) that if the interest is in infant sleep cycling, this particular sleep study indicates that the infant cycles between REM and Non-REM sleep at a rate of about one cycle per hour.

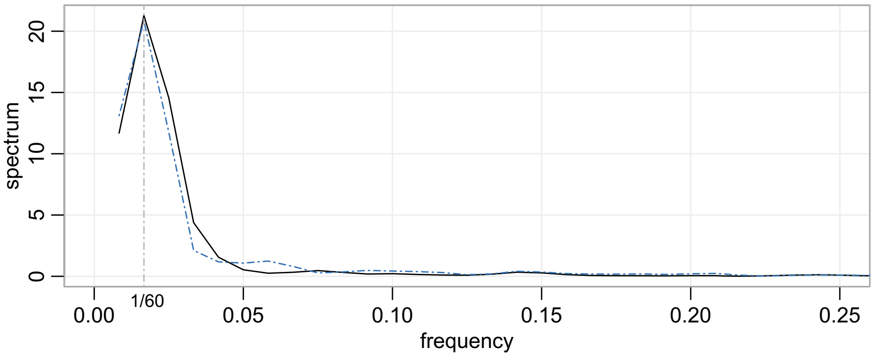


Fig. 23.2 Estimated spectra of the EEG sleep state data in Table 23.1 based on the scaling in (23.1) [solid line] and on the scaling in (23.2) [dashed line]. The peaks in each correspond to a frequency of one cycle every 60 min

The intuition used in the previous example is lost when one considers a long nucleotide DNA sequence. Briefly, a DNA strand can be viewed as a long string of linked nucleotides. Each nucleotide is composed of a nitrogenous base, a five carbon sugar, and a phosphate group where four different bases can be grouped by size, the pyrimidines, thymine (T) and cytosine (C), and the purines, adenine (A), and guanine (G). The nucleotides are linked together by a backbone of alternating sugar and phosphate groups with the $5'$ carbon of one sugar linked to the $3'$ carbon of the next, giving the string direction. DNA molecules occur naturally as a double helix composed of polynucleotide strands with the bases facing inward. The two strands are complementary, so it is sufficient to represent a DNA molecule by a sequence of bases on a single strand. Thus, a strand of DNA can be represented as a sequence of letters, termed base pairs (*bp*), from the finite alphabet {A, C, G, T}. The order of the nucleotides contains the genetic information specific to the organism. Expression of information stored in these molecules is a complex multistage process. One important task is to translate the information stored in the protein-coding sequences (CDS) of the DNA. A common problem in analyzing long DNA sequence data is in identifying CDS that are dispersed throughout the sequence and separated by regions of non-coding (which makes up most of DNA). Table 23.2 shows part of the Epstein–Barr virus (EBV) DNA sequence. The entire sequence consists of over 172,000 bp.

Frequently, geneticists scale according to the purine–pyrimidine alphabet, $\mathtt{A} = \mathtt{G} = 0$ and $\mathtt{C} = \mathtt{T} = 1$, but this is not necessarily of interest for every nucleotide sequence. There are numerous other alphabets of interest, for example, one might focus on the strong–weak hydrogen bonding alphabet $\mathtt{S} = \{\mathtt{C}, \mathtt{G}\} = 0$ and $\mathtt{W} = \{\mathtt{A}, \mathtt{T}\} = 1$; the bp C and G have a strong hydrogen bonding interaction whereas A and T have a weak bonding. In addition, there is no compelling theory that states that any reduction alphabet should be considered. While model calculations as well as experimental data strongly agree that some kind of periodic signal exists in certain—this is DNA sequence, there is a large disagreement about the exact type of periodicity.

Table 23.2 Part of the Epstein–Barr virus (EBV) DNA sequence (read across and down)

AGAATTCGTC	TTGCTCTATT	CACCCTTACT	TTTCTTCTTG	CCCGTTCTCT	TTCTTAGTAT
GAATCCAGTA	TGCCTGCCTG	TAATTGTTGC	GCCCTACCTC	TTTTGGCTGG	CGGCTATTGC
CGCCTCGTGT	TTCACGGCCT	CAGTTAGTAC	CGTTGTGACC	GCCACCGGCT	TGGCCCTCTC
ACTTCTACTC	TTGGCAGCAG	TGGCCAGCTC	ATATGCCGCT	GCACAAAGGA	AACTGCTGAC
ACCGGTGACA	GTGCTTACTG	CGGTTGTCAC	TTGTGAGTAC	ACACGCACCA	TTTACAATGC
ATGATGTTCG	TGAGATTGAT	CTGTCTCTAA	CAGTTCACTT	CCTCTGCTTT	TCTCCTCAGT
CTTTGCAATT	TGCCTAACAT	GGAGGATTGA	GGACCCACCT	TTTAATTCTC	TTCTGTTTGC
ATTGCTGGCC	GCAGCTGGCG	GACTACAAGG	CATTTACGGT	TAGTGTGCCT	CTGTTATGAA
ATGCAGGTTT	GACTTCATAT	GTATGCCTTG	GCATGACGTC	AACTTTACTT	TTATTTCAGT
TCTGGTGATG	CTTGTGCTCC	TGATACTAGC	GTACAGAAGG	AGATGGCGCC	GTTTGACTGT
TTGTGGCGGC	ATCATGTTTT	TGGCATGTGT	ACTTGTCCTC	ATCGTCGACG	CTGTTTTGCA
GCTGAGTCCC	CTCCTTGGAG	CTGTAACTGT	GGTTTCCATG	ACGCTGCTGC	TACTGGCTTT
CGTCCTCTGG	CTCTCTTCGC	CAGGGGGCCT	AGGTACTCTT	GGTGCAGCCC	TTTTAACATT

In addition, there is disagreement about which nucleotide alphabets are involved in the signals, e.g., compare [6, 12].

If we consider the naive approach of arbitrarily assigning numerical values (scales) to the categories and then proceeding with a spectral analysis, the result will depend on the particular assignment of numerical values. The obvious problem of being arbitrary is illustrated as follows: Suppose we observe the sequence ATC TAC ATG . . . , then using the purine–pyrimidine alphabet $\texttt{A} = \texttt{G} = 0$ and $\texttt{C} = \texttt{T} = 1$ yields the numerical sequence 011 101 010 . . ., which is not very interesting. However, if we used the strong–weak bonding alphabet, $\texttt{W} = \{\texttt{A}, \texttt{T}\} = 0$ and $\texttt{S} = \{\texttt{C}, \texttt{G}\} = 1$, then the sequence becomes 001 001 001 . . ., which is very interesting.

In addition, if one considers the sequence {G, A, T, A, G, A, T, A, . . .}, it is repeating every four bp (G, A, T, A, ...). But, the sequence is also repeating every two bp if we consider the sequence in terms of not-A [!A] and A, (!A, A, !A, A, . . .). It should be clear then that one does not want to focus on only one scaling. Instead, the focus should be on finding scalings that bring out all of the interesting features in the data. Rather than choose values arbitrarily, the spectral envelope approach selects scales that help emphasize any periodic feature that exists in a categorical time series of virtually any length in a quick and automated fashion. In addition, the technique can help in determining whether a sequence is merely a random assignment of categories.

23.2 Spectral Envelope

As a general description, the spectral envelope is a frequency based, principal components technique applied to a multivariate time series. In this section, we review the basic concept and its use in the analysis of categorical time series. Technical details can be found in [17]. In addition, various extensions and applications may be found in [8, 18].

In establishing the spectral envelope for categorical time series, we addressed the basic question of how to efficiently discover periodic components in categorical time series. Let $\{X_t;\ t = 0, \pm1, \pm2, \ldots\}$ be a categorical-valued time series with finite state-space $C = \{c_1, c_2, \ldots, c_{k+1}\}$. Assume that X_t is stationary and $p_j = \Pr\{X_t = c_j\} > 0$ for $j = 1, 2, \ldots, k+1$. For $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_{k+1})' \in \mathbb{R}^{k+1}$, denote by $X_t(\boldsymbol{\beta})$ the real-valued stationary time series corresponding to the scaling that assigns the category c_j the numerical value β_j, for $j = 1, 2, \ldots, k+1$. Our goal was to find scalings $\boldsymbol{\beta}$ so that the spectral density, $f(\omega; \boldsymbol{\beta})$, assuming it exists, of the scaled process is in some sense interesting, and to summarize the spectral information by what we called the spectral envelope.

We choose $\boldsymbol{\beta}$ to maximize the power (variance) at each frequency ω, across frequencies $\omega \in (0, 1/2]$, relative to the total power $\sigma^2(\boldsymbol{\beta}) = \text{var}\{X_t(\boldsymbol{\beta})\}$. That is, we choose $\boldsymbol{\beta}(\omega)$, at each ω of interest, so that

$$\lambda(\omega) = \sup_{\boldsymbol{\beta} \not\propto \mathbb{1}} \left\{ \frac{f(\omega; \boldsymbol{\beta})}{\sigma^2(\boldsymbol{\beta})} \right\},$$

where $\mathbb{1}$ is the $(k+1) \times 1$ vector of ones. Note that $\lambda(\omega)$ is not defined if $\boldsymbol{\beta} \propto \mathbb{1}$ because such scalings correspond to assigning each category the same value; in this case, $f(\omega;\ \boldsymbol{\beta}) \equiv 0$ and $\sigma^2(\boldsymbol{\beta}) = 0$. The optimality criterion $\lambda(\omega)$ possesses the desirable property of being invariant under location and scale changes of $\boldsymbol{\beta}$.

As in most scaling problems for categorical data, it was useful to represent the categories in terms of the vectors $\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_{k+1}$, where $\boldsymbol{e}_j$ represents the $(k+1) \times 1$ vector with a one in the jth row, and zeros elsewhere. We then defined a $(k+1)$-dimensional stationary time series $\boldsymbol{Y}_t$ by $\boldsymbol{Y}_t = \boldsymbol{e}_j$ when $X_t = c_j$. The time series $X_t(\boldsymbol{\beta})$ can be obtained from the $\boldsymbol{Y}_t$ time series by the relationship $X_t(\boldsymbol{\beta}) = \boldsymbol{\beta}'\boldsymbol{Y}_t$. Assume that the vector process $\boldsymbol{Y}_t$ has a continuous spectral density matrix denoted by $f_Y(\omega)$. For each ω, $f_Y(\omega)$ is a $(k+1) \times (k+1)$ complex-valued Hermitian matrix. Note that the relationship $X_t(\boldsymbol{\beta}) = \boldsymbol{\beta}'\boldsymbol{Y}_t$ implies that $f_Y(\omega;\ \boldsymbol{\beta}) = \boldsymbol{\beta}' f_Y(\omega)\boldsymbol{\beta} = \boldsymbol{\beta}' f_Y^{re}(\omega)\boldsymbol{\beta}$, where $f_Y^{re}(\omega)$ denotes the real part of $f_Y(\omega)$. The optimality criterion can thus be expressed as

$$\lambda(\omega) = \sup_{\boldsymbol{\beta} \not\propto \mathbb{1}} \left\{ \frac{\boldsymbol{\beta}' f_Y^{re}(\omega)\boldsymbol{\beta}}{\boldsymbol{\beta}' V \boldsymbol{\beta}} \right\}, \tag{23.3}$$

where V is the variance–covariance matrix of $\boldsymbol{Y}_t$. The resulting scaling $\boldsymbol{\beta}(\omega)$ is called the optimal scaling.

In this case, $\boldsymbol{Y}_t$ is a multivariate point process and any particular component of $\boldsymbol{Y}_t$ is the individual point process for the corresponding state (for example, the first component of $\boldsymbol{Y}_t$ indicates whether or not the process is in state c_1 at time t). For any fixed t, $\boldsymbol{Y}_t$ represents a single observation from a simple multinomial sampling scheme. It readily follows that $V = D - \boldsymbol{p}\,\boldsymbol{p}'$, where $\boldsymbol{p} = (p_1, \ldots, p_{k+1})'$, and D

is the diagonal matrix $D = \text{diag}\{p_1, \ldots, p_{k+1}\}$. Since, by assumption, $p_j > 0$ for $j = 1, 2, \ldots, k+1$, it follows that $\text{rank}(V) = k$ with the null space of V being spanned by $\mathbb{1}$. For any $(k+1) \times k$ full rank matrix Q whose columns are linearly independent of $\mathbb{1}$, $Q'VQ$ is a $k \times k$ positive definite symmetric matrix.

With the matrix Q as previously defined, define $\lambda(\omega)$ to be the largest eigenvalue of the determinantal equation

$$|Q' f_Y^{re}(\omega) Q - \lambda Q' V Q| = 0,$$

and let $\boldsymbol{b}(\omega) \in \mathbb{R}^k$ be any corresponding eigenvector, that is,

$$Q' f_Y^{re}(\omega) Q \boldsymbol{b}(\omega) = \lambda(\omega) Q' V Q \boldsymbol{b}(\omega).$$

The eigenvalue $\lambda(\omega) \geq 0$ does not depend on the choice of Q. Although the eigenvector $\boldsymbol{b}(\omega)$ depends on the particular choice of Q, the equivalence class of scalings associated with $\boldsymbol{\beta}(\omega) = Q\boldsymbol{b}(\omega)$ does not depend on Q. A convenient choice of Q is $Q = [\,\mathrm{I_k} \mid \mathbf{0}_\mathrm{k}\,]'$, where $\mathrm{I_k}$ is the $k \times k$ identity matrix and $\mathbf{0}_k$ is the $k \times 1$ vector of zeros. For this choice, $Q' f_Y^{re}(\omega) Q$ and $Q'VQ$ are the upper $k \times k$ blocks of $f_Y^{re}(\omega)$ and V, respectively. This choice corresponds to setting the last component of $\boldsymbol{\beta}(\omega)$ to zero.

The value $\lambda(\omega)$ itself has a useful interpretation; specifically, $\lambda(\omega)d\omega$ represents the largest proportion of the total power that can be attributed to the frequencies $\omega d\omega$ for any particular scaled process $X_t(\boldsymbol{\beta})$, with the maximum being achieved by the scaling $\boldsymbol{\beta}(\omega)$. Because of its central role, $\lambda(\omega)$ was defined to be the *spectral envelope* of a stationary categorical time series.

The name spectral envelope is appropriate since $\lambda(\omega)$ envelopes the standardized spectrum of any scaled process. That is, given any $\boldsymbol{\beta}$ normalized so that $X_t(\boldsymbol{\beta})$ has total power one, $f(\omega; \boldsymbol{\beta}) \leq \lambda(\omega)$ with equality if and only if $\boldsymbol{\beta}$ is proportional to $\boldsymbol{\beta}(\omega)$. That is, *for any scaling of the categories, the standardized spectral density of a scaled sequence is no bigger than the spectral envelope*, with equality only when the numerical assignment is proportional to the optimal scaling, $\boldsymbol{\beta}(\omega)$. The importance of this fact is demonstrated in Fig. 23.3.

Although the law of the process $X_t(\boldsymbol{\beta})$ for any one-to-one scaling $\boldsymbol{\beta}$ completely determines the law of the categorical process X_t, information is lost when one restricts attention to the spectrum of $X_t(\boldsymbol{\beta})$. Less information is lost when one considers the spectrum of $\boldsymbol{Y}_t$. Dealing directly with the spectral density $f_Y(\omega)$ itself is somewhat cumbersome since it is a function into the set of complex Hermitian matrices. Alternatively, one can view the spectral envelope as an easily understood, parsimonious tool for exploring the periodic nature of a categorical time series with a minimal loss of information.

The constraint that $\boldsymbol{\beta}(\omega)$ is real valued leads to restricting attention to the real part of the spectrum as seen in (23.3). If we allow complex-valued scalings, then we would concentrate on the latent roots and vectors of the complex-valued spectral matrix function, $f_Y(\omega)$. In this case, the problem is related to principal component analysis or canonical analysis of time series in the spectral domain as discussed in

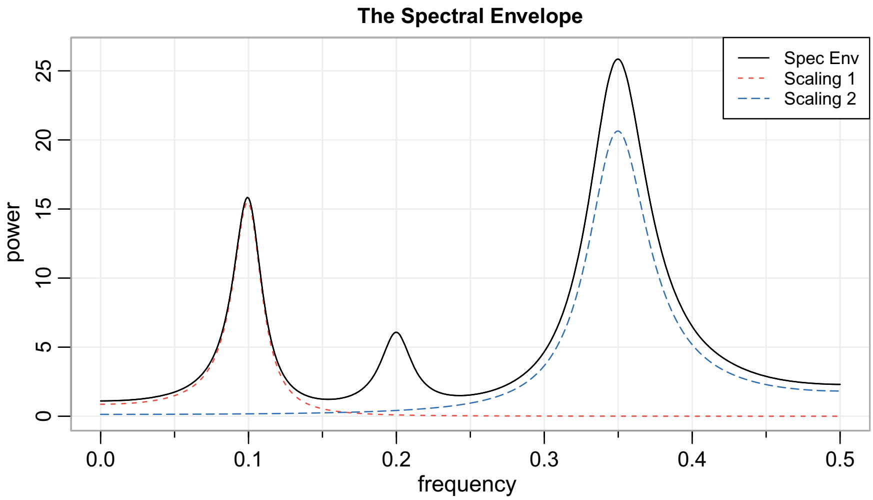


Fig. 23.3 Demonstration of the spectral envelope: The short dashed line indicates a spectral density corresponding to some scaling. The long dashed line indicates a spectral density corresponding to a different scaling. The solid line is the spectral envelope, which can be thought of as throwing a blanket over all possible spectral densities corresponding to all possible scalings of the sequence. Because the spectral density of the first scaling attains the value of the spectral envelope at frequency 0.1, the corresponding scaling is optimal at that frequency. The spectral density of the second scaling is near the spectral envelope at frequency 0.35, hence the corresponding scaling is near-optimal, but can be improved. In addition to finding interesting frequencies (e.g., there is something interesting at frequency 0.2 that neither scaling 1 or 2 discovers), the spectral envelope reveals frequencies for which nothing is interesting (e.g., no matter which scaling is used, there is nothing interesting in this sequence at frequencies below 0.05)

[3]. Although in [3], the problem is formulated in terms of data compression, the problems are similar and the relationship is discussed in more detail in [18]. As a note, we mention that this technique is not restricted to the use of sinusoids. In [16], the use of the Walsh basis of square-wave functions that take only the values ± 1 only is described.

If we observe a finite realization of the stationary categorical time series X_t or, equivalently, the multinomial point process $\boldsymbol{Y}_t$, for $t = 1, \ldots, n$, the theory for estimating the spectral density of a multivariate, real-valued time series is well established and can be applied to estimating $f_Y(\omega)$, the spectral density matrix of $\boldsymbol{Y}_t$. Given an estimate $\widehat{f}_Y(\omega)$ of $f_Y(\omega)$, estimates $\widehat{\lambda}(\omega)$ and $\widehat{\boldsymbol{\beta}}(\omega)$ of the spectral envelope, $\lambda(\omega)$, and the corresponding scalings, $\boldsymbol{\beta}(\omega)$, can then be obtained. Estimation is discussed briefly in the next section.

23.2.1 Estimation

In view of the dimension reduction mentioned in the previous section, the easiest way to estimate the spectral envelope is to fix the scale of the last state at 0, and then select the indicator vectors to be k-dimensional (which we will assume henceforth). That is, to estimate the spectral envelope and the optimal scalings given a stationary categorical sequence, $\{X_t;\ t = 1, \ldots, n\}$, with state-space $C = \{c_1, \ldots, c_{k+1}\}$, perform the following tasks.

- Form $k \times 1$ vectors $\{\boldsymbol{Y}_t,\ t = 1, \ldots, n\}$ as follows:

$$\begin{aligned} \boldsymbol{Y}_t &= \boldsymbol{e}_j \qquad \text{if} \quad X_t = c_j, \quad j = 1, \ldots, k; \\ \boldsymbol{Y}_t &= \boldsymbol{0} \qquad \text{if} \quad X_t = c_{k+1}, \end{aligned}$$

  where now $\boldsymbol{e}_j$ is a $k \times 1$ vector with a 1 in the jth position as zeros elsewhere, and $\boldsymbol{0}$ is the $k \times 1$ vector of zeros.
- Calculate the (fast) Fourier transform of the data,

$$\boldsymbol{d}(\omega_j) = n^{-1/2} \sum_{t=1}^{n} \boldsymbol{Y}_t \exp(-2\pi i t j/n).$$

  Note that $\boldsymbol{d}(\omega_j)$ is a $k \times 1$ complex-valued vector. Calculate the periodogram, $I_n(\omega_j) = \boldsymbol{d}(\omega_j)\boldsymbol{d}^*(\omega_j)$, for $j = 1, \ldots, \lfloor n/2 \rfloor$, where * denotes conjugate transpose.
- Smooth the periodogram as preferred to obtain $\widehat{f}_Y(\omega)$, a consistent estimator, and retain the real part of the spectral matrix estimate. Time series texts such as [14] that cover the spectral domain have an extensive discussion on consistent estimation of a spectral density. Most spectral density estimators can be written in the form

$$\widehat{f}_Y(\omega) = \int_{-1/2}^{1/2} K_n(\omega - \lambda) I_n(\lambda) d\lambda,$$

  where $K_n(\omega)$ is the spectral window, $K_n(\omega) = \frac{1}{b_n} K(\frac{\omega}{b_n})$ for the chosen kernel K. The integral is typically approximated by a sum, and the amount of smoothing is controlled by the bandwidth, b_n, of the window.
- Calculate the $k \times k$ covariance matrix of the data, $S = n^{-1} \sum_{t=1}^{n} (\boldsymbol{Y}_t - \overline{\boldsymbol{Y}})(\boldsymbol{Y}_t - \overline{\boldsymbol{Y}})'$, where $\overline{\boldsymbol{Y}}$ is the sample mean of the data.
- For each $\omega_j = j/n$, for $j = 1, \ldots, \lfloor n/2 \rfloor$, determine the largest eigenvalue and the corresponding eigenvector of the matrix $2n^{-1} S^{-1/2} \widehat{f}_Y^{re}(\omega_j) S^{-1/2}$. Note that $S^{-1/2}$ is the inverse of the unique square root matrix of S.
- The sample spectral envelope $\widehat{\lambda}(\omega_j)$ is the eigenvalue obtained in the previous step. If $\boldsymbol{b}(\omega_j)$ denotes the eigenvector obtained in the previous step, the optimal sample scaling is $\widehat{\boldsymbol{\beta}}(\omega_j) = S^{-1/2}\boldsymbol{b}(\omega_j)$; this will result in k-values, the $(k+1)$-st value being held fixed at zero.

Any standard programming language can be used to do the calculations. The R package `astsa` [15], which supports the text [14], includes a script that calculates the spectral envelope as well as additional scripts to handle various types of data files.

Under the conditions for which $\widehat{f}_Y(\omega)$ has an asymptotic distribution (e.g., see [3]), if $\lambda(\omega)$ is a distinct root (which implies that $\lambda(\omega) > 0$), then, independently, for any collection of Fourier frequencies $\{\omega_i;\ i = 1, \ldots, M\}$, M fixed, and for large n and $\nu_n^2 \sim n\, b_n$ (as $n, \nu_n \to \infty$ but $b_n \to 0$),

$$\nu_n \frac{\widehat{\lambda}(\omega_i) - \lambda(\omega_i)}{\lambda(\omega_i)} \sim \mathrm{AN}(0, 1). \tag{23.4}$$

For example, when estimation is accomplished by a symmetric moving average of the periodogram,

$$\widehat{f}_Y(\omega) = \sum_{q=-r_n}^{r_n} h_q I_n(\omega_{j+q}), \tag{23.5}$$

where $\{\omega_{j+q};\ q = 0, \pm 1, \ldots, \pm r_n\}$ is a band of frequencies and ω_j is the fundamental frequency closest to ω, and such that the weights satisfy $h_q = h_{-q} \geq 0$ and $\sum_{q=-r_n}^{r_n} h_q = 1$, then

$$\nu_n^{-2} = \sum_{q=-r_n}^{r_n} h_q^2.$$

If a simple average is used, $h_q = 1/(2r_n + 1)$, then $\nu_n^2 = (2r_n + 1)$ and the bandwidth is $b_n = \nu_n^2/n$. Based on these results, asymptotic normal confidence intervals and tests for $\lambda(\omega)$ can be readily constructed.

Significance thresholds for consistent spectral envelope estimates can easily be computed using the following approximations. Using a first-order Taylor expansion we have

$$\log \widehat{\lambda}(\omega) \approx \log \lambda(\omega) + \frac{\widehat{\lambda}(\omega) - \lambda(\omega)}{\lambda(\omega)},$$

so that ($n, \nu_n \to \infty, b_n \to 0$)

$$\nu_n[\log \widehat{\lambda}(\omega) - \log \lambda(\omega)] \sim \mathrm{AN}(0, 1). \tag{23.6}$$

It also follows that $\mathrm{E}[\log \widehat{\lambda}(\omega)] \approx \log \lambda(\omega)$ and $\mathrm{var}[\log \widehat{\lambda}(\omega)] \approx \nu_n^{-2}$. If there is no signal present in a sequence of length n, we expect $\lambda(j/n) \approx 2/n$ for $1 < j < n/2$, and hence approximately $(1 - \alpha) \times 100\%$ of the time, $\log \widehat{\lambda}(\omega)$ will be less than $\log(2/n) + (z_\alpha/\nu_n)$ where z_α is the $(1 - \alpha)$ upper tail cutoff of the standard normal distribution. Although this method is a bit crude, from our experience, thresholding at very small α-levels (say, $\alpha = 10^{-4}$ to 10^{-6}, depending on the size of n) works well.

Finally, we mention that inference for estimators of the scaling vectors $\boldsymbol{\beta}(\omega)$ is discussed in [17, Theorem 3.3]. If $\widehat{f}_Y(\omega)$ is a consistent spectral estimator and if for each $i = 1, \ldots, M$ the largest root of $f_Y^{re}(\omega_i)$ is distinct, then

$$\nu_n[\widehat{\boldsymbol{\beta}}(\omega_i) - \boldsymbol{\beta}(\omega_i)] \sim \text{AN}(\mathbf{0}, \Sigma_i),$$

independently for $i = 1, \ldots, M$. The asymptotic covariance structure is given by $\Sigma_i = V^{-1/2}\Omega_i V^{-1/2}$, where

$$\Omega_i = \{\lambda(\omega_i)H(\omega_i)^+ f_Y^{re}(\omega_i)H(\omega_i)^+ - \boldsymbol{a}(\omega_i)\boldsymbol{a}(\omega_i)'\}/2,$$

with $H(\omega_i) = f_Y^{re}(\omega_i) - \lambda(\omega_i)\text{I}_{k-1}$, and $\boldsymbol{a}(\omega_i) = H(\omega_i)^+ f_Y^{im}(\omega_i)V^{1/2}\boldsymbol{b}(\omega_i)$, where $H(\omega_i)^+$ refers to the Moore–Penrose inverse of $H(\omega_i)$. The vector $\boldsymbol{b}(\omega)$ is a normalized version of $\boldsymbol{\beta}(\omega)$ chosen so that $\boldsymbol{b}(\omega)'V^{-1}\boldsymbol{b}(\omega) = 1$ and the first non-zero element of $V^{1/2}\boldsymbol{b}(\omega_i)$ is positive. Notice that the distribution of the scaling vectors depends on the imaginary part of the spectral matrix, however, some simplified approximations are given in [17].

23.2.2 An Example

As a simple example of the kind of analysis that can be accomplished, we consider the gene BNRF1 (3954 bp long) from EBV. Since we are considering the nucleotide

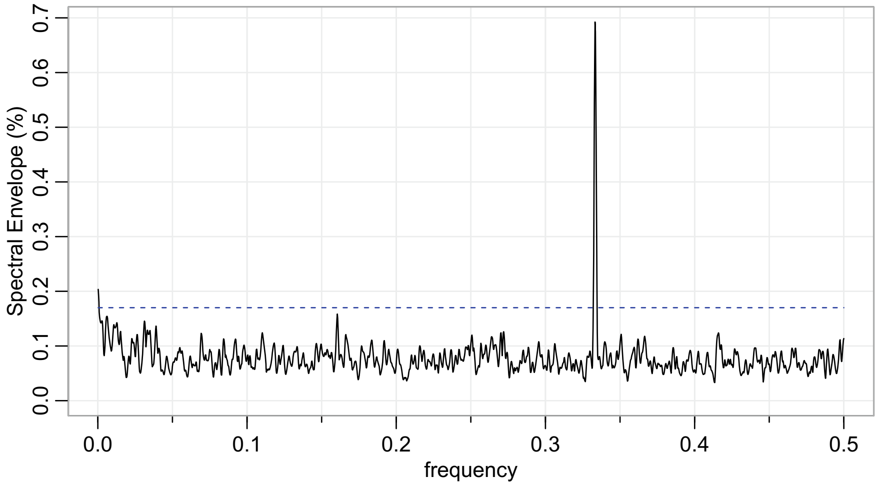


Fig. 23.4 Sample spectral envelope of the gene BNRF1 (3954 bp long) from EBV. There is a large peak in the spectrum at the period of 3 bp

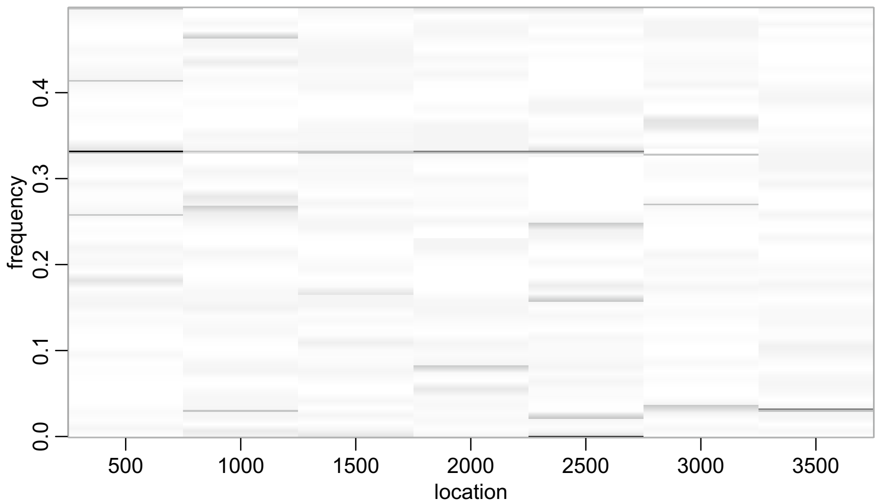


Fig. 23.5 Dynamic spectral envelope estimates for the BNRF1 gene of the EBV based on blocks of 500 bp. The horizontal axis indicates the location of the 500 bp blocks used to calculate the spectral envelope. Darker regions indicate larger values of the spectral envelope

sequence consisting of four bp, we use the following indicator vectors to represent the data:

$$\boldsymbol{Y}_t = (1, 0, 0)' \text{ if } X_t = \texttt{A}; \qquad \boldsymbol{Y}_t = (0, 1, 0)' \text{ if } X_t = \texttt{C};$$
$$\boldsymbol{Y}_t = (0, 0, 1)' \text{ if } X_t = \texttt{G}; \qquad \boldsymbol{Y}_t = (0, 0, 0)' \text{ if } X_t = \texttt{T},$$

so that the scale for the thymine nucleotide, T, is set to zero. Figure 23.4 shows the spectral envelope estimate of the entire coding sequence. The figure also shows a strong signal at frequency 1/3; the corresponding optimal scaling is $\texttt{A} = 0.27$, $\texttt{C} = 0.56$, $\texttt{G} = 0.79$, $\texttt{T} = 0.0$, which indicates the signal is not in terms of any alphabet that collapses the nucleotides such as the purine–pyrimidine (0-1) alphabet, which biotechnologists tend to use, would lead to wrong conclusions.

Figure 23.5 shows a dynamic spectral envelope with a block size of 500. Evidently, even within small segments of the gene, it is not homogeneous. There is, however, a basic cyclic pattern that exists through most of the gene as evidenced by the peak at $\omega = 1/3$ except at the end of the gene. In the next section, we will develop a less ad hoc method.

23.3 Local Analysis

Let a categorical-valued time series $\{X_t;\ t = 1, \ldots, n\}$ consist of an unknown number of segments, m, and let ξ_j be the unknown location of the end of the jth segment, $j = 0, 1, \ldots, m$, with $\xi_0 = 0$ and $\xi_m = n$. Then conditional on m and

$\boldsymbol{\xi} = (\xi_0, \ldots, \xi_m)'$, assume that the process $\{X_t\}$ is piecewise stationary. That is,

$$X_t = \sum_{j=1}^{m} X_{t,j}\, \mathbf{1}_{t,j}, \tag{23.7}$$

where, for $j = 1, \ldots, m$, the indicator processes $\boldsymbol{Y}_{t,j}$ corresponding to $X_{t,j}$ have spectral density $f_j(\omega)$ that may depend on parameters, and $\mathbf{1}_{t,j} = 1$ if $t \in [\xi_{j-1} + 1, \xi_j]$ and 0 otherwise. The piecewise assumption is not very restrictive because slowly varying series may be approximated by a piecewise process, e.g., see [1].

23.3.1 Local Whittle Likelihood

An essential part of the local procedure is the calculation of the local likelihood. In our case, the estimation of the spectral matrix is done nonparametrically via kernel smoothing. Hence, Whittle's form of the likelihood (see [22]) suits our analysis because it depends only on the Fourier transform of the data and a smooth estimate of the spectral matrix function.

Consider a realization $\boldsymbol{x} = \{x_1, \ldots, x_n\}$ from process (23.7), where the breakpoints are known. Let n_j be the number of observations in the jth segment. We assume that the spectra are positive definite, and that each n_j is large enough for the local Whittle likelihood to provide a good approximation. Given a partition of the time series $\boldsymbol{x}$, the jth segment consists of the observations $\boldsymbol{x}_j = \{x_t : \xi_{j-1} + 1 \le t \le \xi_j\}$, $j = 1, \ldots, m$, with underlying spectral densities f_j and Fourier transforms $\boldsymbol{d}_j$ evaluated at frequencies $\omega_{k_j} = k_j/n_j$, for $0 \le k_j \le n_j - 1$. For a given partition $\boldsymbol{\xi}$, the approximate likelihood of the time series is given by

$$L(f_1, \ldots, f_m \mid \boldsymbol{x}, \boldsymbol{\xi}) \approx \prod_{j=1}^{m} (2\pi)^{-n_j/2} \times \prod_{k_j=0}^{n_j-1} \exp\left\{-\frac{1}{2}\left[\log|f_j(\omega_{k_j})| + \boldsymbol{d}_j^*(\omega_{k_j}) f_j^{-1}(\omega_{k_j}) \boldsymbol{d}_j(\omega_{k_j})\right]\right\}, \tag{23.8}$$

where $|\cdot|$ denotes determinant. Conditional on the breakpoints, the local spectral envelope functions can be defined in terms of the local spectral matrix functions in an obvious manner.

23.3.2 Minimum Description Length

Here, we derive an MDL criterion for choosing the best fitting model, where "best" is defined as the model that enables the best compression of the observed series $\boldsymbol{x} = \{x_1, \ldots, x_n\}$.

There are various versions of the MDL principle as put forth by [10, 11] and the version adopted here is a two-part code. Let $\mathcal{F}$ denote the model and $\hat{\mathcal{F}}$ the fitted model. In this case, the first part, denoted by C, represents the *complexity* of the fitted model $\hat{\mathcal{F}}$, and the second part, denoted by $\mathcal{A}$, represents the *accuracy* of the fitted model $\hat{\mathcal{F}}$. The idea of the MDL principle is to find the best pair of C and $\mathcal{A}$ so that via encoding (or compressing), $\boldsymbol{x}$ can be transmitted (or stored) with the least amount of codelength (or memory). To quantify this idea, let $\mathrm{CL}_{\mathcal{F}}(\cdot)$ denote the codelength of an object based on model $\mathcal{F}$. Then we have the decomposition

$$\mathrm{CL}_{\mathcal{F}}(\boldsymbol{x}) = \mathrm{CL}_{\mathcal{F}}(C) + \mathrm{CL}_{\mathcal{F}}(\mathcal{A} \mid C) \tag{23.9}$$

for the data $\boldsymbol{x}$. This approach leads to familiar concepts such as BIC [13] where model accuracy is measured by the negative of the log-likelihood evaluated at the estimated parameters, and complexity is based on the number of parameters in the model and the sample size.

For the complexity term in (23.9), we must consider the various parameters of the model, which includes the number of segments, m; the change points, $\boldsymbol{\xi} = (\xi_1, \ldots, \xi_m)$; and the individual bands in each segment, $B_1, \ldots, B_m$ where $B_j \sim n_j b_{n_j}$ for $j = 1, \ldots, m$ is the number of frequencies included in the smoothing band for segment j as described in (23.5). In this case, we have

$$\mathrm{CL}_{\mathcal{F}}(C) = \mathrm{CL}_{\mathcal{F}}(m) + \mathrm{CL}_{\mathcal{F}}(\xi_1, \ldots, \xi_m \mid m) + \mathrm{CL}_{\mathcal{F}}(B_1, \ldots, B_m \mid m, \boldsymbol{\xi}). \tag{23.10}$$

The values B_j determine the number of distinct bands of frequencies for which $\boldsymbol{f}_j(\omega)$ is estimated. We mention that [5] presented a method based on [10] to choose the bandwidth via stochastic complexity and MDL in the case of stationarity. However, their approach is rarely used because it is overly complex and involves putting a prior on the value of spectral density in each band, which in turn depends on the bandwidth and leads to a somewhat circular argument.

To evaluate (23.10), the codelength for an integer m is $\log_2 m$ bits. For the second term, we note that knowledge of the breakpoints, ξ_j, is equivalent to knowledge of the number of observations in segment j, namely, n_j. Noting that the n_j are bounded by the number of observations, n, we have a bound, $\mathrm{CL}_{\mathcal{F}}(n_j) = \log_2 n$ so that

$$\mathrm{CL}_{\mathcal{F}}(\xi_1, \ldots, \xi_m \mid m) = \mathrm{CL}_{\mathcal{F}}(n_1, \ldots, n_m \mid m) = m \log_2 n.$$

Each bandwidth value will cost about $\log_2 B_j$ bits. In addition, the bandwidth in each segment $j = 1, \ldots, m$ is determined by maximizing the likelihood based on the segment data of n_j observations. For this, we can use a result of Rissanen that states a maximum likelihood estimate of p parameters computed from n_j observations can be effectively encoded with $\frac{1}{2} p \log_2 n_j$ bits, making the third term

$$\mathrm{CL}_{\mathcal{F}}(B_1, \ldots, B_m \mid m, \boldsymbol{\xi}) = \log_2 B_j + \frac{p}{2} \sum_{j=1}^{m} \log_2 n_j,$$

where, in this case, p is the number of parameters in the spectral matrix. For a k-dimensional spectral matrix, there are k real-valued parameters on the diagonal and $k(k-1)/2$ complex-valued parameters on the lower off-diagonals, each with one real and one imaginary part (the upper off-diagonals are the conjugates); hence, $p = k + k(k-1) = k^2$.

For the second term in (23.9), it is shown in [11] that the codelength of the accuracy term, $\mathcal{A}$, is the negative of the $\log_2$ likelihood of the fitted model C. In our case, we use the Whittle likelihood approximation given in (23.8).

Combining the results and working with natural log instead of base 2, we obtain an approximation to the MDL of the model,

$$\mathrm{MDL} = \log m + m \log n + \sum_{j=1}^{m} \log B_j + \frac{k^2}{2} \sum_{j=1}^{m} \log n_j$$
$$+ \sum_{j=1}^{m} \left\{ \frac{n_j}{2} \log(2\pi) + \frac{1}{2} \sum_{k_j=0}^{n_j-1} \left[\log |f_j(\omega_{k_j})| + \boldsymbol{d}_j^*(\omega_{k_j}) f_j^{-1}(\omega_{k_j}) \boldsymbol{d}_j(\omega_{k_j}) \right] \right\}. \tag{23.11}$$

23.3.3 Optimization via Genetic Algorithm

Because the search space is enormous and optimization is a nontrivial task, we use a GA to effectively tackle the problem. A tutorial may be found in [21]. In addition, MATLAB has a toolbox with supporting videos demonstrating GAs that are also good references (see [7]). Our GA is similar to the one specified in [4] who used it to fit local autoregressions to nonstationary univariate time series.

Briefly, GAs are a class of iterative optimization methods that use the principles of evolutionary biology. The algorithm typically begins with some initial randomly chosen population and each generation afterward produces an offspring population using genetic operators. Genetic operators include selection, recombination or crossover, and mutation, which are based on the principle of natural selection to find the best solution while using the principle of diversity to avoid convergence to a local minima.

- *Selection* operators are used to select which offspring survive to the next generation. It is crucial that the fitter individuals are not kicked out of the population, while at the same time diversity should be maintained in the population. Truncation is the simplest selection operator which simply chooses the fittest individuals

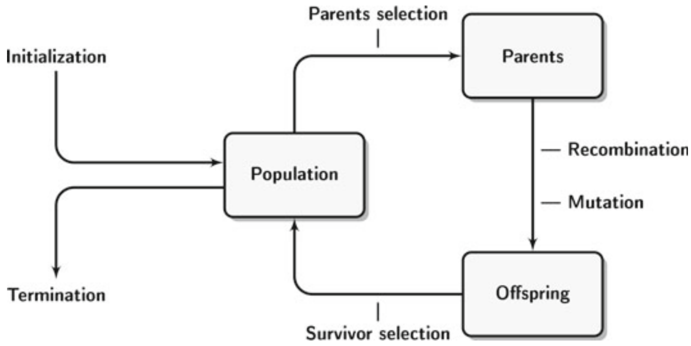


Fig. 23.6 Flow chart of a Genetic Algorithm. The algorithm typically begins with an initial randomly chosen population. Afterward, each generation produces an offspring population using genetic operators. Selection operators are used to select which offspring survive to the next generation. Recombination operators, often referred to as Crossover, are used to mix two or more parents to produce similar, but slightly different offspring. Mutation operators are used to further preserve the diversity of a population to ensure convergence to an optimum

from the parent and offspring population. Tournament selection is another selection operator that randomly sorts the individuals into blocks and chooses the best individual from each block. In Age-Based Selection, there is not a notion of a fitness but it is based on the premise that each individual is allowed in the population for a finite generation where it is allowed to reproduce and then it is kicked out of the population no matter how fit. In Fitness-Based Selection, the children tend to replace the least fit individuals in the population. The selection of the least fit individuals may be done using a variation of any of the selection policies described before, e.g., tournament selection.

- *Recombination* operators, often referred to as *crossover*, are used to mix two or more parents to produce similar, but slightly different offspring. Most crossover operators convert the individual into binary representation to perform the operations. One-point crossover crosses the binary digits at some crossover point of two parents to create two new individuals.
- *Mutation* operators are used to further preserve the diversity of a population to ensure convergence to an optimum. A simple type of mutation involves the addition of a number chosen from a standard normal. Another type of mutation known as flip bit also performs operations on the binary representation of a number where each bit in the representation has some probability of being mutated.

A flow chart of a GA is shown in Fig. 23.6. We give explicit details of the GA we used in the next example.

23.3.4 Another Example

Here, we focus on an analysis of the CDS BNRF1 of EBV, which is roughly 4000 bp long. We selected a subsequence of length $n = 1000$ starting at bp 2500 of the CDS. We choose this section because we know from previous experience (see [18]) that, while most of the CDS contains a signal of period 3, there is a part that appears to be noise. We kept the choice of kernel and corresponding bandwidth simple in that we used the modified Daniell kernel (which is the default in R) with two passes, but allowing the bandwidth to grow. The Daniell kernel corresponds to simple averaging. The modified version simply puts half weights at the ends. For example, if $r = 1$ in (23.5), the modified Daniell weights are ($1/4$, $1/2$, $1/4$). Passing those again yields weights ($1/16$, $4/16$, $6/16$, $4/16$, $1/16$). If one thinks of the first set of weights as a discrete distribution of a random variable with support $\{-1, 0, 1\}$, then the second pass is the distribution of the sum of two independent random variables with that distribution. Thus, in the GA, the value of r in a segment is allowed to grow. In this case, we used an approximation suggested by [19] to obtain the size of the band of the kernel in segment j as $B_j = \nu_{r_j}^2$ in the notation surrounding (23.5). In the case of simple averaging, we have $B_j = 2r_j + 1$.

There are many variations of a GA. For this example, we implemented an Island Model, where instead of running only one search in one giant population, we simultaneously run NI (Number-of-Islands) canonical GAs in NI different sub-populations. The key feature is that a number of individuals are migrated among the islands according to some migration policy. The migration can be implemented in numerous ways (e.g., see [2]) and here we adopted the migration policy that after every M_i generations, the worst M_N chromosomes from the jth island are replaced by the best M_N chromosomes from the $(j-1)$-st island, for $j = 1, \ldots, NI$. For $j = 1$, the best M_N chromosomes are migrated from the NIth island. Here, we used $NI = 40$, $M_i = 5$, $M_N = 2$, and a sub-population size of 40.

Chromosome Representation: The performance of a GA depends on how a possible solution is represented as a chromosome. For our problem, the chromosome carries complete information for any model $\mathcal{F}$, i.e., the number of segments m, the breakpoints ξ_j, and the segment bands B_j. Once these parameters are specified, the Whittle likelihood is uniquely determined. For our problem, a chromosome $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_n)$ is of length n with gene values δ_t defined as $\delta_t = -1$ if there is not a breakpoint at position t, and $\delta_t = B_j$ if $t = \xi_{j-1}$ and the band of the jth piece is B_j. Furthermore, any band size, B_j, is limited to $P_0 = 2r_0 + 1 = 21$ (or 10 fundamental frequencies on either side of the center frequency) and a minimum span on the n_j, ranging from 30 to 70, is specified depending on the size of the band.

Initial Population Generation: The GAs started with an initial population of random chromosomes, and the following strategy was used to generate each of them. First, select a value for $B_1 \in \{0, \ldots, P_0\}$ with equal probabilities and set $\delta_1 = B_1$. Then the next $n_{j_1} - 1$ genes $\delta_2, \ldots, \delta_{n_{j_1}}$ are set to -1 so that the minimum span constraint is imposed for this first piece. The next gene $\delta_{n_{j_1}+1}$ in line will either be initialized as a breakpoint with probability π, or it was assigned -1 with probability

$1 - \pi$. If it is to be initialized as a breakpoint, then we set $\delta_{n_{j_1}} = r_2$, where r_2 is randomly drawn from $\{0, \ldots, P_0\}$. Otherwise, if $\delta_{n_{n_1}}$ is assigned -1, the initialization process will move to the next gene in line and decide if this gene should be a breakpoint gene. This process continues in a similar fashion, and a random chromosome is generated when the process hits the last gene δ_n. In the example, we set $\pi = 10/n$ where n is the length of the sequence.

Crossover and Mutation: Once a set of initial random chromosomes is generated, new chromosomes are generated by either a crossover or a mutation operation. In our implementation, we set the probability for conducting a crossover operation as $1 - \pi$. For the crossover operation, two parent chromosomes are chosen from the current population. The parents are chosen with probabilities inversely proportional to their ranks sorted by their MDL values so that chromosomes having smaller MDL values have a higher chance of being selected. From these two parents, the gene values δ_t of the child chromosome are inherited as follows. First, δ_1 will take on the corresponding value from either the first or second parent with equal probabilities. If the value is -1, then the same gene-inheriting process will be repeated for the next gene in line. Otherwise, the bandwidth is that of the current piece with the minimum span constraint imposed. The same gene-inheriting process will be applied to the next available δ_t.

For mutation, one child is reproduced from one parent. The process starts with $t = 1$ and every δ_t can take on one of the following three values: (i) with probability π_r it will take the corresponding δ_t value from the parent, (ii) with probability π_N it will take the value -1, or (iii) with probability $1 - \pi_r - \pi_N$, it will take a randomly generated bandwidth (subject to the constraints). In our example in the next section, we set $\pi_r = \pi_N = .3$.

Declaration of Convergence: In our example, we use the Island Model in which migration is allowed for every $M_i = 5$ generations. At the end of each migration, the overall best chromosome is noted. If this best chromosome does not change for 10 consecutive migrations, or the total number of migrations exceeds 20, this best chromosome is taken as the solution to this optimization problem.

The GA found two breakpoints at $t = 456$ and 851. Figure 23.7 shows the estimated spectral envelope for each of the three segments as well as for the entire sequence along with significance thresholds. The segment locations appear in the upper right of each plot and the significance threshold used in the figure is 0.0001 for all plots. The first segment shows the typical 3 bp cycle, which was seen in the first example using HVS-BNRF1 and again in the spectral envelope of the entire sequence shown at the bottom right (marked "ALL"). The large values near the zero frequency appear to indicate fractional noise, which is not unusual for DNA sequences, e.g., see [20]. The second and third segments appear to be noise, but the variability in the third segment is slightly larger than the second; in addition, there appears to be fractional noise in the third segment.

The sleep data used in the introduction is included in the R package `astsa` [15] and it is from the third subject included in the data frame `sleep1`. The DNA sequences used throughout this paper may be found online at the National Center

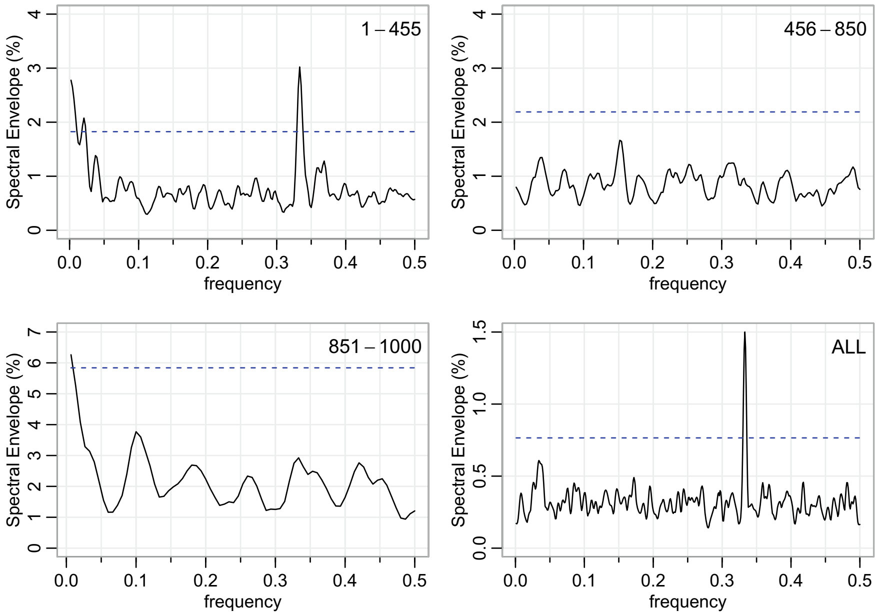


Fig. 23.7 The estimated spectral envelopes for the various segments found by the genetic algorithm in a section of 1000 bp of EBV-BNRF1. The values in the upper right corner are the locations of the segments. The horizontal dashed lines are 0.0001 significance threshold as discussed after (23.6). The graphic on the bottom right is the spectral envelope for the entire 1000 bp and the corresponding threshold is the 0.0001 level

for Biotechnology Information (NCBI). The EBV sequence may be found at [9]. The EBV sequence is also included as a dataset in `astsa` as is the CDS BNRF1 (as `bnrf1ebv`).

References

1. ADAK, S. (1998). Time-dependent spectral analysis of nonstationary time series. *Journal of the American Statistical Association* **93** 1488–1501.
2. ALBA, E. AND TROYA, J. M. (2002). Improving flexibility and efficiency by adding parallelism to genetic algorithms. *Statistics and Computing* **12** 91–114.
3. BRILLINGER, D. (2001). *Time Series: Data Analysis and Theory*. vol 36, Society for Industrial Mathematics.
4. DAVIS, R., LEE, T. AND RODRIGUEZ- YAM, G. (2006). Structural breaks estimation for nonstationary time series models. *Journal of the American Statistical Association* **101** 223–239.
5. HANNAN, E. AND RISSANEN, J. (1988). The width of a spectral window. *Journal of Applied Probability* **25** 301–307.
6. IOSHIKHES, I., BOLSHOY, A., DERENSHTEYN, K., BORODOVSKY, M. AND TRIFONOV, E. N. (1996). Nucleosome DNA sequence pattern revealed by multiple alignment of experimentally mapped sequences. *Journal of Molecular Biology* **262** 129–139.

7. Mathworks (2021). MATLAB Global Optimization Toolbox. https://www.mathworks.com/videos/what-is-a-genetic-algorithm-100904.html
8. McDougall, A., Stoffer, D. and Tyler, D. (1997). Optimal transformations and the spectral envelope for real-valued time series. *Journal of Statistical Planning and Inference* **57** 195–214.
9. NCBI (2021). Epstein-Barr virus (EBV) genome, strain B95-8 - Nucleotide - National Center for Biotechnology Information. https://www.ncbi.nlm.nih.gov/nuccore/V01555
10. Rissanen, J. (1978). Modeling by shortest data description. *Automatica* **14** 465–471.
11. Rissanen, J. (1989). *Stochastic Complexity in Statistical Inquiry*. vol 15, World Scientific.
12. Satchwell, S. C., Drew, H. R. and Travers, A. A. (1986). Sequence periodicities in chicken nucleosome core DNA. *Journal of Molecular Biology* **191** 659–675.
13. Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics* **6** 461–464.
14. Shumway, R. and Stoffer, D. (2017). *Time Series Analysis and Its Applications: With R Examples*. 4th edn, Springer, New York.
15. Stoffer, D. S. (2021). *astsa: Applied Statistical Time Series Analysis*. https://github.com/nickpoison/astsa, r package version 1.14.3
16. Stoffer, D. S., Scher, M. S., Richardson, G. A., Day, N. L. and Coble, P. A. (1988). A Walsh–Fourier analysis of the effects of moderate maternal alcohol consumption on neonatal sleep-state cycling. *Journal of the American Statistical Association* **83** 954–963.
17. Stoffer, D. S., Tyler, D. E. and McDougall, A. J. (1993). Spectral analysis for categorical time series: Scaling and the spectral envelope. *Biometrika* **80** 611–622.
18. Stoffer, D. S., Tyler, D. E. and Wendt, D. A. (2000). The spectral envelope and its applications. *Statistical Science* **15** 224–253.
19. Tukey, J. (1950). The sampling theory of power spectrum estimates. *In: Symposium on Applications of Autocorrelation Analysis to Physical Problems. US Office of Naval Research* 47–67.
20. Voss, R. F. (1993). $1/f$ noise and fractals in DNA-base sequences. *In: Applications of Fractals and Chaos*, *Springer*. 7–20.
21. Whitley, D. (1994). A genetic algorithm tutorial. *Statistics and Computing* **4** 65–85.
22. Whittle, P. (1957). Curve and periodogram smoothing. *Journal of the Royal Statistical Society B* **19** 38–47.

Chapter 24
Topological Inference on Electroencephalography

Yuan Wang

Abstract Statistical inference of electroencephalography (EEG) from diverse clinical groups often requires considerable technicality and computational power. Motivated by topological data analysis, we now take a new analytical angle on EEG signals by characterizing their shape with persistent homology (PH). This paper reviews our recent studies where novel statistical inference procedures are developed for PH features of EEG signals to address clinical questions in brain disorders.

24.1 Introduction

Electroencephalography (EEG) is one of the oldest imaging modalities for recording electrophysiological activities of the brain. Scalp EEG remains popular in clinical evaluation and research studies of brain disorders due to its noninvasive nature, high temporal resolution, and flexibility in setup under various experimental paradigms. EEG activities of patients with brain disorders such as epilepsy and post-stroke aphasia are complex due to the variable neuroanatomical damage/malformation and behavioral symptoms of individual patients. Classical signal processing frameworks typically transform EEG signals from time domain into spectral domain through Fourier and wavelet transforms. The physiological artifacts in standard temporal and spectral features of EEG signals of patients with neuroanatomical impairment can be difficult to separate from underlying brain processes in resting state or in response to stimuli of a task [7, 14]. Moreover, standard group analysis on clinical EEG signals is conducted over temporal or spectral features of signals at a fixed scale, which is useful when we want to infer local signal differences but not when we want to incorporate multiscale features in inference.

A new signal processing approach has been advanced through persistent homology (PH), a major workhorse of topological data analysis (TDA). The topological signal processing approach reveals changes of topological structures across different temporal and spectral scales of signals that are unobservable through the standard

Y. Wang (✉)
University of South Carolina, 915 Greene Street, Columbia, South Carolina, USA
e-mail: wang578@mailbox.sc.edu

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_24

signal processing approach [12, 13, 18, 19]. Current topological signal processing frameworks either extract topological features from a system of signals with the PH algorithm after a sliding window embedding [5, 8, 12, 13, 15], or directly decoding multiscale features in the temporal domain of a signal with the PH algorithm [18, 19]. The former approach has been applied to study financial and video time series, whereas the latter approach has been almost exclusively applied to study EEG signals in brain disorders. In the latter approach, the topological evolution of time segments is summarized through the PH feature persistence landscape (PL) and tests hypotheses on the PLs of signals are based on permuting labels of frequency components in the signals [18] or layers of the PLs [19]. This topological inference framework has been shown in simulation studies to stay robust under certain transformations, such as amplitude translation and frequency scaling of the underlying signal while maintaining sensitivity to topological tearing in the signals. In comparison, statistical tests on standard temporal and spectral features, such as local variance and spectral powers, are empirically sensitive to noisy perturbations and non-topological signal transforms. In addition, the exact inference on the layers in PLs of signals has been empirically shown to drastically improve the computational speed over the standard permutation procedure of randomly permuting frequency components in signals [19].

The studies of [16–22] were motivated by the challenge to perform group analysis on the EEG signals of a diverse group of individuals with epilepsy or post-stroke aphasia. In this paper, we provide a coherent review of the methods proposed in these studies. Section 24.2 covers the basics of PH. Section 24.3 then shows how this helps us extract topological features from signals and perform statistical inference on these features. Section 24.4 highlights the rationale behind simulation studies for evaluating the performance of topological inference methods, while Sect. 24.5 takes stock of interesting applications of these inference methods. Last but not least, directions for future studies are reiterated in Sect. 24.6.

24.2 Preliminary

In simplicial homology, a topological space is represented by combinatorial structures of vertices and edges called *simplicial complexes*. Let $v_1, \ldots, v_p$ be p affinely independent points in the Euclidean space $\mathbb{R}^3$. Then each of the vertices v_i, $i = 1, \ldots, p$, is a *simplex* of the 0th dimension, or 0-simplex. In general, an $(s-1)$-simplex Δ for $s = 1, 2, \ldots,$ is the convex hull of a subset $\{v_{i_1}, \ldots, v_{i_s}\}$ of the p vertices. For instance, an edge joining two vertices is a 1-simplex, a triangle formed by three edges is a 2-simplex, and a tetrahedron formed by four triangles is a 3-simplex (Fig. 24.1). A *face* of Δ is the convex hull of a nonempty subset of $\{v_{i_1}, \ldots, v_{i_s}\}$, e.g., the faces of a tetrahedron are its vertices, edges, and triangles. A *simplicial complex* $\mathcal{K}$ on $\{v_1, \ldots, v_p\}$ is constructed by attaching its simplices combinatorially: a simplex joins $\mathcal{K}$ when all of its faces have joined and the intersection of two simplices in the complex $\mathcal{K}$ must be a face to each of the simplices. A *subcomplex* of $\mathcal{K}$ consists of a subcollection of its simplices attached in the same way.

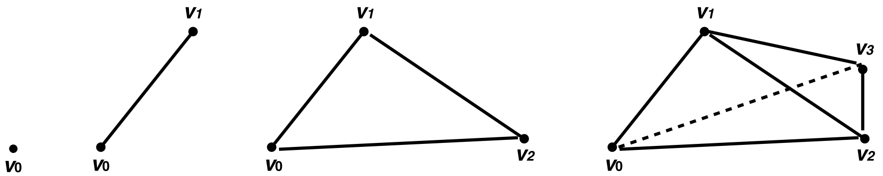


Fig. 24.1 Simplices in $\mathbb{R}^3$; one point v_0 is a 0-simplex, a pair of points v_0 and v_1 joined by an edge is a 1-simplex, a triangle formed by three points v_0, v_1, v_2 joined by three edges is a 2-simplex, and a tetrahedron formed by 4 triangles is a 3-simplex

Suppose we have a simplicial complex $\mathcal{K}$ and a real-valued monotone function $f : \mathcal{K} \to \mathbb{R}$ such that $f(\tau_1) \leq f(\tau_2)$ when τ_1 is a face of τ_2. The monotonicity of f implies that the sublevel set $\mathcal{K} = f^{-1}((-\infty, \lambda])$ is a subcomplex of K for any $\lambda \in \mathbb{R}$. By including faces in $\mathcal{K}$ monotonely with respect to a sequence of λ, we filter through the sequence of subcomplexes of $\mathcal{K}$:

$$\emptyset = \mathcal{K}_0 \subset \mathcal{K}_1 \subset \cdots \subset \mathcal{K}_m = \mathcal{K},$$

where m is bounded by the number of simplices in $\mathcal{K}$. The sequence of subcomplexes is called a *filtration* of $\mathcal{K}$ and the λ is the *filtration value*. The filtration induces a homomorphism chain for each dimension k:

$$0 = H_k(\mathcal{K}_0) \to H_k(\mathcal{K}_1) \to \cdots \to H_k(\mathcal{K}_m) = H_k(\mathcal{K}),$$

where each arrow indicates a homomorphism $H_k^{i,j}$ between the respective k-dimensional homology groups $H_k(\mathcal{K}_i)$ and $H_k(\mathcal{K}_j)$ of $\mathcal{K}_i$ and $\mathcal{K}_j$. The k-dimensional persistent homology group is the image of the homomorphism $H_k^{i,j}$ for $0 \leq i \leq j \leq m$.

The k-dimensional Betti number is defined as

$$\beta_k^{i,j} = rank(H_k^{i,j}),$$

which counts the number of distinct k-dimensional homological features that are born before or at $\mathcal{K}_i$ and die after $\mathcal{K}_j$ [4]. Betti numbers are topological features at fixed scales. The power of PH comes through tracking homological features across scales. If a homological feature of a certain dimension, i.e. cluster or cycle, is born at $\mathcal{K}_i$ and dies at $\mathcal{K}_j$, then $f(\mathcal{K}_j) - f(\mathcal{K}_i)$ is called the *persistence* of the feature. A feature that is born at a finite time and never dies is said to have infinite persistence. Longer persistence indicates a more prominent feature and shorter persistence likely corresponds to noise [2].

24.3 Methods

In this section, we describe PH on univariate signals and statistical inference on PH features of signals.

24.3.1 Persistent Homology on a Signal

The bulk of existing topological signal processing methods extract homological features with respect to a sliding window embedding on a system of one-dimensional signals [5, 8, 12, 13, 15]. The embedding provides a basis for 'inflating' the dimension of the one-dimensional signals so that the standard higher-order homological features can be extracted from the transformed data. The methods used in [16–22] for topological feature extraction is motivated by Morse theory connecting the geometry and topology of an object through its critical points [9]. Although we can only extract the 0th-dimensional homological features (clusters of time segments) this way, they already provide a vast amount of topological information and insight into biomedical problems.

24.3.1.1 Filtration

We have used two types of filtrations for extracting topological features from a univariate signal: sublevel-set filtration, featured in [16, 18, 19, 22] and gradient filtration, featured in [20, 24]. The former filters through a signal vertically, whereas the latter does not restrict the orientation of filtration.

Sublevel-set filtration. Given a smooth signal $h : t \in [0, T] \to \mathbb{R}$, the sublevel set

$$K_\lambda := \{t \in [0, T] | h(t) \leq \lambda\} \tag{24.1}$$

below the horizontal threshold at λ contains time segments corresponding to signal amplitude/electrical potential at or below the threshold. As the threshold moves up and hits $h(t)$ at the unique critical points $(t_{c_1}, h(t_{c_1})), \ldots, (t_{c_M}, h(t_{c_M}))$ satisfying

$$\frac{dh}{dt}(t_{c_i}) = 0 \text{ and } \frac{d^2h}{dt^2}(t_{c_i}) \neq 0, i = 1, \ldots, M, \tag{24.2}$$

time segments emerge and merge in the sublevel set K_λ as the threshold hits local minimums and maximums respectively. This yields the *sublevel-set filtration*. The smoothness condition is easily satisfied with smoothing procedures, which is standard practice for stabilizing the subsequent analysis. We have used the weighted Fourier series in our studies for this purpose (see [18] for a detailed account). But we want

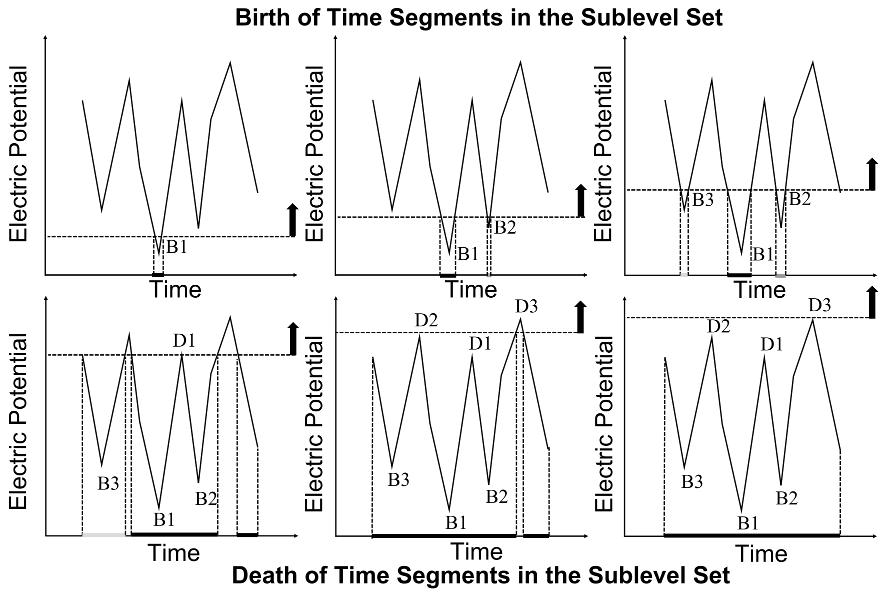


Fig. 24.2 An example of a sublevel-set filtration on a one-dimensional piecewise linear signal. As the horizontal threshold moves up, the time segments born at local minimums B1, B2, and B3 are annihilated at the local maximums D1, D2, and D3, respectively according to the Elder Rule

to point out here that the filtration also works directly on the observed signals taken as piecewise linear functions (Fig. 24.2). Time segments emerge or merge in the sublevel set K_λ as the horizontal threshold moves up and hits $h(t)$ at the unique change points $(t_{c_1}, h(t_{c_1})), \ldots, (t_{c_M}, h(t_{c_M}))$ satisfying

$$\left(h(t_{c_j+1}) - h(t_{c_j})\right)\left(h(t_{c_j}) - h(t_{c_j-1})\right) < 0, \, j = 1, \ldots, M. \tag{24.3}$$

Note that the M points are the discrete analog of the critical points if the function h is smooth. A change point $(t_{c_j}, h(t_{c_j}))$ is analogous to a local minimum if

$$h(t_{c_j+1}) - h(t_{c_j}) > 0 \text{ and } h(t_{c_j}) - h(t_{c_j-1}) < 0,$$

and a local maximum if

$$h(t_{c_j+1}) - h(t_{c_j}) < 0 \text{ and } h(t_{c_j}) - h(t_{c_j-1}) > 0.$$

At a merging juncture, the older of the two time segments lives on and the younger time segment is annihilated—this is known as the *Elder Rule* [4]. In other words, a local minimum where a time segment is born is associated with the birth of the time segment, whereas a local maximum corresponding to two time segments merging is associated with the death of the younger time segment.

Gradient filtration. We have generalized the sublevel-set filtration to filter a signal with a threshold oriented in an arbitrary direction [20]. By treating the smoothed $\hat{g}$ as a two-dimensional curve $\gamma_g : [0, 1] \rightarrow \mathbb{R}^2$, the method quantifies the topological evolution of the arcs on one side of γ_g cut off by a straight line, as expressed by the sublevel set

$$K_\lambda := \{((t, y) \in \gamma_g : y + \kappa t \leq \lambda\}, \quad (24.4)$$

where $y + \kappa t = \lambda$ with a fixed gradient $-\kappa \in \mathbb{R}$ and increasing intercept $\lambda \in \mathbb{R}$ is a straight line in $\mathbb{R}^2$ moving in the direction of increasing λ. Suppose the line touches the curve γ_g tangentially at $(t_{c_1}, g(t_{c_1})), \ldots, (t_{c_M}, g(t_{c_M}))$, i.e. y and t in the line $y + \kappa t = \lambda$ that satisfies

$$y - g(t_{c_i}) = g'(t_{c_i})(t - t_{c_i}), i = 1, \ldots, M,$$

as λ (intercept) of the line increases. The $t_{c_i}, i = 1, \ldots, M$, are critical points where the connectedness of the arcs in the sublevel set K_λ changes as the line moves in the direction of increasing λ (Fig. 24.3 left). Assuming that the line hits at most one $g(t_{c_i})$ at a time, the process yields a sequence of sets of connected arcs $\emptyset = K_0 \subset K_1 \subset \cdots \subset K_M = \gamma_g$. In other words, if $\lambda_1 < \lambda_2 < \cdots < \lambda_M$ are the values of the intercept λ of the line as it hits the $t_{c_i}, i = 1, \ldots, M$ in γ_g, then $K_i = K_{\lambda_i}, i = 1, \ldots, M$, with

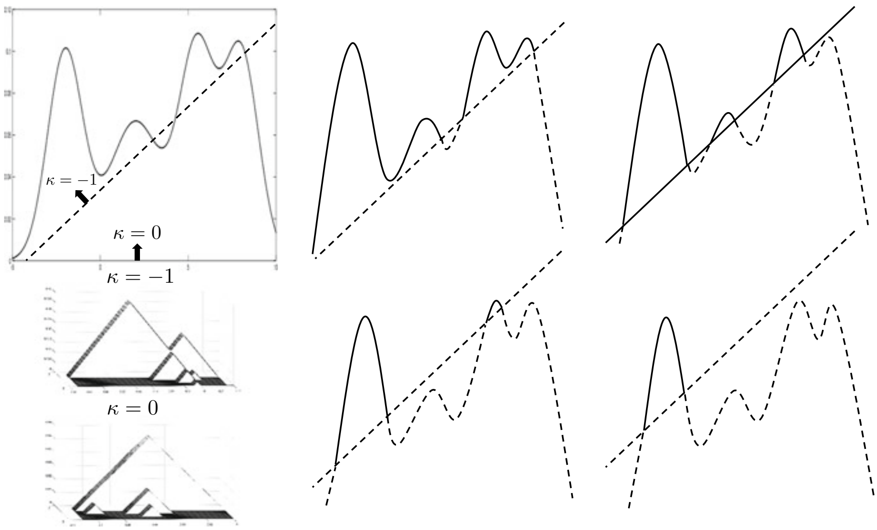


Fig. 24.3 Top and bottom row (right two): Gradient filtration with gradient $-\kappa = 1$. The red arcs form snapshots of the sublevel set in the gradient filtration under different thresholds. As the straight line moves in the direction of increasing λ (intercept of line), arcs emerge (birth) and merge (death) in the sublevel set, where we code the corresponding intercepts of the thresholds with the y-axis as the birth and death times of the arcs. Bottom row (left): The corresponding persistence landscapes (PLs) of the two filtrations with $-\kappa = 1$ and 0

$\lambda_0 = -\infty$. This process forms a *gradient filtration* on γ_g. At a merging junction, a younger arc dies and an elder arc survives–the Elder Rule. In other words, a critical point where an arc is born is associated with the birth of the arc, whereas a critical point corresponding to two arcs merging is associated with the death of the younger arc. Here, we designate the corresponding intercepts λ as the birth and death times of the arcs.

In a gradient filtration, we extract the homological information from γ_g through the connectedness of arcs in its sublevel set K_λ to the right of the threshold $y + \kappa t = \lambda$. This is a generalization of the horizontal sublevel-set filtration tracking through connectedness of time segments corresponding to the signal amplitude/electrical potential of $g(t)$, $t \in [0, T]$, below the horizontal threshold $y = \lambda$ [18, 19, 21]. As λ increases, arcs emerge and merge in the filtration as the threshold $y + \kappa t = \lambda$ hits the points $(t_{c_1}, g(t_{c_1})), \ldots, (t_{c_M}, g(t_{c_M}))$, where the threshold touches the curve γ_g tangentially. This is analogous to time segments emerging and merging in the sublevel-set filtration as the horizontal threshold touches local minimums and maximums of the signal.

24.3.1.2 Persistence Descriptors

If a time segment in a sublevel-set filtration or an arc in a gradient filtration is born at K_i and dies entering K_j, then $pers(\gamma) = \lambda_j - \lambda_i$ is called the *persistence* of the time segment or arc. A time segment or an arc that is born at a finite time and never dies is said to have infinite persistence. A *barcode* encodes the birth and death times of time segments or arcs in a collection of bars $\{(a_i, b_i)\}_{i=1}^{L}$, where the length of a bar corresponds to persistence. A *persistence diagram* (PD) encodes the birth and death times as the x and y coordinates of points above the $y = x$ line, where the vertical drop from a point to the $y = x$ line corresponds to persistence. Barcode and PD do not possess a natural statistical framework and require additional manipulation in complex data analysis. The *persistence landscape* (PL), on the other hand, forms a separable Banach space and has a well-defined statistical framework [1]. A PL can be constructed through a PD or barcode. Given a bar (a_i, b_i) in a barcode $\{(a_i, b_i)\}_{i=1}^{L}$, we define the piecewise linear bump function $r_{(a_i,b_i)} : \mathbb{R} \to \mathbb{R}$ by

$$r_{(a_i,b_i)}(\lambda) = \max(\min(\lambda - a_i, b_i - \lambda), 0), \tag{24.5}$$

where λ denotes the filtration value. The PL $\nu = \{\nu_\ell : 1 \le \ell \le L\}$ of the barcode $\{(a_i, b_i)\}_{i=1}^{L}$ is then defined as

$$\nu_\ell(\lambda) = \begin{cases} \ell\text{-th largest value of } \{r_{(a_i,b_i)}(\lambda)\}_{i=1}^{L} & 1 \le \ell \le L, \\ 0 & \ell > L, \end{cases} \tag{24.6}$$

thus making ν a multi-valued function. Figure 24.4 shows the PL constructed on a barcode obtained from the sublevel-set filtration in Fig. 24.2, where the PL layers

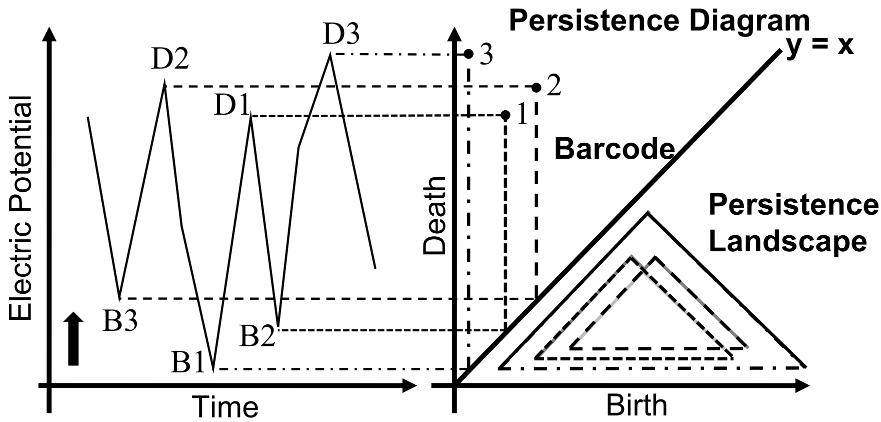


Fig. 24.4 An example of a sublevel-set filtration on a one-dimensional piecewise linear function (left) and the corresponding persistence diagram (black dots), barcode (vertical drops from the black dots to the $y = x$ line), and persistence landscape (red, blue, and green outlines of the intersecting bump functions on the barcode)

are the outlines of the bump functions built on the bars of the barcode. Note that the second and third layers of the PL in Fig. 24.4 trace over the bump functions of two bars as the bump functions intersect. For $1 \leq p \leq \infty$, the *p-landscape distance* between two PLs ν^1 and ν^2 is defined as

$$\Lambda_p(\nu^1, \nu^2) = ||\nu^1 - \nu^2||_p. \tag{24.7}$$

The paper [1] showed that PL is stable with respect to the supremum norm, as follows.

Theorem 24.1 ([1]; ∞-landscape stability theorem) *Let $h_1, h_2 : X \to \mathbb{R}$ be real-valued functions on a topological space X and ν^1, ν^2 be the corresponding persistence landscapes. Then*

$$\Lambda_\infty(\nu^1, \nu^2) \leq ||h_1 - h_2||_\infty,$$

where Λ_∞ is the ∞-landscape distance and

$$||h_1 - h_2||_\infty := \sup_{t \in X} |h_1(t) - h_2(t)|.$$

Suppose we have n trials of EEG signals $f_j, j = 1, \ldots, n$ sampled at $t = t_1, \ldots, t_m \in [0, T]$, from the underlying waveform g:

$$f(t) = g(t) + \varepsilon, t \in [0, T], \tag{24.8}$$

where we assume that the white noise ε follows a Gaussian distribution with mean 0 and variance σ^2. We average across the n trials of EEG signals to obtain an estimate of g:

$$\widehat{g}_n(t_i) := \frac{1}{n}\sum_{j=1}^{n} f_j(t_i),\ i = 1, \ldots, m. \tag{24.9}$$

By Theorem 24.1, the distance between the PLs of g and its estimate $\widehat{g}$ is bounded by the distance between g and $\widehat{g}$:

$$\Lambda_\infty(\bar{\nu}^{\widehat{g}_n}, \bar{\nu}^g) \le ||\widehat{g}_n - g||_\infty. \tag{24.10}$$

We can thus establish an upper bound for $\Lambda_\infty(\bar{\nu}^{\widehat{g}_n}, \bar{\nu}^g)$ (Corollary 24.1) via the supremum norm $||\widehat{g}_n - g||_\infty$ (Theorem 24.2), as follows.

Theorem 24.2 ([22]) *Suppose, for any participant s, a signal $f_j(t_i)$ is sampled at fixed time points t_i, $i = 1, \ldots, m$ for any $j = 1, \ldots, n$, and*

$$\mathbb{E} f_j(t_i) = g(t_i), \forall i = 1, \ldots, m,\ j = 1, \ldots, n.$$

Then

$$\limsup_{n\to\infty} \mathbb{E}||\widehat{g}_n - g||_\infty \le C,$$

for $C > 0$.

Corollary 24.1 ([22]) *For the model (24.8) and estimate $\widehat{g}_n$ defined by (24.9),*

$$\limsup_{n\to\infty} \mathbb{E}\Lambda_\infty(\bar{\nu}^{\widehat{g}_n}, \bar{\nu}^g) \le C,$$

for $C > 0$.

24.3.2 Permutation-Based Topological Inference

We have developed two permutation tests for comparing persistence features in two signals or two groups of signals.

24.3.2.1 Spectral Permutation Test

In [18], we advanced a spectral permutation test based on permuting Fourier coefficients of single-trial signals to resolve an issue of time-based bootstrap breaking topology of characteristic waveforms related to epileptic seizures. The procedure begins by denoising/smoothing EEG signals from two clinical phases, i.e. before vs during seizure, with weighted Fourier series (WFS) derived from a heat diffusion problem. Then we extract PLs from the sublevel-set filtration of the denoised signals. Our goal is to compare the PLs of the denoised signals from the two clinical phases by testing the equality of the PLs ν^1 and ν^2 of the two phases:

$$H_0 : \nu^1 = \nu^2. \tag{24.11}$$

Since we have only one epoch in each clinical phase, we require an inference framework on PLs incorporating a natural resampling approach. Resampling in the time domain by chopping up the signals would break up characteristic waveforms. As we estimate the underlying neural processes in the two phases by truncated WFS, we resample in the frequency domain by randomly permuting the Fourier coefficients in the two WFS respective to their frequency components. Following the resampling step, we reconstruct two WFS in the time domain. We measure the difference between the PLs ν^1 and ν^2 of the WFS by their L_2 distance:

$$L_2(\nu^1, \nu^2) = \left(\int \sum_{k=1}^{N} |\nu^1(k, x) - \nu^2(k, x)|^2 dx \right)^{1/2}, \tag{24.12}$$

where N is the larger of the number of layers of the two PLs. For each permutation, we calculate the $L_2(\nu^{1'}, \nu^{2'})$ between the PLs $\nu^{1'}$ and $\nu^{2'}$ of the two reconstructed WFS. We then measure the statistical significance of $L_2(\nu^1, \nu^2)$ between the PLs by comparing it with the empirical distribution of the $L_2(\nu^{1'}, \nu^{2'})$ from the permutations. Two phases of a signal are said to be *topologically invariant* in the statistical sense if the difference between their PLs is not statistically significant.

This spectral permutation test is obviously limited to single-trial signals and has recently been extended to multi-trial signals in [23]. An interim approach we developed for comparing persistence features representing two groups of multi-trial signals is the exact permutation test [19, 21, 22].

24.3.2.2 Exact Permutation Test of Persistence Features

We adapted an exact inference framework [3] to construct a fast permutation test for comparing persistence and PLs between two groups of multi-trial EEG signals. The idea is to apply a combinatorial procedure in computing the exact p-value of a Komogorov–Smirnov test comparing two monotone increasing functions. It drastically speeds up the permutation test. Two versions of the exact permutation test exist, with the early version based on the raw areas under the layers of the PLs [19, 21], and the later version based on the percentiles of the areas under the layers of the PLs [22]. The early version of the test requires an equal number of layers in the PLs, which is in practice rarely satisfied, and so would require truncation of layers in one of the PLs. But it loses information in the truncated PL. The later version, on the other hand, resolves the issue through percentiles and can also be used for comparing persistence in two groups of multi-trial signals.

Exact inference on persistence. We compare the grand average EEG signals of two groups at each channel through an exact permutation test on their ordered persistence

$$\{pers_1^1, \dots, pers_{L_1}^1\} \text{ and } \{pers_1^2, \dots, pers_{L_2}^2\}.$$

By denoting the ℓ-th percentiles of the two groups of persistence as $\hat{P}_\ell^1$ and $\hat{P}_\ell^2$, $\ell = 1, \dots, L$, with L being a pre-specified integer between 1 and 100. We then test the null hypothesis on the population percentiles

$$H_0 : P_\ell^1 = P_\ell^2, \text{ for all } \ell = 1, \dots, L.$$

The test statistic is

$$D(\psi_1, \psi_2) = \sup_t |\psi_1(t) - \psi_2(t)|, \qquad (24.13)$$

where the monotone step functions $\psi_1(t)$ and $\psi_2(t)$ come from the observed percentiles $\hat{P}_\ell^1$ and $\hat{P}_\ell^2$, $\ell = 1, \dots, L$:

$$\psi_i(t) = \begin{cases} 0 \text{ if } t < \hat{P}_1^i, \\ \ell \text{ if } \hat{P}_\ell^i \le t < \hat{P}_{\ell+1}^i, 1 \le \ell < L, \\ L \text{ if } t \ge \hat{P}_L^i, \end{cases} \qquad (24.14)$$

for $i = 1, 2$. The exact p-value of an observed value d of the test statistic D is given by

$$P(D \ge d) = 1 - \frac{B(L, L)}{\frac{(2L)!}{L!L!}}, \qquad (24.15)$$

where $B(L, L)$ is computed iteratively as

$$B(u, v) = B(u - 1, v) + B(u, v - 1)$$

within the domain $|u - v| < d$, under the boundary condition $B(0, L) = B(L, 0) = 1$.

Exact inference on persistence landscapes. We compare the average PLs of the two groups through the areas

$$\{A_1^1, \dots, A_{L_1}^1\} \text{ and } \{A_1^2, \dots, A_{L_2}^2\}$$

under the reversely ordered layers of the two average PLs. We test the null hypothesis on the ℓ-th population percentiles of these areas

$$H_0 : P_\ell^1 = P_\ell^2, \text{ for all } \ell = 1, \ldots, L,$$

where L is a pre-specified integer between 1 and 100. This way we compare percentiles of the areas under the PL layers [22] instead of the raw areas under the layers [19, 21], which allows comparing the same number of percentiles in two PLs without having to truncate the PL with more layers when the two numbers of layers are not equal. In application, we can use all integer percentiles rather than a select number of percentiles because we want to include all topological information in the inference procedure. The test statistic and p-value are computed as (24.13)–(24.15).

Although the exact permutation approach can be used to compare the persistence features of two groups of signals, it cannot directly permute the labels of the signals, which is a limitation that the new spectral permutation test [23] has overcome.

24.4 Simulations

We have designed the simulation studies in [18, 19, 22] under performance evaluation criteria developed for topological inference. Standard performance metrics for a statistical test are the false positive rate and power. For topological inference, we want to control the rate of *topological false positives* while maintaining decent power; in other words, we require the tests on PLs to stay robust when the underlying signals have similar PH while staying sensitive to dissimilar PH. We know from algebraic topology that the homology of a topological space is preserved under continuous transformation, without tearing or gluing, of the space [6]. Yet, continuity alone does not guarantee the preservation of PH in the sublevel set of a signal. To preserve PH, signal transformations also need to respect the pairing of birth and death times. A theoretical framework for quantifying such transformations under the sublevel-set and gradient filtrations is still in development, so we only simulate baseline transformations such as translation and scaling of amplitude and frequency that are intuitively PH-preserving. On the other hand, it is much easier to construct transformations that destroy the PH of the signal by introducing discontinuities in the signal. We only need to introduce tearing in the signal by multiplying the signal pointwise by a step function with all piecewise time intervals containing at least one but not all critical points of the signal; in other words, we break up the signal at a few time points and consecutively push the signal segments in opposite vertical directions.

We typically simulate up to 1000 sets of signals for the performance evaluation. For each set of signals, a permutation test is applied to compare the original signal(s) and their transformed counterpart(s), yielding a p-value. The false positive rate is calculated as the percentage of p-values that fall under the pre-specified significance level, e.g., 5%, (detecting a false positive) out of the number of simulated sets of signals undergoing PH-preserving transformations. Power is calculated as the percentage of p-values that fall under the significance level (detecting a true positive) out of the number of simulated sets of signals undergoing tearing transformations. Under these settings, we evaluated the performance of the spectral permutation method proposed in [18] that tests the equality between the PLs of two signals. The false positive rates of the proposed tests under amplitude translation and scaling and frequency scaling of a signal stay controlled under the significance level of 5%, while the power of the proposed test when a signal undergoes tearing stays above 90%. Results are consistent with the exact permutation test proposed in [19] on signals simulated under the same settings. We also evaluated the performance of the exact permutation method proposed in [22] that tests for differences in persistence or PLs extracted from two groups of multi-trial signals. We simulated sets of multi-trial signals under the same settings as [18, 19] and the results were largely consistent with the single-trial simulations.

24.5 Applications

Although our permutation-based topological inference approach has been set up for univariate tests, we can still obtain significant insight through these tests by identifying the spatial distribution of neural deficits through mass univariate testing with adjustment for multiple comparisons. In application to single-trial EEG signals [18], we obtained the spatial pattern of insignificant p-values computed by counting the proportion of PLs of reconstructed WFS having L_2 distances exceeding that of the observed PLs before and during seizure and after Bonferroni correction for tests at 8 channels. The spatial pattern indicates that EEG signals in the epileptogenic zone before a seizure attack already has similar topological patterns as those during the seizure, which is consistent with the diagnosis of the left temporal lobe being the epileptogenic zone. The results were also confirmed by the exact permutation test applied to the same dataset in [19]. These permutation tests on persistence features picked up patterns in EEG signals which were overlooked by state-of-the-art methods that previously analyzed this same dataset and showed the evolution of the spectral power and coherence during seizure but not the location of seizure onset [10, 11]. The findings indicate that topological inference may provide a promising metric for seizure localization.

In application to multi-trial EEG signals recorded under an event-related potential (ERP) paradigm [21, 22], we used the exact permutation test to compare persistence features in ERP response of individuals with aphasia and healthy controls with respect to speech onset or in response to two auditory feedback stimuli (upward and downward pitch shifts). The application revealed a spatial pattern of topological difference that highlighted neural deficits in aphasic individuals, with strong topological differences between the aphasia and control groups in the parietal-occipital and occipital regions with respect to speech onset. We used here persistence and PL to summarize the evolution of time segments in the gradient filtration on the ERPs. The spatial pattern of neural deficits was enhanced in the application of the exact permutation test on the gradient filtration in [20], with the parietal-occipital and occipital regions showing significant p-values with respect to speech onset and in response to both auditory feedback stimuli. These applications open up a new avenue for quantifying speech-language deficits through topological studies of concurrent changes in vocal motor behavior and EEG activity.

Note that our summary of the topological evolution in the single-trial EEG signals or ERP response provides a unique account of the relationship between temporal and spectral information across different scales. This enables the persistence and PL tests to detect systematic differences in multiscale, instead of fixed scale, temporal and spectral information of signals. In comparison, spectral analysis of the EEG data in the applications reveals the extensive differences between phases of a seizure attack or between the aphasia and control groups in a diffused pattern of brain regions, instead of pointing to a more focused region of difference like the pattern revealed by the topological analysis.

24.6 Discussion

The topological signal processing and inference approaches reviewed in this paper provide a basis for investigation beyond univariate signals. For instance, we have constructed topological correlation out of PLs extracted across gradient filtration and applied it to study the connectivity of EEG signals recorded in two clinical phases of epileptic seizures [24]. We have also extended the spectral permutation test to multi-trial graph signals in [23]. As pointed out in [22], the exact topological inference framework drastically improves the average computation time over standard permutation test, which will facilitate application to a much larger patient group in the future. The fast computational implementation can be incorporated into statistical analysis of large-scale clinical EEG signals recorded in week-long clinical monitoring settings. Furthermore, the framework can be used for semi-online detection of changes in studies such as induced stroke in rats and cognitive studies on learning.

References

1. Bubenik, P. (2015). Statistical topological data analysis using persistence landscapes. *Journal of Machine Learning Research* **16** 77–102.
2. Carlsson, G. (2009). Topology and data. *Bulletin of the American Mathematical Society* **46** 255–308.
3. Chung, M., Luo, Z., Alexander, A., Davidson, R. and Goldsmith, H. (2018). Exact combinatorial inference for brain images. *Proceedings of the International Conference on Medical Image Computing and Computer Assisted Intervention (MICCAI)* **11070** 629–637.
4. Edelsbrunner, H. and Harer, J. (2010). *Computational Topology*. American Mathematical Society.
5. Gidea, M. and Katz, Y. (2018). Topological data analysis of financial time series. *Physica A: Statistical Mechanics and its Applications* **491** 820–834.
6. Hatcher, A. (2001). *Algebraic Topology*. Cambridge University Press.

7. Leamy, D., Kocijan, J., Domijan, K., Duffin, J., Roche, R., Commins, S., Collins, R. and Ward, T. (2014). An exploration of EEG features during recovery following stroke—implications for BCI-mediated neurorehabilitation therapy. *Journal of NeuroEngineering and Rehabilitation* **11**.
8. Majumdar, S. and Laha, A. (2020). Clustering and classification of time series using topological data analysis with applications to finance. *Expert Systems with Applications* **August** 113868.
9. Milnor, J. (1963). *Morse Theory*. Princeton University Press.
10. Ombao, H., Raz, J., von Sachs, R. and Malow, B. A. (2001). Automatic statistical analysis of bivariate nonstationary time series. *Journal of the American Statistical Association* **96** 543–560.
11. Ombao, H., von Sachs, R. and Guo, W. (2005). Slex analysis of multivariate nonstationary time series. *Journal of the American Statistical Association* **100** 519–531.
12. Perea, J. (2016). Persistent homology of toroidal sliding window embeddings. *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* 6435–6439.
13. Perea, J. and Harer, J. (2015). Sliding windows and persistence: An application of topological methods to signal analysis. *Foundations of Computational Mathematics* **5** 799–838.
14. Sisodiya, S. (2000). Surgery for malformations of cortical development causing epilepsy. *Brain* **123** 1075–1091.
15. Tralie, C. and Perea, J. (2018). (Quasi) periodicity quantification in video data, using topology. *SIAM Journal on Imaging Sciences* **11** 1049–1077.
16. Wang, Y., Ombao, H. and Chung, M. (2015). Topological seizure origin detection in electroencephalographic signals. *Proceedings of the International Symposium on Biomedical Imaging (ISBI)* 351–354.
17. Wang, Y., Chung, M., Dentico, D., Lutz, A. and Davidson, R. (2017). Topological network analysis of electroencephalographic power maps. *Proceedings of the International Workshop on Connectomics in Neuroimaging (CNI)* **10511** 134–142.
18. Wang, Y., Ombao, H. and Chung, M. (2018). Topological data analysis of single-trial electroencephalographic signals. *Annals of Applied Statistics* **12** 1506–1534.
19. Wang, Y., Ombao, H. and Chung, M. (2019). Statistical persistent homology of brain signals. 1125–1129.
20. Wang, Y., Behroozmand, R., Phillip Johnson, L., Bonilha, L. and Fridriksson, J. (2020a). Topology signal processing in neuroimaging studies. *Workshop Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI).*
21. Wang, Y., Behroozmand, R., Phillip Johnson, L. and Fridriksson, J. (2020b). Topology highlights neural deficits of post-stroke aphasia patients. *Proceedings of the IEEE International Symposium on Biomedical Imaging (ISBI)* 754–757.
22. Wang, Y., Behroozmand, R., Phillip Johnson, L., Bonilha, L. and Fridriksson, J. (2021). Topological signal processing and inference of event-related potential response. *Journal of Neuroscience Methods* **363** 109324.
23. Wang, Y., Chung, M. and Fridriksson, J. (2022). Spectral permutation test on persistence diagrams. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* 1461–1465.
24. Yin, J. and Wang, Y. (2022). Topological correlation of brain signals. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)* 1411–1415.

Chapter 25
UMVU Estimation for Time Series

Xiaofei Xu, Masanobu Taniguchi, and Naoya Murata

Abstract This paper introduces the sufficiency, completeness, and uniformly minimum variance unbiased estimation for Gaussian circular ARMA processes. We propose a uniformly most powerful (UMP) test by monotone likelihood ratio for the coefficient parameter of the Gaussian circular AR(1) models. The numerical study shows good performance of the UMP test with composite null and composite alternative hypotheses for the Gaussian circular AR(1) models.

25.1 Introduction

Autoregressive moving average (ARMA) model is one of the most important time series models which account for the serial dependence among the historical, present, and future values of time series. Dating back to the 1940s, a foundation of inference for AR-type models has been developed by [7] where the maximum-likelihood estimators (MLEs) are proved to be consistent. A comprehensive introduction to estimation and hypothesis testing theories of time series models was given by [2]. See also [3, 4] for more discussion on various univariate and multivariate ARMA models and their applications in various fields.

Numerous estimators and hypothesis testings for time series models have been studied in the literature. For independent observations, we refer the readers to [6, 8]

X. Xu
Wuhan University, 299 Ba Yi Road, Wuchang District, Wuhan, Hubei, P.R. China
e-mail: xiaofeix@whu.edu.cn

M. Taniguchi (✉)
Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
e-mail: taniguchi@waseda.jp

N. Murata
Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
e-mail: n38murata25@asagi.waseda.jp

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_25

for a good overview of methods and theories of hypothesis testing. Among them, the uniformly most powerful (UMP) test, which maximizes the power of all alternatives in the class of alternatives even when there is more than one, has been of interest in the literature. Ideally one would like to use a UMP test. However, the UMP tests rarely exist in general frameworks with two-sided hypotheses or with multiparameters (see [8]). A UMP test was proved to exist for certain two-sided hypotheses problems for one-parameter exponential families (see [6]). Many UMP testing problems thus reduce to consider one of testings with a simple null hypothesis against a simple or composite alternative. For the ARMA(p, q) model with observations $\{X_t,\ t = 1, \ldots, n\}$ and noises $\{\epsilon_t,\ t = 1, \ldots, n\}$, a circular assumption that $X_{-\ell} \equiv X_{n-\ell}$ for $\ell = 0, 1, \ldots, p-1$ and $\epsilon_{-\ell} \equiv \epsilon_{n-\ell}$ for $\ell = 0, 1, \ldots, q-1$ was introduced by [1]. In this paper, we introduce the sufficiency, completeness, and uniformly minimum variance unbiased (UMVU) estimation for Gaussian circular ARMA processes. Then a UMP test for the coefficient parameter of Gaussian circular AR(1) models is discussed. Numerical studies show good performance of the UMP test for the testing problem with composite null and alternative hypotheses.

The rest of the paper is organized as follows. In Sect. 25.2, we introduce a UMVU estimator which is sufficient and complete for Gaussian circular ARMA models. In Sect. 25.3, we derive a UMP test with composite null and composite alternative hypotheses, assuming that the underlying family has monotone likelihood ratio. In Sect. 25.4, we conduct a numerical simulation to investigate the finite sample performance of a UMP test for the Gaussian circular AR(1) models. Section 25.5 concludes this paper.

25.2 Methodology

Suppose that $\{X_t,\ t = 1, \ldots, n\}$ is a Gaussian circular ARMA(p, q) process with zero mean and spectral density $f_{\boldsymbol{\theta}}(\lambda)$, where the parameter vector is denoted by $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_d)^\top \in \boldsymbol{\Theta} \subset \mathbb{R}^d$. Let

$$\mathrm{I}(\lambda_j) = \frac{1}{2\pi n} \left| \sum_{t=0}^{n-1} X_t e^{-it\lambda_j} \right|^2,$$

where $\lambda_j = 2\pi j/n$. Under the circular assumption, it was shown in [1] that $\mathrm{I}(\lambda_j)$, for $j = 1, \ldots, n$, is independently distributed as $f_{\boldsymbol{\theta}}(\lambda_j)\chi_2^2/2$. Letting $Y_j = f_{\boldsymbol{\theta}}(\lambda_j)\chi_2^2/2$, we can see that the joint probability density of $\mathbf{Y}_n \equiv (Y_1, \ldots, Y_n)^\top$ is given by

$$L(\boldsymbol{\theta}) = \left(\sqrt{2\pi f_{\boldsymbol{\theta}}(\lambda_0)\mathrm{I}(\lambda_0)} \prod_{j=1}^{\lfloor (n-1)/2 \rfloor} f_{\boldsymbol{\theta}}(\lambda_j)\right)^{-1} \exp\left\{-\frac{\mathrm{I}(\lambda_0)}{2 f_{\boldsymbol{\theta}}(\lambda_0)} - \sum_{j=1}^{\lfloor (n-1)/2 \rfloor} \frac{\mathrm{I}(\lambda_j)}{f_{\boldsymbol{\theta}}(\lambda_j)}\right\}.$$

Denote the log-likelihood by $\ell(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta})$.

Assumption 25.1 $\ell(\boldsymbol{\theta})$ is differentiable with respect to θ_k, $k = 1, \ldots, d$, and $\frac{\partial}{\partial\theta_k}\ell(\boldsymbol{\theta})$ is square-integrable for $k = 1, \ldots, d$.

Suppose that $\mathbf{T}(\mathbf{Y}_n) = (T_1(\mathbf{Y}_n), \ldots, T_d(\mathbf{Y}_n))^\top$ is an unbiased estimator of $\boldsymbol{\theta}$, i.e.,

$$E(\mathbf{T}(\mathbf{Y}_n)) = \boldsymbol{\theta}, \quad \boldsymbol{\theta} \in \Theta.$$

Then, given a set of values ϕ_k, $k = 1, \ldots, d$, with $(\phi_1, \ldots, \phi_d)^\top \in \boldsymbol{\Theta} \subset \mathbb{R}^d$, we have, for $k = 1, \ldots, d$,

$$\int_{-\infty}^{\infty} T_k(\mathbf{y}_n)\left\{\frac{L(\theta_1, \ldots, \theta_{k-1}, \phi_k, \theta_{k+1}, \ldots, \theta_d) - L(\boldsymbol{\theta})}{(\phi_k - \theta_k)L(\boldsymbol{\theta})}\right\}L(\boldsymbol{\theta})d\mathbf{y}_n = 1, \quad (25.1)$$

which, as $\phi_k \to \theta_k$, leads to

$$\int_{-\infty}^{\infty} T_k(\mathbf{y}_n)\left\{\frac{\partial}{\partial\theta_k}\log L(\boldsymbol{\theta})\right\}L(\boldsymbol{\theta})d\mathbf{y}_n = 1.$$

Hence we have

$$\mathrm{Cov}\left[T_k(\mathbf{Y}_n), \frac{\partial}{\partial\theta_k}\log L(\boldsymbol{\theta})\right] = 1, \qquad k = 1, \ldots, d.$$

By Schwarz's inequality,

$$V_{\boldsymbol{\theta}}(T_k(\mathbf{Y}_n)) \geq F_k(\boldsymbol{\theta})^{-1}, \quad k = 1, \ldots, d, \quad (25.2)$$

where $F_k(\boldsymbol{\theta}) = V_{\boldsymbol{\theta}}(\frac{\partial}{\partial\theta_k}\log L(\boldsymbol{\theta}))$. The equality in (25.2) holds, i.e., $\mathbf{T}(\mathbf{Y}_n)$ is efficient, if and only if

$$\mathbf{T}(\mathbf{Y}_n) = A_1(\boldsymbol{\theta})\frac{\partial}{\partial\boldsymbol{\theta}}\log L(\boldsymbol{\theta}) + a_2(\boldsymbol{\theta}) \quad \text{a.e.} \quad (25.3)$$

for some $d \times d$ matrix $A_1(\boldsymbol{\theta})$ and $d \times 1$ vector $a_2(\boldsymbol{\theta})$.

Integrating (25.3) with respect to $\boldsymbol{\theta}$, and solving this with respect to $L(\boldsymbol{\theta})$, we have

$$L(\boldsymbol{\theta}) = \exp\left[G_1(\boldsymbol{\theta})^\top\mathbf{T}(\mathbf{Y}_n) + g_2(\boldsymbol{\theta}) + U(\mathbf{Y}_n)\right] \quad \text{a.e.}$$

for some $d \times 1$ vector $G_1(\boldsymbol{\theta})$, and scalar-valued $g_2(\boldsymbol{\theta})$ and $U(\mathbf{Y}_n)$. We can see that $\mathbf{T}(\mathbf{Y}_n)$ is sufficient (see [6, Factorization theorem, p. 54]) and complete (see [6, Theorem 1, p. 149]).

Note that

$$\mathrm{I}(\lambda_j) = \frac{1}{2\pi}\sum_{l=-n+1}^{n-1}\hat{R}(\ell)e^{-i\ell\lambda_j}, \quad (25.4)$$

where $\hat{R}(\ell) = \frac{1}{n}\sum_{t=1}^{n-|l|} Y_t Y_{t+|\ell|}$.

Let $\boldsymbol{\theta} = (R(0), \ldots, R(d-1))^\top,\ d-1 \leq p-q$, where $R(j) = E(Y_t Y_{t+j})$ and $T_k(\mathbf{Y}_n) = \frac{2\pi}{n}\sum_{j=1}^{n} e^{i\lambda_j k} Y_j,\ k = 1, \ldots, d$. If $T_k(\mathbf{Y}_n)$ satisfies (25.3), there exist real-valued c_1, c_2, and c_3 such that

$$\sum_{j=1}^{n} e^{i\lambda_j k} Y_j = c_1 \sum_{j=1}^{n} \frac{-1}{f_{\boldsymbol{\theta}}(\lambda_j)} \frac{\partial}{\partial\theta_k} f_{\boldsymbol{\theta}}(\lambda_j) + c_2 \sum_{j=1}^{n} \frac{1}{f_{\boldsymbol{\theta}}(\lambda_j)^2} \frac{\partial}{\partial\theta_k} f_{\boldsymbol{\theta}}(\lambda_j) Y_j + c_3 \quad \text{a.e.}$$

Hence,

$$e^{i\lambda_j k} = \frac{c_2}{f_{\boldsymbol{\theta}}(\lambda_j)^2} \frac{\partial}{\partial\theta_k} f_{\boldsymbol{\theta}}(\lambda_j) \quad \text{a.e.}$$

From [5, p. 310], we obtain the following theorem.

Theorem 25.1 *Under Assumption 25.1, it holds that*
(i) $(T_1(\mathbf{Y}_n), \ldots, T_d(\mathbf{Y}_n))^\top = \left(\hat{R}(0), \ldots, \hat{R}(d-1)\right)^\top \overset{\text{say}}{=} \mathbf{S}$.
(ii) $\mathbf{S}$ *is sufficient and complete.*

Remark 25.1 $\hat{R}(\ell)$ is not unbiased for $R(\ell)$. However, $\hat{R}^*(\ell) := \frac{n}{n-|\ell|}\hat{R}(\ell)$ is unbiased for $R(\ell)$.

Remark 25.2 The conclusion (ii) in Theorem 25.1 is straightforward if we use the theory of exponential family.

Corollary 25.1 $\left(\hat{R}^*(0), \ldots, \hat{R}^*(d-1)\right)^\top$ *is a UMVU estimator of* $(R(0), \ldots, R(d-1))^\top$.

25.3 Tests by Monotone Likelihood Ratio

Let $P_\theta(\mathbf{Y}_n)$ be the joint probability density function (pdf) of $\mathbf{Y}_n$. The family $\{P_\theta(\mathbf{Y}_n), \theta \in \boldsymbol{\Theta}\}$ is said to have monotone likelihood ratio if there exists a real-valued function $\mathrm{T}(\mathbf{Y}_n)$ such that

(i) for any $\theta < \theta'$, the pdf P_θ and $P_{\theta'}$ are distinct, and
(ii) $P_{\theta'}(\mathbf{Y}_n)/P_\theta(\mathbf{Y}_n)$ is a non-decreasing function of $\mathrm{T}(\mathbf{Y}_n)$.

Then we have the following lemma.

Lemma 25.1 ([6, Theorem 3.4.1, p. 65]) *Suppose that $P_\theta(\mathbf{Y}_n)$ has monotone likelihood ratio in $T(\mathbf{Y}_n)$. Then for testing $H_0 : \theta \le \theta_0$ against $H_1 : \theta > \theta_0$, there exists a UMP test, which is given by*

$$\varphi(\mathbf{Y}_n) = \begin{cases} 1 & when \;\; T(\mathbf{Y}_n) > c \\ r & when \;\; T(\mathbf{Y}_n) = c \\ 0 & when \;\; T(\mathbf{Y}_n) < c, \end{cases}$$

where c and r are determined by $E_{\theta_0}(\varphi(\mathbf{Y}_n)) = \alpha$.

Now, let $\{Y_t\}$ be generated from the Gaussian circular AR(1) model:

$$Y_t = \theta Y_{t-1} + \epsilon_t,$$

where $\{\epsilon_t\}$ is a sequence of i.i.d. variables from $N(0, 1)$. The log-likelihood of $\mathbf{Y}_n$ is

$$-\int_{-\pi}^{\pi} |1 - \theta e^{i\lambda}|^2 \big\{\hat{R}(0) + e^{i\lambda}\hat{R}(1) + \cdots + e^{i(n-1)\lambda}\hat{R}(n-1)\big\} d\lambda.$$

Then the MLE of θ is given by $\hat{R}(1)/\hat{R}(0) = \mathrm{T}(\mathbf{Y}_n)$ (say). Repeating the discussion of (25.1)–(25.4), we have $L(\boldsymbol{\theta}) = \exp[G_1(\theta)\mathrm{T}(\mathbf{Y}_n) + g_2(\theta) + U(\mathbf{Y}_n)]$, which implies that $\mathrm{T}(\mathbf{Y}_n)$ is sufficient and complete. The likelihood ratio of $L(\boldsymbol{\theta})$ and $L(\boldsymbol{\theta}')$ is proportional to

$$\begin{aligned}
&\int_{-\pi}^{\pi} \big\{|1 - \theta e^{i\lambda}|^2 - |1 - \theta' e^{i\lambda}|^2\big\}\big\{\hat{R}(0) + e^{i\lambda}\hat{R}(1) + \cdots + e^{i(n-1)\lambda}\hat{R}(n-1)\big\} d\lambda \\
&\quad = \int_{-\pi}^{\pi} \big\{\theta^2 - (\theta')^2 + 2(-\theta + \theta')\cos\lambda\big\}\big\{\hat{R}(0) + \cdots + e^{i(n-1)\lambda}\hat{R}(n-1)\big\} d\lambda \\
&\quad = (\theta - \theta')(\theta + \theta')\hat{R}(0) + (-\theta + \theta')\hat{R}(1).
\end{aligned}$$

Hence, if we take $\mathrm{T}(\mathbf{Y}_n) = \hat{R}(1)/\hat{R}(0)$, we can apply Lemma 25.1 to obtain the following theorem.

Theorem 25.2 *The UMP test of $H_0 : \theta \le \theta_0$ against $H_1 : \theta > \theta_0$ at a significance level α has the rejection region $\mathrm{T}(\mathbf{Y}_n) > c$, where c is determined by*

$$\alpha = \mathrm{Pr}_{\theta_0}(\mathrm{T}(\mathbf{Y}_n) > c).$$

25.4 Simulation

In this section, we conduct a numerical simulation to investigate the finite sample performance of the proposed UMP test statistic $\mathrm{T}(\mathbf{Y}_n) = \hat{R}(1)/\hat{R}(0)$ of the hypothesis test for the Gaussian circular AR(1) model:

$$H_0 : \theta \leq \theta_0 \;\; against \quad H_1 : \theta > \theta_0.$$

We investigate the test performance under different parameters θ and sample sizes n for a Gaussian circular AR(1) model as an illustration. As stated in previous sections, the circular assumption in [1] requires that $Y_{-t} = Y_{n-t}$, for $t = 0, \ldots, p-1$, and $\epsilon_{-t} = \epsilon_{n-t}$, for $t = 0, \ldots, q-1$, for stationary ARMA(p, q) models. We now generate the Gaussian circular AR(1) model satisfying assumption $Y_0 = Y_n$. Let

$$\mathbf{M} = \begin{bmatrix} \mathbf{0} & 1 \\ \mathbf{I}_{n-1} & 0 \end{bmatrix}_{n \times n}$$

be an $n \times n$ matrix, where $\mathbf{I}_{n-1}$ is an $(n-1) \times (n-1)$ identify matrix. The Gaussian circular AR(1) model for $\mathbf{Y}_n$ is

$$\mathbf{Y}_n - \theta \mathbf{M} \mathbf{Y}_n = \boldsymbol{\epsilon}, \quad \boldsymbol{\epsilon} \sim N(\mathbf{0}_n, \mathbf{I}_n \sigma^2),$$

that is,

$$\begin{aligned} Y_1 - \theta Y_n &= \epsilon_1, \\ Y_t - \theta Y_{t-1} &= \epsilon_t, \;\; \text{for } t = 2, \ldots, n. \end{aligned} \tag{25.5}$$

Based on the model (25.5) we generate the Gaussian circular AR(1) model with $\theta_0 = 0.1, 0.3, 0.5, 0.7$, and 0.9, respectively. Since the distribution of the test statistic $\mathrm{T}(\mathbf{Y}_n)$ is not available, we consider using a Monte Carlo simulation to obtain the kernel density estimate (KDE) for the density of $\mathrm{T}(\mathbf{Y}_n)$. We compute the critical value c as $1 - \alpha$ quantile of its KDE, satisfying

$$\alpha = \Pr_{\theta_0}(\mathrm{T}^{(s)}(\mathbf{Y}_n) > c), \tag{25.6}$$

where $\mathrm{T}^{(s)}(\mathbf{Y}_n)$ denotes the value of the statistic in the sth iteration. We set $\alpha = 0.05$, $\theta_0 = 0.5$, $n = 100$, and replication $S = 1000$. It is known that the KDE suffers from the problem of bandwidth selection. Figure 25.1 shows the effect of bandwidth to the KDE of $\mathrm{T}(\mathbf{Y}_n)$. The left plot displays the KDEs with bandwidths $h = 0.005, 0.015, 0.025, 0.1$, and the right plot shows the KDE using optimal bandwidth selected by four methods: "Rule-of-thumb variation" (nrd), "Unbiased cross-validation" (ucv), "Biased cross-validation" (bcv), and "Sheather–Jones method" (SJ) (see [9]), respectively. It is not surprising that bandwidth has significant effect to the KDE of $\mathrm{T}(\mathbf{Y}_n)$ with a too small (large) bandwidth leading to undersmoothing (oversmoothing) KDE. Meanwhile, the optimal bandwidth works well for the density estimate and the performance is not sensitive to the bandwidth selection techniques. Hence, we adopt the default method "Rule-of-thumb variation" in R software for the bandwidth selection in the following study without loss of generality. Figure 25.2 displays the KDE ("Rule-of-thumb variation" method) of $\mathrm{T}(\mathbf{Y}_n)$ with n =100, and $\theta_0 = 0.1$, 0.5, and 0.9, respectively. It shows that when $\theta = 0.1$ and 0.5, the KDE

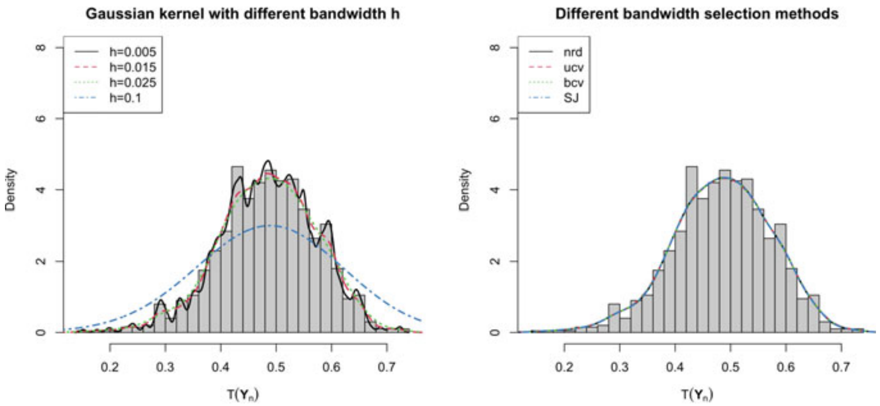


Fig. 25.1 The KDE of $\mathrm{T}(\mathbf{Y}_n)$ for the Gaussian circular AR(1) model with $\theta_0 = 0.5$ and $n = 100$. The left panel displays the KDE using different bandwidth h, the right panel shows the KDE with optimal bandwidth selected by different techniques

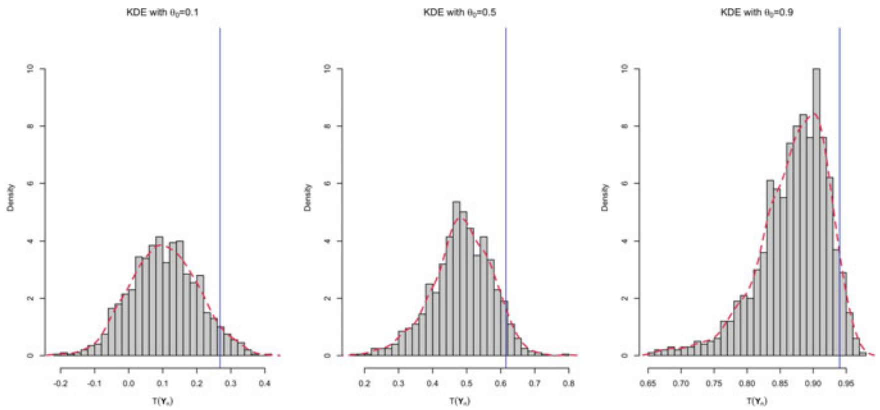


Fig. 25.2 The histogram and KDE of $\mathrm{T}(\mathbf{Y}_n)$ for the Gaussian circular AR(1) model with $\theta_0 = 0.1$, 0.5, and 0.9, respectively, and n =100. The KDE is obtained using the default Gaussian kernel with the optimal bandwidth selected by "Rule-of-thumb variation" technique in R software. The blue vertical line indicates the 95% quantile of empirical distribution

of $\mathrm{T}(\mathbf{Y}_n)$ displays more symmetric pattern, while it exhibits obvious left skewness when θ_0 is close to 1 ($\theta_0 = 0.9$).

Next, we generate new observations from the Gaussian circular AR(1) model with different AR coefficients θ. In specific, for $\theta_0 = 0.1$, 0.3, 0.5, and 0.7, we set θ as θ_0, $\theta_0 \pm 0.01$, and $\theta_0 \pm 0.2$, respectively. While for $\theta_0 = 0.9$, we set θ as θ_0, $\theta_0 \pm 0.01$, $\theta_0 - 0.2$, and $\theta = \theta_0 + 0.08$, respectively. The setup of the parameter values is shown in Table 25.1. The empirical size and power of the proposed test are measured by the rejection probability, defined by

Table 25.1 The test results based on the Gaussian circular AR(1) model with various θ_0, θ, and n. The average of test statistic $\mathrm{T}(\mathbf{Y}_n)$ and rejection probability ("Ratio" column) is reported

		$n = 50$		$n = 100$		$n = 500$		$n = 1000$	
θ_0	θ	$\mathrm{T}(\mathbf{Y}_n)$	Ratio	$\mathrm{T}(\mathbf{Y}_n)$	Ratio	$\mathrm{T}(\mathbf{Y}_n)$	Ratio	$\mathrm{T}(\mathbf{Y}_n)$	Ratio
0.1	−0.1	−0.099	0.002	−0.098	0.000	−0.100	0.000	−0.100	0.000
	0.09	0.083	0.042	0.084	0.030	0.091	0.030	0.091	0.029
	0.1	0.093	0.042	0.096	0.036	0.099	0.061	0.100	0.051
	0.11	0.093	0.034	0.103	0.044	0.108	0.083	0.110	0.102
	0.3	0.285	0.395	0.289	0.604	0.299	1.000	0.299	1.000
0.3	0.1	0.098	0.000	0.092	0.000	0.099	0.000	0.099	0.000
	0.29	0.272	0.047	0.280	0.061	0.286	0.019	0.290	0.029
	0.3	0.291	0.051	0.292	0.068	0.298	0.031	0.301	0.059
	0.31	0.294	0.058	0.306	0.088	0.309	0.063	0.308	0.087
	0.5	0.472	0.465	0.484	0.711	0.498	1.000	0.498	1.000
0.5	0.3	0.285	0.002	0.290	0.000	0.300	0.000	0.298	0.000
	0.49	0.461	0.041	0.478	0.033	0.486	0.035	0.489	0.027
	0.5	0.477	0.065	0.484	0.042	0.498	0.071	0.499	0.061
	0.51	0.478	0.070	0.494	0.057	0.506	0.114	0.508	0.116
	0.7	0.659	0.591	0.678	0.798	0.694	1.000	0.697	1.000
0.7	0.5	0.471	0.000	0.485	0.000	0.496	0.000	0.500	0.000
	0.69	0.652	0.033	0.670	0.035	0.686	0.026	0.686	0.017
	0.7	0.659	0.038	0.682	0.051	0.695	0.059	0.698	0.039
	0.71	0.669	0.057	0.693	0.066	0.706	0.092	0.707	0.112
	0.9	0.849	0.731	0.872	0.921	0.895	1.000	0.897	1.000
0.9	0.7	0.661	0.000	0.679	0.000	0.697	0.000	0.698	0.000
	0.89	0.840	0.049	0.862	0.021	0.884	0.010	0.888	0.004
	0.9	0.849	0.049	0.873	0.043	0.894	0.047	0.897	0.050
	0.91	0.855	0.085	0.882	0.087	0.905	0.133	0.908	0.203
	0.98	0.939	0.645	0.954	0.754	0.974	1.000	0.977	1.000

$$\sum_{s=1}^{S} I(\mathrm{T}^{(s)}(\mathbf{Y}_n) \geq c)/S,$$

with I being the indicator. We set $S = 1000$ and $n = 50, 100, 500, 1000$, respectively.

Table 25.1 reports the average value of the test statistic $\mathrm{T}(\mathbf{Y}_n)$ and the empirical size/power of the UMP test. The results show that the proposed UMP test generally delivers good performance under various combinations of θ_0, θ, and n. For example, under H_0 with $\theta \leq \theta_0$ and $\theta - \theta_0 = -0.2$, the rejection probability is close or equal to 0 for all experiments. Under the alternative hypothesis with $\theta - \theta_0 = 0.2$, the power equals 1 when $n \geq 500$.

The results also show that the test performance improves under greater divergence between θ and θ_0 and/or with larger sample sizes. For example, when $n = 50$ and H_1 holds, if the divergence, $\theta - \theta_0$, increases from 0.01 to 0.2, the power of the test also increases from 0.034 to 0.395, which indicates larger rejection probability and better test performance. Moreover, when $\theta_0 = 0.1$ and $\theta = 0.3$ (i.e., under H_1), the power of the test is 0.395 when $n = 50$, and increases to 1 when $n = 500$ or 1000. For the case of $\theta_0 = 0.9$ and $\theta = 0.89$ (i.e., under H_0), the rejection probability decreases from 0.049 to 0.004 when n increases from 50 to 1000.

When $|\theta - \theta_0| = 0.01$, i.e., the divergence is very small, the test performance is still acceptable if n is large. For example, for the case $\theta_0 = 0.5$ and $\theta = 0.51$ under H_1, the power of the test is 0.07 for $n = 50$ and 0.057 for $n = 100$, which are only slightly larger than the nominal level 0.05. On the other hand, when $n = 500$ or 1000, the power becomes 0.114 or 0.116, which is significantly larger than the nominal level. When $\theta = \theta_0$, the power varies around the nominal level. In summary, the UMP test statistic $\mathrm{T}(\mathbf{Y}_n)$ delivers good performance for the Gaussian circular AR(1) model.

25.5 Conclusion

In this paper, we have introduced the sufficiency, completeness, and UMVU estimator for the Gaussian circular ARMA models. We have proposed a UMP test by monotone likelihood ratio for the coefficient parameter of the Gaussian circular AR(1) models. In numerical study, the proposed UMP test provides good performance with satisfied empirical size and power under different combinations of AR coefficients and sample sizes for a Gaussian circular AR(1) model. We found that the UMP test has better performance under the greater divergence between θ and θ_0 or with larger sample sizes.

Acknowledgements We thank the editor and anonymous referees for their valuable time and insightful comments, which have helped improve this paper. This research was supported by JSPS KAKENHI Kiban (S) Grand-in-Aid No. 18H05290.

References

1. Anderson, T. (1977). Estimation for autoregressive moving average models in the time and frequency domains. *The Annals of Statistics* **5** 842–865.
2. Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. John Wiley & Sons, New York.
3. Chan, N. H. (2004). *Time Series: Applications to Finance*. John Wiley & Sons, New York.
4. Hamilton, J. D. (1994). *Time Series Analysis*. Princeton University Press.
5. Kakizawa, Y. and Taniguchi, M. (1994). Asymptotic efficiency of the sample covariances in a gaussian stationary process. *Journal of Time Series Analysis* **15** 303–311.

6. LEHMANN, E. L. AND ROMANO, J. P. (1969). *Testing Statistical Hypotheses*. Wiley : New York. Japanese translation by Shibuya, M. and Takeuchi, K. Iwanami publisher.
7. MANN, H. B. AND WALD, A. (1943). On the statistical treatment of linear stochastic difference equations. 173–220.
8. ROMANO, J. P., SHAIKH, A. M. AND WOLF, M. (2010). Hypothesis testing in econometrics. *Annu Rev Econ* **2** 75–104.
9. SHEATHER, S. J. AND JONES, M. C. (1991). A reliable data-based bandwidth selection method for kernel density estimation. *Journal of the Royal Statistical Society: Series B (Methodological)* **53** 683–690.

Chapter 26
A New Generalized Estimator for AR(1) Model Which Improves MLE Uniformly

Yujie Xue and Masanobu Taniguchi

Abstract For the first-order autoregressive model, Ochi (Journal of Time Series Analysis 4:57–67, 1983) introduced a generalized estimator of the coefficient α with two constants c_1 and c_2, which includes Daniels' estimator, least-squares estimator, and Durbin's estimator. From Fujikoshi and Ochi (Annals of the Institute of Statistical Mathematics 36:119–128, 1984), compared with a third-order approximated estimator of maximum likelihood estimator, it was shown that the modified maximum likelihood estimator (MLE) is better than the modified Ochi's estimator in the third-order sense if we modify the both estimators to be "third-order asymptotically median unbiased". In this paper, we propose a new estimator when the c_1 and c_2 depend on α, i.e., $c_1(\alpha)$ and $c_2(\alpha)$. Then we show that it improves the MLE uniformly in the sense of the third-order mean square error "without bias-adjustments". Because α is unknown, the feasible estimator with $c_1(\hat{\alpha})$ and $c_2(\hat{\alpha})$ is proposed with the replacement of α bt a consistent estimator $\hat{\alpha}$.

26.1 Introduction

In time series analysis, an autoregressive (AR) model is treated as one of the basic models used for forecasting when there is some correlation between values. Many linear and nonlinear models (autoregressive integrated moving average model, autoregressive conditional heteroscedasticity model, the threshold autoregression model, etc.) have been introduced based on this model as well. As a simple but effective model, AR models have been broadly applied in many fields such as finance, seismology, and brain-computer interface. To estimate the AR parameters, several techniques such as Kalman filter, Yule-Walker, expectation-maximization, least-square, Burg,

Y. Xue (✉)
Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
e-mail: yujiexue23@asagi.waseda.jp

M. Taniguchi
Waseda University, 3-4-1 Okubo, Shinjuku-ku, Tokyo 169-8555, Japan
e-mail: taniguchi@waseda.jp

Y. Liu et al. (eds.), *Research Papers in Statistical Inference for Time Series and Related Models*, https://doi.org/10.1007/978-981-99-0803-5_26

forward-backward algorithm, etc. were used. Especially, for the AR(1) coefficient, Ochi [5] proposed an estimator $\hat{\alpha}(c_1, c_2)$ by using two constants for the edge points of the data of size n. Yule-Walker's estimator [2], the least-square estimator, and Burg's estimator [3] of the AR(1) coefficient are examples with special choices of those two constants. Besides, Anderson [1] introduced a third-order approximated estimator $\tilde{\alpha}_{ML}$ for the maximum likelihood estimator (MLE) $\hat{\alpha}_{ML}$ whose explicit form is not available. In this paper, firstly, in Sect. 26.2 we review the comparison between $\hat{\alpha}(c_1, c_2)$ and $\tilde{\alpha}_{ML}$ when both are modified to be the third-order asymptotically median unbiased (AMU). It is shown that the modified MLE $\tilde{\alpha}^*_{ML}$ is better than the modified Ochi's estimator $\hat{\alpha}^*_n(c_1, c_2)$ uniformly in the third-order sense. In Sect. 26.3, we propose a new estimator whose constants c_1 and c_2 are functions of α, i.e., $\hat{\alpha}(c_1(\alpha), c_2(\alpha))$. It is shown that the new estimator improves the MLE uniformly in the sense of the third-order mean square error (MSE). Because α is unknown, we propose a feasible estimator $\hat{\alpha}(c_1(\hat{\alpha}), c_2(\hat{\alpha}))$, where $\hat{\alpha}$ is a consistent estimator. Lastly, in Sect. 26.4, some numerical studies are provided to illustrate the theoretical finding.

26.2 A Generalized Estimator of the AR Coefficient

Suppose that $\{y_t\}$ is generated by

$$y_t = \alpha y_{t-1} + u_t,$$

where $|\alpha| < 1$ and $u_t \sim N(0, \sigma^2)$. To estimate α, Ochi [5] introduced the estimator

$$\hat{\alpha}(c_1, c_2) = \frac{X_1}{X_2 + c_1 y_1^2 + c_2 y_n^2},$$

where c_1 and c_2 are constants, and $X_1 = \sum_{k=2}^{n} y_k y_{k-1}$, $X_2 = \sum_{k=2}^{n-1} y_k^2$. The estimator $\hat{\alpha}(1/2, 1/2)$ is Daniels' estimator, $\hat{\alpha}(1, 0)$ is the least-squares estimator, and $\hat{\alpha}(0, 1)$ is Durbin's estimator. Hence $\hat{\alpha}(c_1, c_2)$ includes a variety of famous estimators. Ochi [5] derived the third-order Edgeworth expansion of the distribution of $\sqrt{n}\{\hat{\alpha}(c_1, c_2) - \alpha\}$.

As for the another estimator of α, the MLE $\hat{\alpha}_{ML}$ is commonly used, but the explicit form of it is not available. Anderson [1] showed that

$$\tilde{\alpha}_{ML} = \left(1 - \frac{1}{n}\right) \frac{X_1}{X_2}$$

is the third-order approximation for $\hat{\alpha}_{ML}$.

Let $\tilde{\alpha}_{ML}^*$ and $\hat{\alpha}_n^*(c_1, c_2)$ be the modified third-order AMU estimators of $\tilde{\alpha}_{ML}$ and $\hat{\alpha}(c_1, c_2)$, respectively. Fujikoshi and Ochi [4] showed

$$\lim_{n\to\infty} n[P\{-a < \sqrt{n}(\tilde{\alpha}_{ML}^* - \alpha) < b\} - P\{-a < \sqrt{n}(\hat{\alpha}_n^*(c_1, c_2) - \alpha) < b\}] \geq 0$$

for any $a > 0$, $b > 0$. Hence the MLE is better than $\hat{\alpha}(c_1, c_2)$ in the sense of the third-order probability concentration around the true value, if we modify the both estimators to be third-order AMU.

26.3 A New Estimator Which Improves MLE Uniformly

In the previous section, we showed that, if we modify and $\hat{\alpha}(c_1, c_2)$ to be third-order AMU, the modified $\hat{\alpha}_{ML}^*$ is better than the modified $\hat{\alpha}^*(c_1, c_2)$ uniformly. However, comparison in the class of the third-order AMU may be unnatural. We should compare estimators "without bias or median adjustments". In view of this, we propose a new estimator which improves the MLE uniformly in the sense of the third-order MSE.

Now, we observe that

$$\begin{aligned}
\hat{\alpha}_n(c_1, c_2) &= \frac{X_1}{X_2\left\{1 + \frac{c_1 y_1^2}{X_2} + \frac{c_2 y_n^2}{X_2}\right\}} \\
&= \frac{X_1}{X_2}\left[1 - \frac{c_1 y_1^2}{X_2} - \frac{c_2 y_n^2}{X_2} + O_p\left(\frac{1}{n^2}\right)\right] \\
&= \left(1 - \frac{1}{n}\right)\frac{X_1}{X_2}\left(1 - \frac{1}{n}\right)^{-1}\left[1 - \frac{c_1 y_1^2}{X_2} - \frac{c_2 y_n^2}{X_2}\right] + O_p\left(\frac{1}{n^2}\right) \\
&= \tilde{\alpha}_{ML}\left[1 + \frac{1}{n} + \frac{1}{n^2} + \cdots\right]\left[1 - \frac{c_1 R(0) + c_1 y_1^2 - c_1 R(0)}{X_2}\right. \\
&\qquad \left. - \frac{c_2 R(0) + c_2 y_n^2 - c_2 R(0)}{X_2}\right] + O_p\left(\frac{1}{n^2}\right),
\end{aligned} \tag{26.1}$$

where $R(0) = \mathrm{E}y_t^2$. Note that

$$\begin{aligned}
X_2 &= \mathrm{E}X_2 + X_2 - \mathrm{E}X_2 \\
&= (n-2)R(0) + X_2 - \mathrm{E}X_2 \\
&= nR(0)\left[1 - \frac{2}{n} + \frac{X_2 - \mathrm{E}X_2}{nR(0)}\right],
\end{aligned} \tag{26.2}$$

which leads to

$$\frac{1}{X_2} = \frac{1}{nR(0)}\left[1 + \frac{2}{n} + O_p\left(\frac{1}{\sqrt{n}}\right)\right].$$

From (26.1) and (26.2), it follows that

$$\hat{\alpha}(c_1, c_2) = \tilde{\alpha}_{ML}\left[1 - \frac{c_1}{n} - \frac{c_2}{n} + \frac{1}{n}\right] + O_p\left(\frac{1}{n\sqrt{n}}\right). \tag{26.3}$$

Let $\tilde{c} = c_1 + c_2 - 1$. Applying Shiraishi et al. [6, Theorem 2(i)] to Eq. (26.3) (here, we set $s(\alpha) = \tilde{c}\alpha$, which is the difference between the second-order bias of $\tilde{\alpha}_{ML}$ and that of $\hat{\alpha}(c_1, c_2)$), we have

$$n^2[\mathrm{E}(\tilde{\alpha}_{ML} - \alpha)^2 - \mathrm{E}\{\hat{\alpha}(c_1, c_2) - \alpha\}^2] = \frac{2\tilde{c}}{g} + 2\mu_{ML}\tilde{c}\alpha - \tilde{c}^2\alpha^2 + o(1),$$

where $g = \mathrm{E}Z_\alpha^2$, and Z_α is the derivative of the log-likelihood of $\{y_t\}$ with respect to α. From Taniguchi [7, p. 31], we know $\mu_{ML} = -2\alpha$ and $g = \frac{1}{1-\alpha^2}$, hence, we obtain

$$n^2[\mathrm{E}(\tilde{\alpha}_{ML} - \alpha)^2 - \mathrm{E}\{\hat{\alpha}(c_1, c_2) - \alpha\}^2] = \phi(\tilde{c}) + o(1),$$

where

$$\phi(\tilde{c}) = \tilde{c}(2 - 6\alpha^2 - \tilde{c}\alpha^2) \leq \frac{(1 - 3\alpha^2)^2}{\alpha^2}$$

(the equality holds iff $\tilde{c} = \frac{1-3\alpha^2}{\alpha^2}$; hereafter, the i.i.d. case of $\alpha = 0$ is excluded). This motivates us to select Ochi's constants c_1 and c_2 as a function of α. That is, we suggest making use of $c_1(\alpha)$ and $c_2(\alpha)$, in such a way that the resulting estimator $\hat{\alpha}(c_1(\alpha), c_2(\alpha))$ improves the MLE. In summary, we have:

Theorem 26.1 *The estimator $\hat{\alpha}(c_1(\alpha), c_2(\alpha))$ improves the MLE $\tilde{\alpha}_{ML}$ uniformly in the sense of the third-order MSE, whenever $c_1(\alpha) + c_2(\alpha) - 1 = (1 - 3\alpha^2)/\alpha^2$.*

In Theorem 26.1, $(c_1(\alpha), c_2(\alpha))$ cannot be identified; however, if we take, for simplicity, $c_1(\alpha) = c_2(\alpha)$, i.e., $c_1(\alpha) = c_2(\alpha) = \frac{1}{2\alpha^2} - 1 = c(\alpha)$ (say), then the resulting (infeasible) estimator $\hat{\alpha}(c(\alpha), c(\alpha))$ improves the MLE uniformly. We propose a feasible estimator $\hat{\alpha}(c(\hat{\alpha}), c(\hat{\alpha}))$, where $\hat{\alpha}$ is a $\sqrt{n}$-consistent estimator of α. Then, we have

Theorem 26.2 *The estimator $\hat{\alpha}(c(\hat{\alpha}), c(\hat{\alpha}))$ improves the MLE uniformly in the sense of the third-order MSE.*

Table 26.1 The MSEs ($\times 10^{-3}$) when $n = 50, 100, 300$, where $\widehat{\alpha 1}$ stands for $\hat{\alpha}(c_1(\alpha), c_2(\alpha))$, and $\widehat{\alpha 2}$ and $\widehat{\alpha 3}$ stand for $\hat{\alpha}(c_1(\hat{\alpha}(1,0)), c_2(1,0))$ and $\hat{\alpha}(c_1(\hat{\alpha}(1/2,1/2)), c_2(\hat{\alpha}(1/2,1/2)))$, respectively

n	α	−0.9	−0.7	−0.5	−0.3	−0.1	0.1	0.3	0.5	0.7	0.9
50	$\tilde{\alpha}_{ML}$	6.100	12.766	15.708	17.641	17.626	19.121	16.899	15.353	12.874	6.006
	$\widehat{\alpha 1}$	6.238	12.677	15.955	16.097	8.440	8.657	16.512	15.572	12.671	6.057
	$\widehat{\alpha 2}$	6.229	12.750	16.213	16.429	8.364	8.540	16.934	15.797	12.758	6.064
	$\widehat{\alpha 3}$	6.235	12.755	16.210	16.428	8.362	8.539	16.930	15.795	12.762	6.073
100	$\tilde{\alpha}_{ML}$	2.053	5.481	7.310	9.033	9.104	9.457	8.554	7.092	5.919	2.701
	$\widehat{\alpha 1}$	2.039	5.460	7.363	8.650	6.291	6.424	8.515	7.231	5.848	2.545
	$\widehat{\alpha 2}$	2.047	5.499	7.522	8.977	6.565	6.602	8.797	7.357	5.886	2.555
	$\widehat{\alpha 3}$	2.048	5.500	7.522	8.976	6.565	6.601	8.796	7.354	5.887	2.556
300	$\tilde{\alpha}_{ML}$	0.692	1.760	2.808	3.251	3.303	3.320	3.073	2.624	1.808	0.668
	$\widehat{\alpha 1}$	0.676	1.760	2.801	3.205	2.989	3.054	3.052	2.624	1.808	0.644
	$\widehat{\alpha 2}$	0.676	1.760	2.802	3.234	3.116	3.173	3.058	2.625	1.812	0.645
	$\widehat{\alpha 3}$	0.676	1.760	2.802	3.234	3.116	3.173	3.058	2.625	1.812	0.645

26.4 Numerical Studies

In the simulations, we generate the Gaussian AR(1) model

$$y_t = \alpha y_{t-1} + u_t,$$

with $\alpha = -0.9, -0.7, -0.5, -0.3, -0.1, 0.1, 0.3, 0.5, 0.7, 0.9$, and $u_t \sim N(0,1)$. Then we calculate the MSEs of $\tilde{\alpha}_{ML}$, $\hat{\alpha}(c(\alpha), c(\alpha))$, and $\hat{\alpha}(c(\hat{\alpha}), c(\hat{\alpha}))$, using $\hat{\alpha} = \hat{\alpha}(1,0)$ (the least-squares estimator) or $\hat{\alpha}(1/2,1/2)$ (Daniels' estimator). $\hat{c}_1^*$ and $\hat{c}_2^*$ are calculated by the data of size 10000. Table 26.1 shows that the MSEs decrease as the sample size n increases, in which $\widehat{\alpha 1}$ stands for $\hat{\alpha}(c(\alpha), c(\alpha))$, and $\widehat{\alpha 2}$ and $\widehat{\alpha 3}$ stand for $\hat{\alpha}(c(\hat{\alpha}(1,0)), c(1,0))$ and $\hat{\alpha}(c(\hat{\alpha}(1/2,1/2)), c(\hat{\alpha}(1/2,1/2)))$, respectively. The proposed estimator has a better performance, compared to the MLE.

Acknowledgements The authors thank the reviewer for the valuable and constructive comments on the paper. This paper is supported by the Research Institute for Science & Engineering, Waseda University and the Japan Society for the Promotion of Science (JSPS) funding: Kiban(S)(18H05290).

References

1. Anderson, T. W. (1971). *The Statistical Analysis of Time Series*. Wiley: New York.
2. Daniels, H. E. (1956). The approximate distribution of serial correlation coefficients. *Biometrika* **43** 169–185.
3. Durbin, J. (1980). The approximate distribution of partial serial correlation coefficients calculated from residuals from regression on Fourier series. *Biometrika* **67** 335–349.

4. Fujikoshi, Y. and Ochi, Y. (1984). Asymptotic properties of the maximum likelihood estimate in the first order autoregressive process. *Ann. Inst. Statist. Math.* **36** 119–128.
5. Ochi, Y. (1983). Asymptotic expansions for the distribution of an estimator in the first-order autoregressive process. *J. Time Series Anal.* **4** 57–67.
6. Shiraishi, H., Taniguchi, M. and Yamashita, T. (2018). Higher-order asymptotic theory of shrinkage estimation for general statistical models. *J. Multivar. Anal.* **166** 198–211.
7. Taniguchi, M. (1991). *Higher Order Asymptotic Theory for Time Series Analysis*. Springer-Verlag: Berlin.