Springer Proceedings in Mathematics & Statistics

Soumendra Lahiri
Anton Schick
Ashis SenGupta
T.N. Sriram *Editors*

Contemporary Developments in Statistical Theory

A Festschrift for Hira Lal Koul

Springer

Springer Proceedings in Mathematics & Statistics

Volume 68

For further volumes:
http://www.springer.com/series/10533

Springer Proceedings in Mathematics & Statistics

This book series features volumes composed of select contributions from workshops and conferences in all areas of current research in mathematics and statistics, including OR and optimization. In addition to an overall evaluation of the interest, scientific quality, and timeliness of each proposal at the hands of the publisher, individual contributions are all refereed to the high quality standards of leading journals in the field. Thus, this series provides the research community with well-edited, authoritative reports on developments in the most exciting areas of mathematical and statistical research today.

Soumendra Lahiri • Anton Schick
Ashis SenGupta • T.N. Sriram
Editors

Contemporary Developments in Statistical Theory

A Festschrift for Hira Lal Koul

Springer

Editors
Soumendra Lahiri
North Carolina State University
Raleigh
North Carolina
USA

Ashis SenGupta
Applied Statistics Unit
Indian Statistical Institute
Kolkata
India

Anton Schick
Department of Mathematical Sciences
Binghampton University
Binghamton
New York
USA

T.N. Sriram
Department of Statistics
University of Georgia
Athens
Georgia
USA

ISSN 2194-1009 ISSN 2194-1017 (electronic)
ISBN 978-3-319-02650-3 ISBN 978-3-319-02651-0 (eBook)
DOI 10.1007/978-3-319-02651-0
Springer Cham Heidelberg New York Dordrecht London

Library of Congress Control Number: 2013956340

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Acknowledgments

The Editors thank the following colleagues who served as referees for one or more papers appearing in this volume:

M.Z. Anis, Indian Statistical Institute, Kolkata, India
J. Bertoin, Universität Zürich, Switzerland
Arijit Chakrabarti, Indian Statistical Institute, Kolkata, India
H.V. Kulkarni, Shivaji University, India
K.B. Kulasekera, Clemson University, USA

Thomas Lee, University of California at Davis, USA
Robert Lund, Clemson University, USA
William McCormick, University of Georgia, USA
Ian McKeague, Columbia University, USA
Joseph McKean, Western Michigan University, USA

D. Paindaveine, Universit e Libre de Bruxelles, Belgium
Cheolwoo Park, University of Georgia, USA
R.N. Rattihalli, University of Hyderabad, India
David Ruppert, Cornell University, USA
H. Sang, University of Mississippi, USA

K. Sato, Nagoya University, Japan
S. S. Sundarraman, New Jersey Institute of Technology, USA
Wolfgang Wefelmeyer, University of Cologne, Germany
Hongtu Zhu, University of North Carolina, Chapel Hill, USA

Contents

Contributors

I. V. Basawa Department of Statistics, University of Georgia, Athens, GA, USA

Manish C. Bhattacharjee Center for Applied Mathematics & Statistics, Department of Mathematical Sciences, New Jersey Institute of Technology, NJ, Newark, United States

V. Čekanavičius Department of Mathematics and Informatics, Vilnius University, Lithuania

Fuxia Cheng Department of Mathematics, Illinois State University, Normal, IL, USA

Somnath Datta Department of Biostatistics and Bioinformatics, University of Louisville, Louisville, Kentucky

Miguel A. Delgado Universidad Carlos III de Madrid, Madrid, Spain

Sunil K. Dhar Center for Applied Mathematics & Statistics, Department of Mathematical Sciences, New Jersey Institute of Technology, NJ, Newark, United States

J. Carlos Escanciano Indiana University, Bloomington, USA

Jiaying Gu University of Illinois, IL, Urbana, USA

Anirban Das Gupta Purdue University and Stanford University, USA

Jan Hannig Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, North Carolina, USA

Marc Hallin ECARES, Université libre de Bruxelles, Bruxelles, Belgium

ORFE, Princeton University, Princeton, NJ, USA

S. Y. Hwang Department of Statistics, Sookmyung Women's University, Seoul, Korea

Richard A. Johnson Department of Statistics, University of Wisconsin-Madison, Madison, WI, USA

Iain M. Johnstone Purdue University and Stanford University, USA

Jana Jurečková Faculty of Mathematics and Physics, Department of Statistics, Charles University in Prague, Prague, Czech Republic

Ingrid Van Keilegom Institut de statistique, Université catholique de Louvain, Louvain-la-Neuve, Belgium

Estate Khmaladze Victoria University of Wellington, Wellington, New Zealand

Roger Koenker University of Illinois, IL, Urbana, USA

Soumendra Lahiri North Carolina State Universy, North Carolina, Raleigh, USA

Fang Li Department of mathematical Sciences, Indiana University Purdue University at Indianapolis, Indianapolis, IN, USA

Linyuan Li Department of Mathematics and Statistics, University of New Hampshire, Durham, NH, USA

Kewei Lu Department of Mathematics and Statistics, University of New Hampshire, Durham, NH, USA

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

Ursula U. Müller Department of Statistics, Texas A&M University, College Station, TX, USA

Jan Picek Department of Applied Mathematics, Technical University in Liberec, Liberec, Czech Republic

Lianfen Qian Florida Atlantic University, Boca Raton, FL, USA

Lyudmila Sakhanenko Department of Statistics and Probability, Michigan State University, East Lansing, MI, USA

A. K. Md. Ehsanes Saleh School of Mathematics and Statistics, Carleton University, Ottawa, Canada

Anton Schick Department of Mathematical Sciences, Binghampton University, New York, Binghamton, USA

Ashis SenGupta Applied Statistics Unit, Indian Statistical Institute, Kolkata, India

Derek L. Sonderegger Department of Mathematics and Statistics, Northern Arizona Univeristy, Flagstaff, USA

Weixing Song Department of Statistics, Kansas State University, Manhattan, Kansas, USA

T.N. Sriram Department of Statistics, University of Georgia, Georgia, Athens, USA

Robert G. Staudte La Trobe University, Melbourne, Australia

Yvik Swan Faculté des Sciences, de la Technologie et de la Communication, Université de Luxembourg, Luxembourg, Grand Duchy of Luxembourg

P. Vellaisamy Department of Mathematics, Indian Institute of Technology Bombay, Mumbai, India

Thomas Verdebout EQUIPPE and INRIA, Université Lille 3, Villeneuve-d'Ascq Cedex, France

Yuankun Wang Department of Mathematics and Statistics, Lancaster University, United Kingdom

Guiping Yang Teva Pharmaceuticals, Malvern, USA

Song Yang Office of Biostatistics Research, National Heart, Lung, and Blood Institute, Bethesda, MD, USA

In-Kwon Yeo Department of Statistics, Sookmyung women's University, Seoul, Korea

Wei Zhang Wenzhou University, Zhejiang, China

Haitao Zheng Department of Statistics, South-west Jiaotong University, Chengdu, China

Chapter 1
Professor Hira Lal Koul's Contribution to Statistics

Soumendra Lahiri, Anton Schick, Ashis SenGupta and T.N. Sriram

Professor Hira Koul received his Ph.D. in Statistics from the University of California, Berkeley in 1967 under the supervision of Professor Peter Bickel. He has the unique distinction of being the first doctoral student of Professor Bickel. True to his training at Berkeley, in the initial years of his research career, he focused on developing asymptotic theory of statistical inference. He pioneered the approach of *Asymptotic Uniform Linearity* (AUL) as a theoretical tool for studying properties of the empirical process based on residuals from a semiparametric model. This approach has been widely employed by several authors in studying the asymptotic properties of tests of composite hyptheses, and has been a particularly powerful tool for deriving limit laws of goodness-of-fit tests. At around the same time, he also developed the theory of weighted empirical processes which played a fundamental role in the study of asymptotic distribution of robust estimators (e.g., Rank-based estimators and M-estimators) in linear regression models. An elegant account of the theory of weighted empirical processes for independent as well as dependent random variables is given in his monographs on the topic (Koul (1992, 2002)).

He has made significant contributions to several different areas of Statistics, including Asymptotic theory of efficient estimation, Bootstrap, Long-range dependence, Measurement Error, Robustness, Sequential Analysis, Survival Analysis, Nonlinear Time series, among others. In all his work, a common thread has been the

S. Lahiri (✉)
North Carolina State Universy, North Carolina, Raleigh, USA
e-mail: snlahiri@ncsu.edu

A. Schick
Department of Mathematical Sciences, Binghampton University,
New York, Binghamton, USA
e-mail: anton@math.binghamton.edu

A. SenGupta
Applied Statistics Unit, Indian Statistical Institute, Kolkata, India
e-mail: amsseng@gmail.com

T. N. Sriram
Department of Statistics, University of Georgia, Georgia, Athens, USA
e-mail: tn@stat.uga.edu

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_1,

use of rigorous mathematical arguments to derive useful statistical theory for estimation and testing. Here we highlight some of his major contributions to selected issues and problems to give a glimpse of the breadth and impact of his research. Building on his work on empirical processes, he developed asymptotic theory of minimum distance estimation in semi-parametric models. He also initiated the use of weighted empirical processes and repeatedly demonstrated its usefulness in studying limit distributions of classes of robust estimators, particularly the M- and R-estimators in regression models and in complex time series models. Starting in the 1980s, jointly with Professors V. Susarla and J. van Ryzin, he initiated the study of regression models in the presence of censoring and introduced the celebrated *Koul–Susarla–van Ryzin estimator* of the regression parameters in their 1980 *Annals of Statistics* paper. In contrast to its competitors, the Koul–Susarla–van Ryzin estimator is explicitly defined and easy to compute, which made it a popular choice among practitioners. Professor Koul further continued his work on censored data by establishing the Local Asymptotic Normality (LAN) property and results on asymptotic efficient estimation in semiparametric models.

Starting in the late 1980s, Professor Koul developed an interest in time series and Econometrics. He has made fundamental contributions to nonparametric and robust inference under complex temporal dependence structures, notably under long range dependence (LRD). In addition to developing asymptotic distributional theory for classes of robust estimators under LRD, jointly with Professor D. Surgailis, he derived higher order asymptotic expansions for M-estimators, which provided critical information into the structure of the successive smaller order terms. More recently, together with his long time collaborators Professors L Giraitis and D. Surgailis, he proved a Central Limit Theorem for periodogram based statistics under LRD requiring a weak Lindeberg-type condition. This is a highly effective tool for investigating asymptotic properties of such statistics, one that is bound to be used by researchers working with time series under LRD for years to come. The recent monograph, Koul, Giraitis and Surgailis (2013) gives an authoritative and detailed account of the statistical inference for time series under LRD, and contains many of Professor Koul's important results on the topic.

Many of Professor Koul's publications appeared in top-tier statistics journals. Given below is a chronological list of his publications to date.

Books:

1. *Weighted Empirical and Linear Models.* (1992). Lecture Notes-Monograph Series, 21, Institute of Mathematical Statistics, Hayward, California.
2. *Weighted Empirical Processes in Dynamic Nonlinear Models*. 2nd Edition. (2002). Lecture Notes Series in Statistics, 166, Springer, New York, N.Y., USA.
3. *Large Sample Inference For Long Memory Processes* (2013). Imperial College Press. London, UK. (with L. Giraitis and D. Surgailis).

Papers:

1. Asymptotic behavior of the Wilcoxon type condence regions for the multiple linear regression. (1969). *Ann. Math. Statist.* **40** 1950–1979.
2. A class of ADF tests for the subhypotheses in the multiple linear regression. (1970). *Ann. Math. Statist.* **41** 1273–1281.
3. Some convergence theorems for ranks and weighted empirical cumulatives. (1970). *Ann. Math. Statist.* **41** 1768–1773.
4. Asymptotic normality of random rank statistics. (1970). *Ann. Math. Statist.* **41** 2144–2149.
5. Asymptotic behavior of a class of condence regions based on ranks in regression. (1971). *Ann. Math. Statist.* **42** 466–476.
6. Some asymptotic results on random rank statistics. (1972). *Ann. Math. Statist.* **43** 842-859.
7. Asymptotic normality of signed rank statistics. (1972). *Z. Wahrscheinlichkeitstheorie, Verw. Geb.* **22** 293–300. (with R. G. Staudte, Jr.)
8. Weak convergence of weighted empirical cumulatives based on ranks. (1972). *Ann. Math. Statist.* **43** 832–841. (with R.G. Staudte, Jr.)
9. The Bahadur eciency of the Reimann-Vincze statistics. (1974). *Studia Scientiacarum Mathmaticarum Hungarica* **9** 399–403. (with M.P. Quine)
10. Asymptotic normality of H-L estimators based on dependent data. (1975). *J. Inst. Statist. Math.* **27** 429–441.
11. Power bounds for Smirnov test statistics in testing the hypothesis of symmetry. (1976). *Ann. Statist.* **4** 924–935. (Joint with R. G. Staudte, Jr.)
12. L^1 - rate of convergence for linear rank statistics. (1976). *Ann. Statist.* **4** 771–774. (with R.V. Erickson)
13. Behavior of robust estimators in the regression model with dependent errors. (1977). *Ann. Statist.* **5** 681–699.
14. A test for new better than used. (1977). *Communications: Statist. Theor. Meth.* **A6** 563–573.
15. A class of tests for new better than used. (1978). *Can. J. Statist.* **6** 249–471.
16. Testing for new is better than used in expectation. (1978). *Communications; Statist. Theory Meth.* **A7** 685–701.
17. Weighted empirical processes and the regression model. (1979). An invited paper for *J. of the Indian Statist. Assoc.* **17** 83–91.
18. Asymptotic tests of composite hypothesis for nonergodic type stochastic processes. (1979). *J. of Stoch. Proc. and Application* **9(3)** (with I.V. Basawa).
19. Some weighted empirical inferential procedures for a simple regression model. (1980). *Colloq. Math. Soc. Janos Bolyai* **32** 537–565.
20. Testing for new better than used in expectation with incomplete data. (1980). *J. Amer. Statist. Assoc.* **75** 952–956. (with V. Susarla).
21. A simulation study of some estimators of regression coefficients using censored data (1980). In *Proceedings of the annual meeting, American Statistical Association*. (with V. Susarla and J. Van Ryzin).
22. Regression analysis with randomly right censored data. (1981). *Ann. Statist.* **9** 1276–1288. (with V. Susarla and J. Van Ryzin).
23. A limit theorem for testing with randomly censored data. (1981). In *Survival Analysis, IMS Lecture Notes* **2** 189–205. (with V. Susarla).
24. Multi-step estimation of regression coecients in a linear model with censored survival data. (1981). In *Survival Analysis, IMS Lecture-Notes Monograph Series* **2** 85–100. (with V. Susarla and J. Van Ryzin).
25. Least square regression analysis with censored survival data. (1982). In *Topics in Applied Statistics* 151–165. (Eds: Chaubey, Y.P. & Dwivedi). T.D. Marcel Dekker, N.Y. (with V. Susarla and J. Van Ryzin).
26. Asymptotically minimax tests of composite hypotheses for nonergodic type processes. (1983). *J. of Stoch. Proc. & Applications* **14**. (with I.V. Basawa).

27. Minimum distance estimation in a linear regression. (1983). *Ann. Statist.* **11** 921–932. (with T. Dewet).
28. Adaptive estimation in regression. (1983). *Statistics and Decisions* **1** 379–400. (with V. Susarla).
29. Estimators of scale parameters in linear regression. (1983). *Statist. and Probab. Letters* **1** 273–277. (with V. Susarla).
30. LAN for randomly censored linear regression. (1984). *Statistics and Decision, Supplement Issue* **1** 17–30. (with W. H. Wang).
31. Test of goodness-of fit in linear regression. (1984). *Colloq. Math. Soc. Jonos. Bolyai* **45** 279–315.
32. Minimum distance estimation in multiple linear regression model. (1985). *Sankhya, Ser. A* **47** 57–74.
33. Minimum distance estimation in linear regression with unknown error distribution. (1985). *Statist. and Probab. Letters* **3** 1–8.
34. On a Kolmogorov-Smirnov type aligned test in linear regression. (1985). *Statist. & Probab. Letters* **3** 111–115. (with P.K. Sen).
35. Minimum distance estimation and goodness-of fit tests in first order autoregression. (1986). *Ann. Statist.* **14** 1194–1213.
36. An estimator of the scale parameter for the rank analysis of linear models under general score functions. (1987). *Scand. J. Statist.* **14** 131–143. (with G. Sievers and J. McKean).
37. Ecient estimation of location with censored data. (1988). *Statistics and Decisions* **4** 349–360. (with A. Schick and V. Susarla).
38. Large sample statistics based on quadratic dispersion. (1988). *Int. Statist. Rev.* **56** 199–219. (with I. V. Basawa).
39. Minimum distance estimation of scale parameter in the two sample problem: Censored and Uncensored Data. (1989). In *Recent Developments in Statistics and Their Applications* 117–134. (Eds—J. Klein and J. Lee). Freedom Press. (with S. Yang).
40. A quadraticity limit theorem useful in linear models. (1989). *Probab. Theory and Relat. Fields.* **82** 371–386.
41. Weak convergence of residual empirical process in explosive autoregression. (1989). *Ann. Statist.* **17** 1784–1794. (with S. Levental).
42. Weakly adaptive estimators in explosive autoregression. (1990). *Ann. Statist.* **18** 939–960. (with G. Pug.)
43. Weak convergence of a weighted residual empirical process in autoregression. (1991). *Statist. and Decis.* **9** 235–262. (with P. K. Sen).
44. Robustness of minimum distance estimation in linear regression against errors-in-variables model. (1991). In the *Proceedings of International Symposium on Nonparametric Statistics and Related Fields* 163–177. (Ed: A. K. Md. E. Saleh). Elsevier Science Publishers.
45. A weak convergence result useful in robust autoregression. (1991). *J. Statist. Planning and Infer.* **29** 291–308.
46. M-estimators in linear regression models with long range dependent errors. (1992). *Statist. and Probab. Letters* **14** 153–164.
47. Locally asymptotically minimax minimum distance estimators in linear regression. (1992). In the *Proceedings of the symposium on Order Statist. and Nonparmetrics in honor of A.E. Sarhan*, Alexendria, Egypt. (Eds - P.K. Sen and I.A. Salama). 405–417.
48. R-estimation of the parameters of autoregression models. (1993). *Ann. Statist.* **21** 534–551. (with A.K.Md.E. Saleh).
49. Bahadur representations for some minimum distance estimators in linear models. (1993). In *Statist. and Probab: A Raghu Raj Bahadur Festschrift*. 349–364. (Eds. J.K. Ghosh, S.K. Mitra, K.R. Parthasarathy, and B.L.S. Prakas Rao). Wiley Eastern Lmtd, Publishers. (with Z. Zhu.)
50. Asymptotics of R-, MD- and LAD-estimators in linear regression models with long range dependent errors. (1993). *Probab. Theory and Relat. Fields* **95** 535–553. (with K. Mukherjee).

51. Weak convergence of randomly weighted dependent residual empiricals with applications to autoregression. (1994). *Ann. Statist.* **22** 540–562. (with M. Ossiander).
52. On bootstrapping M-estimated residual processes in multiple linear regression models. (1994). *J. Mult. Analysis.* **49** 255–265. (with S. Lahiri).
53. Regression quantiles and related processes under long range dependent errors. (1994). *J. Mult. Analysis.* **51** 318–317. (with K. Mukherjee).
54. Minimum distance estimation of the center of symmetry with randomly censored data. (1995). *Metrika* **42** 79–97. (with S. Yang).
55. Auto-regression quantiles and related rank-score processes. (1995). *Ann Statist.* **23** 670–689. (with A.K. Md. Ehsanes Saleh).
56. Bahadur-Kiefer representations for GM-estimators in auto-regression models. (1995). *J. of Stoch. Proc. and Applications* **57** 167–189. (with Z. Zhu).
57. Asymptotics normality of Regression Estimators with long memory errors. (1996). *Statist. and Probab. Letters* **29** 317–335. (with L. Giraitis and D. Surgailis).
58. Asymptotics of some estimators and sequential empiricals in non-linear time series. (1996). *Ann. Statist.* **24** 380–404.
59. Adaptive estimation in a random coefficient autoregressive model. (1996). *Ann Statist.* **24** 1025–1054. (with A. Schick).
60. Asymptotics of M-estimators in non-linear regression with long range dependent errors. (1996). In the *proceedings of the Athens Conference on Apllied Probab. & Time Series, II, honoring E.J. Hannan: Lecture Notes in Statist.* **115** 272–290. (Eds.—P. M. Robinson and M. Rosenblatt). Springer Verlag, New York.
61. Efficient estimation in non-linear time series models. (1997). *Bernoulli* **3** 247–277. (with A. Schick).
62. Note on convergence rate of semiparametric estimators of the dependence index. (1997). *Ann. Statist.* **25** 1725–1739. (with Peter Hall and Berwin Turlach).
63. Testing for the equality of two nonparametric regression curves. (1997) *J. Statist. Planning & Inference* **65** 293–314. (with Anton Schick).
64. Asymptotic expansion of M-estimators with long memory errors. (1997). *Ann. Statist.* **25** 818–850. (with D. Surgailis).
65. Estimation of the dependence parameter in linear regression with long-range dependent errors. (1997). *J. Stoch. Proces. and Appl.* **71** 207–224. (with L. Giraitis).
66. Lack-of fit tests in regression with non-random design. (1998). In *Applied Statist. Science III; Nonparametric statistics & related fields: a volume honoring A.K.Md.E. Saleh.*. pp 53–70. (Eds: S. Ahmad, M. Ahsanullah & B. Sinha). Nova Sci. Publishers, Inc. (with W. Stute).
67. Regression model tting with long memory errors. (1998). *J. Statist. Planning & Inference* **71** 35–56. (with W. Stute).
68. Nonparametric model checks in time series. (1999). *Ann. Statist.* **27** 204–237. (with W. Stute).
69. Inference about the ratio of scale parameters in a two sample setting with current status data. (1999). *Statist. & Probab. Letters* **45** 359–370. (with A. Schick.)
70. Estimation of the dependence parameter in non-linear regression with random design and long memory errors. (2000). In *Perspectives in Statistical Sciences* (Eds - D. Basu, J.K. Ghosh, P.K. Sen & B.K. Sinha) pp. 191–208, Oxford University Press.
71. Asymptotic normality of the Whittle estimator in linear regression models with long memory errors. (2000). *Statist. Inference for Stochast. Processes* **3** 129–147. (with D. Surgailis).
72. Second order behaviour of M-estimators in linear regression with long memory param- eter. (2000). *J. Statist. Planning & Inference* **91** 399–412. (with D. Surgailis).
73. Asymptotics of empirical processes of long memory moving averages with innite vari- ance. (2001). *J. Stochastic Procsses & App.* **91** 309–336. (with D. Surgailis).
74. Asymptotics of maximum likelihood estimator in a two-phase linear regression model. February 2001. (2002). *J. Statist. Planning & Inference* **108** 99–119. (with L. Qian).

75. Robust estimators in regression models with long memory errors. In *Theory and Applications of Long Range Dependence*. 339–354. (Eds—G. Oppenheim, P. Doukhan and M. S. Taqqu). Birkhauser (2002). (with D. Surgailis).
76. Fitting a two phase linear regression model. (2000). *J. Indian Statist. Assoc.* **38** 331–353.
77. Asymptotics of M-estimators in two phase linear regression models. (2003). *J. Stochastic Processes & Applications* **103** 123–154. (with L. Qian & D. Surgailis).
78. Testing for superiority among two regression curves. (2003). *J. Statist. Planning & Inference* **117** 15–33. (with Anton Schick).
79. Asymptotic expansion of the empirical process of long memory moving averages. (2002). An invited review article for the book *Empirical Process Techniques for Dependent Data.* (Eds–Dehling, H.G., Mikosch, T. and Sorensen M.). Birkhauser pp. 213–239. (with D. Surgailis).
80. On weighted and sequential residual empiricals in ARCH models with some applications. (with Kanchan Mukherjee). Included in the monograph Weighted empirical processes in dynamic nonlinear models, second edition. (2002). Springer Lecture Notes, 166.
81. Asymptotic distributions of some scale estimators in nonlinear models. (2002). *Metrika* **55** 75–90.
82. Asymptotics of M-estimators in non-linear regression with long memory design. (2003). *Statist. & Probab. Letters* **61** 237–252. (with Baillie, R.T.)
83. Minimum distance estimation in a unit root autoregressive model. (2004). *J. Indian Statistical Assoc.* **41** 285–307. (with U. Naik-Nimbalkar).
84. Uniform reduction principle and some implications. (2004). (Invited paper) *J. Indian Statist. Assoc.* **21** 309–338. (with D. Surgailis).
85. Minimum distance regression model checking. (2004). *J. Statist. Planning & Inference* **119** 109–142. (with Pingping Ni).
86. Regression model checking with a long memory covariate process. (2004). *Econometric Theory* **20** 485–512. (with R.T. Baillie and D. Surgailis).
87. Martingale transforms goodness-of fit tests in regression models. (2004). *Ann. Statist.* **32** 995–1034. (with E. Khmaladze).
88. Model diagnosis for SETAR time series. (2005). *Statistica Sinica* **15** 795–817. (with W. Stute and Li, F.)
89. Testing for superiority among two time series. (2005). *Statist. Inference for Stochast. Processes* **6** (with Fang Li).
90. Goodness-of-fit testing in regression: A finite sample comparison of bootstrap methodology and Khmaladze transformation. (2005) *Statist. & Probab. Letters* **74** 290–302. (with Lyudmila Sakhanenko).
91. Fitting an error distribution in some heteroscedastic time series models. (2006). *Ann. Statist.* **34** 994–1012. (with Shiqing Ling).
92. Goodness-of-fit testing in interval censoring case 1. (2006). *Statist. & Probab. Letters* **76** 709–718. (with Tingting Yi).
93. Regression model fitting for the interval censored 1 responses. (2006). *Austrian J. Statist.* **35** 143–156. (with Tingting Yi).
94. Model Checks of Higher Order Time Series. (2006). *Statist. & Probab. Letters* **76** 1385–1396. (with W. Stute, M. Presedo Quindimil, and W. Gonzalez Manteiga).
95. Model Diagnostics via Martingale Transforms: A Brief Review. In *Frontiers in Statistics.* (2006), pp 183–206. Imperial College Press, London, UK. (Eds—J. Fan and H. L. Koul).
96. Nonparametric regression with heteroscedastic long memory errors. (2007). —it J. Statist. Planning & Inference **137** 379–404. (with H. Guo).
97. Serial auto-regression and regression rank scores statistics. (with Marc Hallin and Jana Jurechkova). An invited paper in *Advances in Statistical Modeling and Inference* (2007), 335–362. World Scientific, Singapore. (Editor: V. Nair).
98. Regression model checking with Berkson measurement errors. (2008). *J. Statist. Plan- ning & Inference* **138** 1615–1628. (with Weixing Song).

99. Asymptotic inference for some regression models under heteroscedasticity and long memory design and errors. (2008). *Ann. Statist.* **36** 458–487. (with H. Guo).
100. Minimum distance inference in unilateral autoregressive lattice processes. (2008). *Statistica Sinica* **18** 617–631. (with Marc Genton).
101. Testing of a sub-hypothesis in linear regression models with long memory covariates and errors. (2008). *Applications of Mathematics* **53** 235–248. (with Donatas Surgailis).
102. Minimum empirical distance goodness-of-fit tests for current status data. (2008). *J. Indian Statistical Association* **46** 79–124. (with D. Aggarwal).
103. Minimum distance regression model checking with Berkson measurement errors. (2009). *Ann. Statist.* **37** 132–156. (with Weixing Song).
104. Testing of a sub-hypothesis in linear regression models with long memory errors and deterministic design. (2009). *J. Statistical Planning & Inference* **139** 2715–2730. (with D. Surgailis).
105. Testing the tail index in autoregressive models. (2009). *Annals of Institute of Statistical Mathematics* **61** 579–598. (with J. Jurechkova and J. Picek).
106. Goodness-of-fit problem for errors in non-parametric regression: distribution free approach. (2009). *Ann. Statist.* **37** 3165–3185. (with E.V. Khmaladze).
107. Model checking in partial linear regression models with Berkson measurement errors. (2010). *Statistica Sinica* **20** 1551–1579. (with Weixing Song).
108. A class of minimum distance estimators in AR(p) models with infinite error variance. (2010). In *Nonparametrics and Robustness in Modern Statistical Inference and Time Series Analysis: A Festschrift in honor of Professor Jana Jureckov. IMS Collections* **7** 143–152. (Eds.—Antoch, J., Huskova, M. & Sen, P.K.) (with Xiaoyu Li).
109. Goodness of fit testing under long memory. (2010). *J. Statist. Planning & Inference* **140** 3742–3753. (with D. Surgailis).
110. Conditional variance model checking. (2010). *J. Statist. Planning & Inference* **140** 1056–1072. (with Weixing Song).
111. Khmaladze transformation. In *International Encyclopedia of Statistical Science*. Springer Verlag, Berlin, (*DOI 10.1007/978-3642-04898-2*). (2010).
112. Minimum distance lack-of-fit tests in fixed design. (2011). *J. Statist. Planning & Inference* **141** 65–79.
113. A goodness-of-fit test for GARCH innovation density. (2012). *Metrika* **75** 127–149. (with Nao Mimoto).
114. Lack-of-fit testing of the conditional mean function in a class of Markov duration models. (2012). *Econometric Theory* **28** 1283–1312. (with Indeewara Perera and Meryvn Silvaphulle). (*DOI: http://dx.doi.org/10.1017/ S0266466612000102*).
115. A class of goodness of fit tests in a linear errors-in-variables model. (2012). *J. French Statist. Soc.* **153** 52–70. (with Weixing Song).
116. Complete Case Analysis Revisited. (2012). *Ann. Statist.* **40** 3031–3049. (with U. Mueller-Harknett and Anton Schick).
117. Goodness-of-fit tests for long memory moving average marginal density. (2013). *Metrika* **76** 205–224. (Joint with N. Mimoto & D. Surgailis).
118. On asymptotic distributions of weighted sums of periodograms. (2013). *Bernoulli* (To appear). (with L. Giraitis).
119. Asymptotic normality for weighted sums of linear processes. (Submitted) (with K.M. Abadir, W. Distaso, L. Giraitis).
120. Automatic studentization in nonparametric regression. (Submitted) (with V. Dalla and L. Giraitis).
121. Large sample results for varying kernel regression estimates. (Submitted) (with Weixing Song).
122. Model checking in Tobit regression via nonparametric smoothing. (Submitted). (with Weixing Song and Shan Liu).
123. Minimum distance lack-of-fit tests under long memory errors. (Submitted). (with D. Surgailis).

Chapter 2
Martingale Estimating Functions for Stochastic Processes: A Review Toward a Unifying Tool

S. Y. Hwang and I. V. Basawa

SY Hwang is currently Head of the Department and the Director of Research Institute of Natural Sciences, Sookmyung Womens University.

IV Basawa is a Prof. Emeritus in the Department of Statistics at the University of Georgia.

2.1 Introduction

Various methods of estimation such as least squares, method of moments, maximum likelihood, pseudolikelihood, and quasilikelihood have been studied extensively in the literature. Historically, each of the estimating methods was developed individually to suit particular situations and at varying points of time. Large sample theory (covering consistency and limit distributions of the estimates) was also developed for each of the methods using diverse tools and limit theorems suited to the individual method. Most of the early work on estimation was devoted to independent observations. More recently, methods and theory of estimation (inference in general) have been extended to cover dependent observations in stochastic processes. See, for instance, Basawa and Prakasa Rao (1980a, b), Basawa and Koul (1988), and Basawa (1983, 2001). Martingale estimating functions provide a unified framework which covers various estimation methods under a single setting. See, among others, Godambe (1985); Bibby and Sorensen (1995); Wefelmeyer (1996); Basawa et al. (1997); and Heyde (1997). More recent research on large sample theory for estimating functions is focused on developing a unified approach to establish asymptotic optimality and large sample comparison of estimates obtained from estimating functions. Martingale limit theorems have proved useful when establishing large sample

S. Y. Hwang (✉)
Department of Statistics, Sookmyung Women's University, Seoul, Korea
e-mail: shwang@sookmyung.ac.kr

I. V. Basawa
Department of Statistics, University of Georgia, Athens, GA, USA
e-mail: ishwar@stat.uga.edu

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_2,

properties of estimates for dependent observations. This paper is an overview of current research by the authors (Hwang and Basawa 2011b) on martingale estimating functions and asymptotic optimality of parameter estimates for stochastic processes.

Section 2.2 presents a general formulation of martingale estimating functions which includes conditional least squares, quasilikelihood, maximum likelihood, and pseudolikelihood, as special cases. As an illustration, various estimates for a general class of generalized autoregressive conditional heteroskedasticity (GARCH)-type processes are presented in a unified way. Asymptotic optimality for a certain class of martingale estimating functions (MEFs) for ergodic processes is established via the convolution theorems in Sect. 2.3. Applications to conditional linear autoregressive processes, GARCH-type processes, and bifurcating autoregressive processes are then presented as examples of ergodic processes. Section 2.4 covers the extension of results of Sect. 2.3 to the nonergodic case. Branching Markov processes and explosive autoregressive processes are discussed to illustrate the nonergodic case. Finally, Sect. 2.5 gives a brief summary of the results and some concluding remarks.

2.2 Martingale Estimating Functions: A Formulation

Let $\{X_t, t = 0, 1, \ldots.\}$ denote a discrete time stochastic process defined on a probability space. Suppose that the probability measure P_θ associated with $\{X_t\}$ is indexed by a $(k \times 1)$ vector parameter θ. Assume that θ takes values in Θ which is an open subset of the k-dimensional Euclidean space. It is noted that P_θ needs not be parametric (in the sense that θ determines the underlying distribution). Rather, most of the theory in the paper is applicable to semiparametric or even nonparametric cases with a restriction depending on the parameter θ, allowing additional (infinite dimensional) nuisance parameter. Based on a sample of size n observations $X_1, X_2, \ldots, X_n$, we are concerned with estimating the parameter vector θ. Consider the following $(k \times 1)$ estimating function (EF) $U_n(\theta)$ given by

$$U_n(\theta) = \sum_{t=1}^{n} u_t(\theta) \tag{2.1}$$

where $\{u_t(\theta)\}$ is a sequence of martingale differences, i.e., $E\,(u_t(\theta)|F_{t-1}) = 0$. Here, F_t denote the σ-field generated by $X_t, X_{t-1}, \ldots, X_1$. We shall refer to $U_n(\theta)$ as the MEF. Assume for the moment that $\{X_t\}$ is strictly stationary and ergodic. Nonergodic cases will be discussed separately in Sect. 2.4. Fix $\theta \in \Theta$ and the local neighborhood $N_\delta(\theta)$ of the radius $\delta > 0$ about θ is defined by

$$N_\delta(\theta) = \{\, \theta^* \,;\, \sqrt{n}\,|(\theta^* - \theta)| < \delta\}. \tag{2.2}$$

where and throughout, the vector (or matrix) norm will be simply denoted by $|\cdot|$, viz., for any vector or matrix A, $|A|^2 = tr(A^T A) = tr(AA^T)$. Here, A^T is the transpose A. The neighborhood $N_\delta(\theta)$ is to be further specified in Sect. 2.4 for discussing "non-ergodic" cases. It is assumed that the $(k \times 1)$ vector $u_t(\theta)$ is differentiable (with

respect to θ). As to partial derivatives of column vector $U_n(\theta)$, $\partial U_n(\theta)/\partial\theta^T$ will be used as usual to denote $(k \times k)$ matrix of partial derivatives. In this paper, we shall confine ourselves to the case that $U_n(\theta)$ is regular in the following sense.

(C1: Regular MEF) For any radius $\delta > 0$, $U_n(\theta)$ satisfies, as $n \to \infty$

$$n^{-1} sup \mid \partial U_n(\theta^*)/\partial\theta^T - \partial U_n(\theta)/\partial\theta^T \mid = o_p(1)$$

where the sup is taken over $\theta^* \in N_\delta(\theta)$ and $o_p(1)$ stands for a term converging to zero in probability.

We now define a collection U of all regular MEFs $U_n(\theta)$. Some useful elements contained in U are illustrated below. As special cases of regular MEFs, Godambe (1985) considered the following "linear" MEF $G_n(\theta)$.

$$G_n(\theta) = \sum_{t=1}^{n} w_{t-1}(\theta) a_t(\theta) \tag{2.3}$$

where $a_t(\theta)$ is a prespecified martingale difference vector of dimension d and $w_{t-1}(\theta)$ is a $(k \times d)$ weight matrix whose components are F_{t-1} measurable. Godambe (1985) generated the Godambe-class of "linear" MEFs $G_n(\theta)$ by varying the "coefficients" $w_{t-1}(\theta)$ while $a_t(\theta)$, the innovation, being fixed. We shall refer to the Godambe-class as L which is clearly a subset of U. The Godambe-class L is known to be useful for the case when the likelihood is not known. Refer to, for instance, Hwang and Basawa (2011b).

Conditional Least Squares (LS) Let $m_t(\theta)$ denote the conditional mean of X_t given F_{t-1}, that is, $m_t(\theta) = E(X_t|F_{t-1})$. Consider $U_n(\theta) = \sum_{t=1}^{n} u_t(\theta)$ with

$$u_t(\theta) = \left(\frac{\partial m_t(\theta)}{\partial\theta}\right) \cdot (X_t - m_t(\theta)) \tag{2.4}$$

which is referred to as a LS-score (cf. Klimko and Nelson (1979)).

Quasilikelihood (QL) Let the conditional variance of X_t given F_{t-1} be denoted by $h_t(\theta) = Var(X_t|F_{t-1})$. Note that $m_t(\theta)$ and $h_t(\theta)$ are F_{t-1}-measurable and the parameter vector θ will be suppressed in $m_t(\theta)$ and $h_t(\theta)$ for notational simplicity. Consider a QL score (see, e.g., Godambe (1985))

$$u_t(\theta) = \left(\frac{\partial m_t}{\partial\theta}\right) \cdot h_t^{-1} \cdot (X_t - m_t)\,. \tag{2.5}$$

If we choose the innovation $a_t(\theta) = (X_t - m_t)$, LS and QL scores (2.4) and (2.5) belong also to the Godambe-class L.

Maximum-likelihood (ML) As an important member of U, one may consider the maximum likelihood (ML) score function by choosing $u_t(\theta)$ as the derivative of the log-conditional density of X_t given F_{t-1}, viz.,

$$u_t(\theta) = \frac{\partial ln p_t(\theta)}{\partial\theta} : (k \times 1)\text{vector} \tag{2.6}$$

where $p_t(\theta)$ denotes the conditional density of X_t given F_{t-1} and the property of $E\,(u_t(\theta)|F_{t-1}) = 0$ follows from the differentiability under the integral sign. We refer to, among others, Basawa et al. (1976) and Hwang and Basawa (1993) for a broad treatment of ML asymptotics in stochastic processes.

Pseudo-likelihood (PL) It is usually the case that the true likelihood is unknown to researchers, and thus we need to presume a tractable likelihood for the data which is called a PL. A PL may be a falsely specified likelihood. A pseudomaximum likelihood estimator is obtained by maximizing the objective function of PL score. Often, the PL is taken via Gaussian errors, standardized t-distributions with unknown degrees of freedom, and generalized error distributions (refer to, for instance, Tsay (2010, Chap. 10)). It is obvious that the PL-estimator reduces to the maximum likelihood (ML) estimator provided the PL coincides with the (unknown) true likelihood. It is interesting to note that even when the true likelihood is different from the PL, the PL estimator continues to be consistent and asymptotically normal under some regularity conditions (cf., Gourieroux (1997, Chap. 4), Hwang et al. (2013b)).

To better understand the members in the class U of regular MEFs, it will be illustrative to consider a general class of conditionally heteroscedastic processes. A general GARCH-type process is defined by

$$X_t = \sqrt{h_t}\, e_t \tag{2.7}$$

where $h_t(\theta) = Var(X_t|F_{t-1})$ and $\{e_t\}$ is independent and identically distributed (iid) with mean zero and variance unity. If we take $h_t = \alpha_0 + \alpha_1 X_{t-1}^2 + \beta_1 h_{t-1}$ with $\theta = (\alpha_0, \alpha_1, \beta_1)^T$, the process $\{X_t\}$ is called the GARCH of order one. Various GARCH-type models making some variations to the standard GARCH have been suggested and investigated in the literature. We refer to, for instance, a recent paper of Choi et al. (2012) and references therein for a broad class of nonlinear (asymmetric) GARCH processes. As an illustration of the asymmetric GARCH, consider the following threshold- GARCH process (T-GARCH) defined by

$$h_t = \alpha_0 + \alpha_{11} X_{t-1}^{+2} + \alpha_{12} X_{t-1}^{-2} + \beta_1 h_{t-1}$$

where X^+ and X^- denote the positive and negative functions respectively, that is, $X^+ = max(X, 0)$ and $X^- = max(-X, 0)$. If $\alpha_{11} = \alpha_{12}$, then the T-GARCH model reduces to the standard GARCH(1,1). Here, the functional form of h_t is not specified and therefore we are concerned with a broad class of GARCH-type processes. Suppose that no distributional assumptions on e_t are made, other than $E(e_t) = 0$ and $Var(e_t) = 1$. Then, the ML-score is not applicable in estimating the parameters. One may employ, e.g., a QL and a PL. For the QL, consider the martingale differences $\{X_t^2 - h_t\}$, and then generate the following Godambe-class of estimating functions defined by

$$\sum_{t=1}^{n} w_{t-1}(\theta)\left(X_t^2 - h_t(\theta)\right) \tag{2.8}$$

Assume the finite fourth order moment of e_t, i.e., $E(e_t^4) < \infty$. The QL score which enjoys certain optimal property within the Godambe class (2.8) is given by (refer to Godambe (1985); Heyde (1997); Hwang and Basawa (2011b))

$$QL(\theta) = \sum_{t=1}^{n} \frac{\partial h_t(\theta)}{\partial \theta} \left[(\zeta - 1)h_t^2(\theta)\right]^{-1} \left[X_t^2 - h_t(\theta)\right] \tag{2.9}$$

where $\zeta = E(e_t^4)$. Note that ζ turns out to be 3 when e_t is $N(0, 1)$. The QL estimator is obtained by solving $QL(\theta) = 0$. Now, we turn to the PL-score. Let us denote the pseudo density of the innovation e_t by $f(\cdot)$. The PL of the data is given by $\prod_{t=1}^{n} p\,(X_t|F_{t-1}) = \prod_{t=1}^{n} f\left[X_t/\sqrt{h_t}\right] h_t^{-1/2}$ where $p\,(X_t|F_{t-1})$ denotes a pseudo-conditional density of X_t given the past F_{t-1}, and $h_t = h_t(\theta)$. The PL estimator is obtained by solving $PL(\theta) = 0$ where

$$PL(\theta) = \sum_{t=1}^{n} l_t(\theta) \text{ with } l_t(\theta) = \partial lnf(e_t)/\partial\theta - \frac{1}{2} h_t^{-1} \partial h_t/\partial\theta \tag{2.10}$$

where $e_t = X_t/\sqrt{h_t(\theta)}$. In particular, if the PL is chosen based on the Gaussian likelihood, that is, if we take $f(\cdot)$ as $N(0, 1)$, it can be shown that the PL score $PL(\theta)$ reduces to

$$PL(\theta) = \sum_{t=1}^{n} \frac{x_t^2 - h_t}{2\, h_t^2} \frac{\partial h_t}{\partial \theta} \tag{2.11}$$

which is proportional to the QL score in (2.9). Consequently, it is interesting to note that the PL based on the Gaussian innovation is essentially the same as the QL based on the martingale differences $\{X_t^2 - h_t\}$. See Proposition 1 of Hwang et al. (2013b). We also note that the conditional least squares (CL) score is given by

$$CL(\theta) = \sum_{t=1}^{n} \frac{\partial h_t(\theta)}{\partial \theta} \left[X_t^2 - h_t(\theta)\right]. \tag{2.12}$$

The question that arises naturally is which (if any) MEF produces the "best" estimator within the class U. In the next section, via establishing the convolution theorem, the ML score is shown to be optimal within the class U in the sense of the minimum limit variance.

2.3 Convolution Theorems and Asymptotic Optimality

Consider any MEF $U_n(\theta) = \sum_{t=1}^{n} u_t(\theta) \in U$. Suppose that both the process $\{X_t\}$ and $\{u_t(\theta)\}$ are strictly stationary and ergodic. Nonergodic cases will be discussed in Sect. 2.4. One can obtain an estimator, say $\hat{\theta}_n$ of θ as a solution of $U_n(\theta) = 0$. The question regarding existence of strongly consistent solution $\hat{\theta}_n$ and its limit

distribution is addressed in the following theorem. In general, $u_t(\theta)$ may involve unobservable random variables. As an illustration, consider the GARCH-type process in (2.7) for which $u_t(\theta)$ may contain unobservable values $X_{-1},\ X_{-2},\ h_0,\ h_{-1}, \ldots$. If we treat these unobservable random variables as constants, $\{u_t(\theta)\}$ can not be strictly stationary (but it is asymptotically stationary). Now, we are willing to extend $\{X_t,\ t = 0, 1, 2, \ldots\}$ to two sided strictly stationary process $\{X_t,\ t = 0, \pm 1, \pm 2, \ldots\}$. Then, one may regard unobservable values (such as $X_{-1},\ X_{-2},\ h_0,\ h_{-1}$) involved in $u_t(\theta)$ as stationary random variables so that one can view the two-sided $\{u_t(\theta),\ t = 0, \pm 1, \pm 2, \ldots\}$ as a strictly stationary process. From now on, we shall treat $\{u_t(\theta)\}$ as being strictly stationary and ergodic. Define $(k \times k)$ matrices A and B such that

$$A = E\left(-\frac{\partial u_t(\theta)}{\partial \theta^T}\right) \tag{2.13}$$

and

$$B = Var\,[u_t(\theta)] = E\big(u_t(\theta)u_t^T(\theta)\big) \tag{2.14}$$

where the expectation is taken under the stationary distribution. It is noted that B is a symmetric matrix while A is permitted to be asymmetric, depending on the specification of $u_t(\theta)$.

Theorem 3.1 For any fixed $U_n(\theta) \in U$, we have as $n \to \infty$;

(1) With probability tending to one, there exists a strongly consistent estimator $\hat{\theta}_n$ such that $U_n\big(\hat{\theta}_n\big) = 0$.

(2)

$$\sqrt{n}\big(\hat{\theta}_n - \theta\big) \xrightarrow{d} N\big(0, M^{-1}\big) \quad \text{with} \quad M = A^T B^{-1} A. \tag{2.15}$$

Proof The proof is omitted since it follows essentially from standard arguments such as in, for instance, Basawa et al. (1976) and Klimko and Nelson (1979). Refer also to recent reference of Hwang et al. (2013a, Theorem 1) and Hwang et al. (2013b, Theorem 1).

Suppose that the true likelihood is known to us. The ML score, in particular, is denoted by $S_n(\theta)$;

$$S_n(\theta) = \sum l_t(\theta) \tag{2.16}$$

where $l_t(\theta)$ denotes $\frac{\partial ln p_t(\theta)}{\partial \theta}$ with $p_t(\theta)$ being the conditional density of X_t given F_{t-1}. See Eq. (2.6). Define $(k \times k)$ symmetric matrix

$$C = E\big(l_t(\theta) \cdot l_t^T(\theta)\big) = E\big(-\partial l_t(\theta)/\partial \theta^T\big). \tag{2.17}$$

Note that the covariance matrix between $u_t(\theta)$ and $l_t(\theta)$ is given by the matrix A, that is,

$$A = E\big(-\partial u_t(\theta)/\partial \theta^T\big) = E\big(u_t(\theta) \cdot l_t^T(\theta)\big). \tag{2.18}$$

The ML estimator, say $\hat{\theta}_{ML}$, is obtained from solving $S_n(\theta) = 0$. For $\hat{\theta}_{ML}$, it is obvious that $A = B = C$ and thus it follows readily from Theorem 3.1-(2) that

$$\sqrt{n}\big(\hat{\theta}_{ML} - \theta\big) \xrightarrow{d} N\big(\,0\,,\; C^{-1}\big). \tag{2.19}$$

It will be shown that the ML estimator $\hat{\theta}_{ML}$ is asymptotically optimal in the sense of having the "smallest" covariance matrix among all the estimators $\hat{\theta}_n$ within U. To do this, we state the following convolution theorem due to Hwang and Basawa (2011a).

Theorem 3.2 (Convolution theorem within the class U) Consider any $U_n(\theta) \in U$ and the consistent solution $\hat{\theta}_n$ of $U_n(\theta)$ addressed in Theorem 3.1. Then, $\sqrt{n}\big(\hat{\theta}_n - \theta\big)$ can be expressed as a sum of two independent random variables, say, $N_1(\theta)$ and $N_2(\theta)$ where $N_1(\theta)$ and $N_2(\theta)$ follow $N\big(0, (A^T B^{-1} A)^{-1} - C^{-1}\big)$ and $N\big(0, C^{-1}\big)$ in limit, respectively.

Remark The convolution theorem implies that $\big(A^T B^{-1} A\big)^{-1} - C^{-1}$ is nonnegative definite. Comparing (2.15) and (2.19), the ML estimator $\hat{\theta}_{ML}$ attains the "smallest" variance–covariance matric C^{-1} within the MEF class U.

Proof Since the proof follows essentially the same lines as in Theorem 3.3 of Hwang and Basawa (2011a), we provide outlines only, omitting details. A martingale central limit theorem gives

$$n^{-1/2} \begin{pmatrix} U_n(\theta) \\ S_n(\theta) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} B & A \\ A^T & C \end{pmatrix} \right) \tag{2.20}$$

and

$$\begin{pmatrix} \sqrt{n}(\hat{\theta}_n - \theta) \\ n^{-1/2} S_n(\theta) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} (A^T B^{-1} A)^{-1} & I_p \\ I_p & C \end{pmatrix} \right). \tag{2.21}$$

Consider the following expression

$$\sqrt{n}(\hat{\theta} - \theta) = N_{1n}(\theta) + N_{2n}(\theta)$$

where

$$N_{1n}(\theta) = \sqrt{n}(\hat{\theta} - \theta) - C^{-1} n^{-1/2} S_n(\theta) \quad \text{and} \quad N_{2n}(\theta) = C^{-1} n^{-1/2} S_n(\theta).$$

Equivalently, we have

$$\begin{pmatrix} N_{1n}(\theta) \\ N_{2n}(\theta) \end{pmatrix} = \begin{pmatrix} I & -C^{-1} \\ 0 & C^{-1} \end{pmatrix} \begin{pmatrix} \sqrt{n}(\hat{\theta}_n - \theta) \\ n^{-1/2} S_n(\theta) \end{pmatrix}. \tag{2.22}$$

Then, it follows from (2.21) that $\begin{pmatrix} N_{1n}(\theta) \\ N_{2n}(\theta) \end{pmatrix}$ is asymptotically normal with mean zero and variance–covariance matrix given by

$$\begin{pmatrix} (A^T B^{-1} A)^{-1} - C^{-1} & 0 \\ 0 & C^{-1} \end{pmatrix}.$$

This completes the proof.

Although the ML estimator $\hat{\theta}_{ML}$ is asymptotically optimal in the sense of having "the smallest variance", it is usually the case in stochastic processes that the exact likelihood is unknown to researchers and therefore $\hat{\theta}_{ML}$ is not available. Without the knowledge of the likelihood, one may consider the QL score $QL(\theta)$ instead of the ML score. Godambe (1985) established certain "optimality" of the $QL(\theta)$ within the Godambe class L of linear MEFs $G_n(\theta)$ defined in (2.3), viz.,

$$G_n(\theta) = \sum_{t=1}^{n} w_{t-1}(\theta) a_t(\theta) \tag{2.23}$$

in which we consider the scalar innovation $a_t(\theta)$ for the simplicity of presentation. Define $(k \times k)$ matrices H and J;

$$H = E\big(- \partial(w_{t-1}a_t)/\partial\theta^T \big) = E\big(- w_{t-1}(E_{t-1}Da_t)^T \big) \tag{2.24}$$

and

$$J = Var\big(w_{t-1}a_t\big) = E\big(w_{t-1}w_{t-1}^T E_{t-1}a_t^2\big) \tag{2.25}$$

where Da_t represents $\partial a_t/\partial\theta$ and θ is suppressed in $w_{t-1}(\theta)$ and $a_t(\theta)$. Here and in what follows E_{t-1} denotes the conditional expectation given F_{t-1}. Let $\hat{\theta}_n$ denote the consistent solution of $G_n(\theta) = 0$ for $G_n(\theta) \in G$. It then follows from (2.15) that

$$\sqrt{n}\big(\hat{\theta}_n - \theta\big) \xrightarrow{d} N\big(0, (H^T J^{-1} H)^{-1}\big). \tag{2.26}$$

The QL score due to Godambe (1985) is obtained by

$$QL_{(\theta)} = \sum_{t=1}^{n} w_{t-1}^{O}(\theta) a_t(\theta) \tag{2.27}$$

for which $w_{t-1}^{O}(\theta) = E_{t-1}\big[\partial a_t(\theta)/\partial\theta\big]/E_{t-1}\big[a_t^2(\theta)\big]$. Let us denote the consistent solution from $QL(\theta) = 0$ by $\hat{\theta}_{QL}$. It is not difficult to see that

$$\sqrt{n}\big(\hat{\theta}_{QL} - \theta\big) \xrightarrow{d} N\big(\,0\,,\; K^{-1}\big) \tag{2.28}$$

where

$$K = E\left[\big(E_{t-1}Da_t\big)\big(E_{t-1}Da_t\big)^T/E_{t-1}a_t^2\right]. \tag{2.29}$$

By establishing the following convolution theorem within the restricted Godambe class L, Hwang and Basawa (2011b) argued that the matrix $(H^T J^{-1} H)^{-1} - K^{-1}$ is nonnegative definite, which implies that $\hat{\theta}_{QL}$ is better than $\hat{\theta}_n$ in the sense of the "smaller" asymptotic variance.

Theorem 3.3 (Convolution theorem within the Godambe class L) Split $\hat{\theta}_n$ into two components, viz.,

$$\sqrt{n}(\hat{\theta}_n - \theta) \;=\; Y_{1n}(\theta) \;+\; Y_{2n}(\theta) \tag{2.30}$$

where

$$Y_{1n}(\theta) \;=\; \sqrt{n}(\hat{\theta}_n - \theta) \;-\; K^{-1}QL(\theta)/\sqrt{n}\;;\; Y_{2n}(\theta) \;=\; K^{-1}QL(\theta)/\sqrt{n}. \tag{2.31}$$

Then, $Y_{1n}(\theta)$ and $Y_{2n}(\theta)$ are asymptotically independent normal random vectors. Specifically, we have

$$\begin{pmatrix} Y_{1n}(\theta) \\ Y_{2n}(\theta) \end{pmatrix} \xrightarrow{d} N\Big(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} (H^T J^{-1} H)^{-1} - K^{-1} & 0 \\ 0 & K^{-1} \end{pmatrix}\Big). \tag{2.32}$$

Proof Refer to Theorem 3 of Hwang and Basawa (2011b).

We have discussed asymptotic optimality for $\hat{\theta}_{ML}$ and $\hat{\theta}_{QL}$ within the class U and L, respectively. When the likelihood is known, we may go ahead and use $\hat{\theta}_{ML}$. If the likelihood is not available, $\hat{\theta}_{QL}$ will be a good choice as an alternative to $\hat{\theta}_{ML}$. Illustrative examples follow.

2.3.1 Conditionally Linear (AR(1)) Processes (CLAR(1))

Grunwald et al. (2000) introduced a class of conditionally linear AR(1) models (CLAR(1)) defined by

$$m_t(\theta) = \theta_1 + \theta_2 X_{t-1}, \quad -\infty < \theta_1 < \infty,\ |\theta_2| < 1. \tag{2.33}$$

where we do not require the knowledge of the likelihood. Note that the conditional mean $m_t(\theta)$ is linear in terms of the parameter $\theta = (\theta_1, \theta_2)^T$. Grunwald et al. (2000) argued that the CLAR(1) class contains a large number of models in the literature, including standard AR(1) process, random coefficient AR(1) model and various integer-valued thinning models as special cases. Hwang and Basawa (2009, 2011b) reviewed the class in the context of estimating function approach. Assume that the conditional variance $h_t \;=\; Var(X_t|F_{t-1})$ is known. To construct Godambe class, set $a_t(\theta) \;=\; X_t \;-\; m_t(\theta)$. It is easily seen that

$$\hat{\theta}_{QL} = \begin{pmatrix} \sum h_t^{-1} & \sum X_{t-1} h_t^{-1} \\ \sum X_{t-1} h_t^{-1} & \sum X_{t-1}^2 h^{-1} \end{pmatrix}^{-1} \begin{pmatrix} \sum X_t h_t^{-1} \\ \sum X_t X_{t-1} h_t^{-1} \end{pmatrix} \tag{2.34}$$

and the limit distribution is given by

$$\sqrt{n}(\hat{\theta}_{QL} - \theta) \xrightarrow{d} N\big(0, K^{-1}\big)$$

which gives the "minimum" variance–covariance matrix K^{-1} within the Godambe class. Refer to Hwang and Basawa (2011b) for further details. The least squares estimator $\hat{\theta}_n$ belonging to the Godambe class is seen to be

$$\hat{\theta}_n = \begin{pmatrix} n & \sum X_{t-1} \\ \sum X_{t-1} & \sum X_{t-1}^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum X_t \\ \sum X_t X_{t-1} \end{pmatrix}.$$

It is obvious from Theorem 3.3 that $\hat{\theta}_{QL}$ is better than $\hat{\theta}_n$.

2.3.2 A General GARCH-Type Processes

Revisit a general GARCH-type process

$$X_t = \sqrt{h_t}\, e_t \tag{2.35}$$

where $\{e_t\}$ is a sequence of iid random errors with mean zero and unit variance, having density $f_e(\cdot)$. Here, h_t denotes the conditional variance of X_t given F_{t-1}. We here do not specify the functional form of h_t and accordingly a broad class of conditionally heteroscedastic time series $\{X_t\}$ with a general form of h_t will be discussed. Hwang et al. (2013a) investigated general GARCH-type processes in order to compare various MEFs. Discussions below are adapted from Hwang et al. (2013a). First consider the conditional least squares score $CL(\theta)$ given in (2.12). It is noted that $Var\big(X_t|F_{t-1}\big) = h_t(\theta)$ and

$$Var\big(X_t^2|F_{t-1}\big) = \big(\zeta - 1\big)h_t^2(\theta) \tag{2.36}$$

where $\zeta = E\big(e_t^4\big)$. In order to determine the limiting distribution of $\hat{\theta}_{LS}$, one can obtain

$$H = E\big(u_t(\theta)u_t^T(\theta)\big) = (\zeta - 1)E\left[h_t^2\left(\frac{\partial h_t(\theta)}{\partial \theta}\right)\left(\frac{\partial h_t(\theta)}{\partial \theta}\right)^T\right]$$

and

$$J = E\left(-\frac{\partial u_t(\theta)}{\partial \theta}\right) = E\left[\left(\frac{\partial h_t(\theta)}{\partial \theta}\right)\left(\frac{\partial h_t(\theta)}{\partial \theta}\right)^T\right].$$

We thus have

$$\sqrt{n}\big(\hat{\theta}_{LS} - \theta\big) \xrightarrow{d} N\big(0, H^{-1}JH^{-T}\big). \tag{2.37}$$

Next, we consider the QL score $QL(\theta)$ given in (2.9). The QL estimator $\hat{\theta}_{QL}$ is obtained from $QL(\theta) = 0$. It can be verified that

$$K = (\zeta - 1)^{-1}E\left[h_t^{-2}\left(\frac{\partial h_t(\theta)}{\partial \theta}\right)\left(\frac{\partial h_t(\theta)}{\partial \theta}\right)^T\right] \tag{2.38}$$

which in turn yields

$$\sqrt{n}(\hat{\theta}_{QL} - \theta) \xrightarrow{d} N\big(0, K^{-1}\big). \tag{2.39}$$

Note that $\hat{\theta}_{QL}$ is better than $\hat{\theta}_{LS}$ in the sense of Theorem 3.3. Let f_e denote the true density of e_t. The ML estimator $\hat{\theta}_{ML}$ is obtained from $S_n(\theta) = 0$ where $S_n(\theta)$ is defined in (2.9). It then follows that

$$\sqrt{n}(\hat{\theta}_{ML} - \theta) \xrightarrow{d} N\big(0, C^{-1}\big)$$

where $C = E\big(l_t(\theta) l_t^T(\theta)\big) = E\big(-\partial l_t(\theta)/\partial\theta^T\big)$. Although $\hat{\theta}_{ML}$ is optimum in the sense of "minimum variance" within U, it is noted that $\hat{\theta}_{ML}$ requires a specification of the density f_e (of e_t) while $\hat{\theta}_{QL}$ can be easily implemented regardless of f_e. Refer to Hwang et al. (2013a) for further details.

2.3.3 *Bifurcating Autoregressive Processes (BAR)*

Cowan and Stuadte (1986) introduced (BAR) processes indexed by a binary-splitting tree for cell lineage study. A first-order BAR(1) process $\{X_t, t = 1, 2, \dots\}$ is defined recursively by

$$\begin{aligned} X_{2t} &= \theta X_t + \epsilon_{2t} \\ X_{2t+1} &= \theta X_t + \epsilon_{2t+1} \end{aligned} \tag{2.40}$$

where $|\theta| < 1$ and $\{\epsilon_t\}$ is a sequence of iid random errors with mean zero and variance σ_ϵ^2. In a BAR model, observations are indexed by a bifurcating tree where each individual (mother) in one generation produces two individuals (sisters) in the next generation. For each individual, an observation X is recorded. The BAR and various BAR type models have been studied by several authors including, among others, Hwang and Basawa (2009, 2011a) and Hwang and Kang (2012). Let $t(1)$ denote the first ancestor (i.e., mother) of the individual t. It can be shown that $t(1) = [t/2]$ where $[\,\cdot\,]$ denotes the greatest integer function. The BAR(1) in (2.40) can be rewritten in terms of a single equation as (cf. Hwang and Kang 2012)

$$X_t = \theta X_{t(1)} + \epsilon_t, \; t \geq 2. \tag{2.41}$$

Consider the following nonlinear BAR process defined by

$$X_t = m_t + \sqrt{h_t} \cdot e_t \tag{2.42}$$

where the variance of e_t is set to be unity and $m_t = m(X_{t(1)})$ and $h_t = h(X_{t(1)})$ stand respectively for the conditional mean and variance function defined by

$$m_t = E\big(X_t | X_{t(1)}\big) \text{ and } h_t = Var\big(X_t | X_{t(1)}\big) \tag{2.43}$$

It is noted that m_t and h_t are $X_{t(1)}-$measurable. Suppose that the distribution of e_t is not known and then we rely on a QL score. Let

$$a_t(\theta) = \begin{pmatrix} X_t - m_t(\theta) \\ (X_t - m_t(\theta))^2 - h_t(\theta) \end{pmatrix} \tag{2.44}$$

and consider the Godambe class L given in (2.3) with $k = 2$. Denote the (2×2) conditional variance–covariance matrix of $a_t(\theta)$ given the first ancestor $X_{t(1)}$ by $V_t(\theta)$, viz.,

$$V_t(\theta) = E\big[a_t(\theta)a_t^T(\theta)|X_{t(1)}\big] = \begin{pmatrix} h_t(\theta) & \mu_{3t}(\theta) \\ \mu_{3t}(\theta) & \mu_{4t}(\theta) - h_t^2(\theta) \end{pmatrix}. \tag{2.45}$$

where $\mu_{3t}(\theta) = E\big[(X_t - m_t(\theta))^3|X_{t(1)}\big]$ and $\mu_{4t}(\theta) = E\big[\big(X_t - m_t(\theta)\big)^4|X_{t(1)}\big]$ (cf. Hwang et al. (2013b)). The conditional expectation of the derivative matrix of $a_t(\theta)$ is given by

$$E\big[\partial a_t(\theta)/\partial\theta^T|X_{t(1)}\big] = \begin{pmatrix} -(\partial m_t(\theta)/\partial\theta)^T \\ -(\partial h_t(\theta)/\partial\theta)^T \end{pmatrix} : (2 \times k). \tag{2.46}$$

We now have a QL score which is optimum within L

$$\begin{aligned} QL(\theta) = \sum_{t=1}^{n} w_{t-1}^{O}(\theta)a_t(\theta) &= \sum_{t=1}^{n} \big(E_{t-1}\big[\partial a_t(\theta)/\partial\theta^T\big]\big)^T \big(E_{t-1}\big[a_t(\theta)a_t^T(\theta)\big]\big)^{-1} a_t(\theta). \\ &= -\sum_{t=1}^{n} \left(\frac{\partial m_t(\theta)}{\partial\theta}, \frac{\partial h_t(\theta)}{\partial\theta}\right) V_t^{-1}(\theta)a_t(\theta) \end{aligned} \tag{2.47}$$

where $V_t(\theta)$ is given in (2.45). It then follows from (2.28) and Theorem 3.3 that

$$\sqrt{n}\big(\hat{\theta}_{QL} - \theta\big) \xrightarrow{d} N\big(0, K^{-1}\big) \tag{2.48}$$

where

$$K = E\left[\left(\frac{\partial m_t(\theta)}{\partial\theta}, \frac{\partial h_t(\theta)}{\partial\theta}\right) V_t^{-1}(\theta) \left(\frac{\partial m_t(\theta)}{\partial\theta}, \frac{\partial h_t(\theta)}{\partial\theta}\right)^T\right]. \tag{2.49}$$

Refer to, for instance, Hwang and Basawa (2011b) for further details. A weighted least squares seems to be simple to use. A weighted least squares estimator $\hat{\theta}_{WL}$ of θ is obtained by minimizing $\sum_{t=1}^{n}\left(\frac{X_t - m_t(\theta)}{\sqrt{h_t(\theta)}}\right)^2$. As an illustration, we consider $m_t(\theta) = \theta_1 X_{t(1)}^{+} + \theta_2 X_{t(1)}^{-}$ and $h_t(\theta) = \alpha_0 + \alpha_1 X_{t(1)}$, that is, we examine a simple heteroscedastic threshold BAR process. See Hwang and Kang (2012) for details on this model. Suppose that $\theta = (\theta_1, \theta_2)$ is the parameter of interest and the secondary parameter $\alpha = (\alpha_0, \alpha_1)$ is known so that $h_t(\theta)$ is free from the parameter θ of interest. The WL- score is given by

$$U_n(\theta) = \begin{pmatrix} \sum (X_t - m_t(\theta))X_{t(1)}^{+}/h_t \\ \sum (X_t - m_t(\theta))X_{t(1)}^{-}/h_t \end{pmatrix}, \text{with } m_t(\theta) = \theta_1 X_{t(1)}^{+} + \theta_2 X_{t(1)}^{-} \tag{2.50}$$

which in turn gives

$$\hat{\theta}_{WL} = \left[diag\left(\sum h_t^{-1} \left(X_{t(1)}^{+} \right)^2, \sum h_t^{-1} \left(X_{t(1)}^{-} \right)^2 \right) \right]^{-1} \begin{pmatrix} \sum X_t X_{t(1)}^{+} / h_t \\ \sum X_t X_{t(1)}^{-} / h_t \end{pmatrix} \tag{2.51}$$

where $diag(\cdot, \cdot)$ represent a diagonal matrix. The limit distribution of $\hat{\theta}_{WL}$ can be obtained via Theorem 3.2 (cf. Hwang and Kang (2012)). It is noted that the WL-score (2.50) is a member of Godambe MEFs L generated by the innovation vector $a_t(\theta)$ in (2.44) and thus $\hat{\theta}_{QL}$ provides the "smaller" variance than $\hat{\theta}_{WL}$, due to Theorem 3.3.

2.4 Non-Ergodic Martingale Estimating Functions

As discussed in Sect. 2.3, for ergodic stationary processes, a constant norm (e.g., $\sqrt{n}$) is used to get asymptotic normal distributions of the various estimators obtained from MEFs. On the other hand, for nonergodic type processes, limit distributions of standard estimators are mixed-normal when a nonrandom norm is used. Instead, a random norm is required to get normal limit distributions for nonergodic processes. Asymptotics of various statistics normalized by random norms in a broad context have recently been discussed by Pena et al. (2009) and Hwang et al. (2013a). Refer to Basawa and Scott (1983) for various nonergodic processes including normal mixture models, explosive autoregressive processes and branching processes. In this section we consider large sample estimation based on MEFs for a class of nonergodic processes. Via establishing a convolution theorem using a random norm, it will be shown that the ML estimator continues to be asymptotically optimum in a sense of "minimum variance" within a class of estimators obtained from nonergodic MEFs. Most of the contents in this section are adapted from those in Hwang et al. (2013a) and Hwang and Basawa (2011a) and therefore we provide streamlined outlines only, omitting some details.

Consider the following MEF $U_n(\theta)$ arising from possibly nonstationary process

$$U_n(\theta) = \sum_{t=1}^{n} u_t(\theta) : (k \times 1) \text{ vector} \tag{2.52}$$

Let ξ_n denote the sum of conditional variance–covariance matrix, i.e.,

$$\xi_n = \sum_{t=1}^{n} Var\left(u_t(\theta) | F_{t-1} \right) = \sum_{t=1}^{n} E\left(u_t(\theta) u_t^T(\theta) | F_{t-1} \right) : (k \times k) \text{ matrix.} \tag{2.53}$$

It is assumed that $|\xi_n| \to \infty$ almost surely, as n tends to infinity. The local neighborhood $N_\delta(\theta)$ about θ defined earlier in (2.2) needs to be modified as

$$N_\delta(\theta) = \{\theta^*;\ |\xi_n^{1/2}(\theta^* - \theta)| < \delta\} \tag{2.54}$$

and in turn the supremum appearing in (C1) of the regular MEF is to be taken over the local neighborhood of (2.54). Now, we collect all the regular MEFs into Ψ. Consider (arbitrary) member $U_n(\theta) \in \Psi$ and assume that

(C2) There exists a nonsingular (non-random) matrix G such that

$$G = plim\big[-\xi_n^{-1/2}(\partial U_n(\theta)/\partial\theta^T)\xi_n^{-1/2}\big]. \tag{2.55}$$

where *plim* denotes "limit in probability".

In particular, for the ML score $S_n(\theta) = \sum l_t(\theta)$, G reduces to I_k, the identity matrix of order k. Specifically, define

$$\eta_n = \sum Var\big(l_t(\theta)|F_{t-1}\big) = \sum E\big(l_t(\theta)l_t^T(\theta)|F_{t-1}\big) = \sum E\big(-\partial l_t(\theta)/\partial\theta^T|F_{t-1}\big) \tag{2.56}$$

and note that

$$I_k = plim\left[-\eta_n^{-1/2}\big(\partial S_n(\theta)/\partial\theta^T\big)\eta_n^{-1/2}\right]. \tag{2.57}$$

Let $\hat{\theta}_n$ be a solution of the martingale estimating equation $U_n(\theta) = 0$. The limit distribution of $\hat{\theta}_n$ is identified in the following theorem.

Theorem 4.1 Under (C1) and (C2), we have

$$\xi_n^{1/2}\big(\hat{\theta}_n - \theta\big) \xrightarrow{d} N\big(0, G^{-1}G^{-T}\big) \tag{2.58}$$

where G is defined in (2.55). In addition, we conclude

$$\eta_n^{1/2}\big(\hat{\theta}_{ML} - \theta\big) \xrightarrow{d} N\big(0, I_k\big). \tag{2.59}$$

Note that the norming (random) matrices in (2.58) and (2.59) are different as given by $\xi_n^{1/2}$ and $\eta_n^{1/2}$, respectively.

(C3) There exists a nonrandom and nonsingular matrix C which is the limiting $(k \times k)$ covariance matrix between $\xi_n^{-1/2}U_n(\theta)$ and $\eta_n^{-1/2}S_n(\theta)$, viz.,

$$C = plim\big[\xi_n^{-1/2}U_n(\theta)\eta_n^{-1/2}S_n(\theta)^T\big]. \tag{2.60}$$

We now define the "ratio" matrix Γ between G and C.

$$\Gamma = C^{-1}G \tag{2.61}$$

It can then be verified that the ML estimate using the norming matrix $\xi_n^{1/2}$ has the limiting variance–covariance matrix given by $\Gamma^{-1}\Gamma^{-T}$, viz.,

$$\xi_n^{1/2}\big(\hat{\theta}_{ML} - \theta\big) \xrightarrow{d} N\big(0, \Gamma^{-1}\Gamma^{-T}\big). \tag{2.62}$$

Refer to Hwang et al. (2013a) for details. It will be shown that $G^{-1}G^{-T} - \Gamma^{-1}\Gamma^{-T}$ is non-negative definite and therefore we deduce that $\hat{\theta}_{ML}$ is optimal within the regular

class Ψ of MEFs in the sense of having the "minimum" variance–covariance matrix. A convolution theorem for non-ergodic MEFs (due to Hwang et al. (2013a)) is now presented. Decompose

$$\xi_n^{1/2}\big(\hat{\theta}_n - \theta\big) = Y_{1n}(\theta) + Y_{2n}(\theta)$$

where

$$Y_{1n}(\theta) = \xi_n^{1/2}\big(\hat{\theta}_n - \theta\big) - \Gamma^{-1}\eta_n^{-1/2}S_n(\theta) \tag{2.63}$$

and

$$Y_{2n}(\theta) = \Gamma^{-1}\eta_n^{-1/2}S_n(\theta). \tag{2.64}$$

Theorem 4.2 (A convolution theorem for non-ergodic MEFs) Under some regularity conditions, for any $\hat{\theta}_n$ obtained from $U_n(\theta) \in \Psi$, $\xi_n^{1/2}\big(\hat{\theta}_n - \theta\big)$ can be expressed as a sum of two asymptotically independent components which are distributed as $N\big(0, G^{-1}G^{-T} - \Gamma^{-1}\Gamma^{-T}\big)$ and $N\big(0, \Gamma^{-1}\Gamma^{-T}\big)$, respectively. Specifically,

$$\xi_n^{1/2}\big(\hat{\theta}_n - \theta\big) = Y_{1n}(\theta) + Y_{2n}(\theta)$$

$$\begin{pmatrix} Y_{1n}(\theta) \\ Y_{2n}(\theta) \end{pmatrix} \xrightarrow{d} N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} G^{-1}G^{-T} - \Gamma^{-1}\Gamma^{-T} & 0 \\ 0 & \Gamma^{-1}\Gamma^{-T} \end{pmatrix}\right)$$

where $Y_{1n}(\theta)$ and $Y_{2n}(\theta)$ are defined in (2.63) and (2.64).

To illustrate Theorems 4.1 and 4.2, two nonergodic processes are discussed.

2.4.1 *Branching Markov processes (BMP)*

A BMP is a tree-indexed process where the tree index is a branching process $\{Z_t, t = 0, 1, 2, \ldots\}$ with Z_t denoting the $t-$th generation size. BMPs were investigated by Hwang and Basawa (2009, 2011a). Let $X_t(j), j = 1, 2, \ldots, Z_t$ and $t = 0, 1, 2, \ldots$, denote observation on the $j-$th individual in the $t-$th generation. Figure 2.1 illustrates a sample path of BMP (see Hwang and Basawa (2009)). We assume that $\{Z_t\}$ follows a standard supercritical Galton-Watson (G-W) branching process for which $E(Z_1) = m > 1$ and $Var(Z_1) = \sigma^2 > 0$ where m and σ^2 are the offspring mean and variance respectively. It is well known that there exists a random variable W to which Z_n/m^n converges almost surely as $n \to \infty$, and $P(W > 0) = 1$. To clarify the ancestral path of $X_t(j)$, use the notation $X_{t-1}(t(j))$ to denote the observation on the immediate mother of the $X_t(j)$. Here, the subscript $t-1$ is used for denoting $(t-1)$th generation. We refer to Hwang and Basawa (2009, 2011a) for various examples of BMP. A simple BMP is a branching AR model defined by

$$X_t(j) = \theta_0 + \theta_1 X_{t-1}(t(j)) + \epsilon_t(j) \tag{2.65}$$

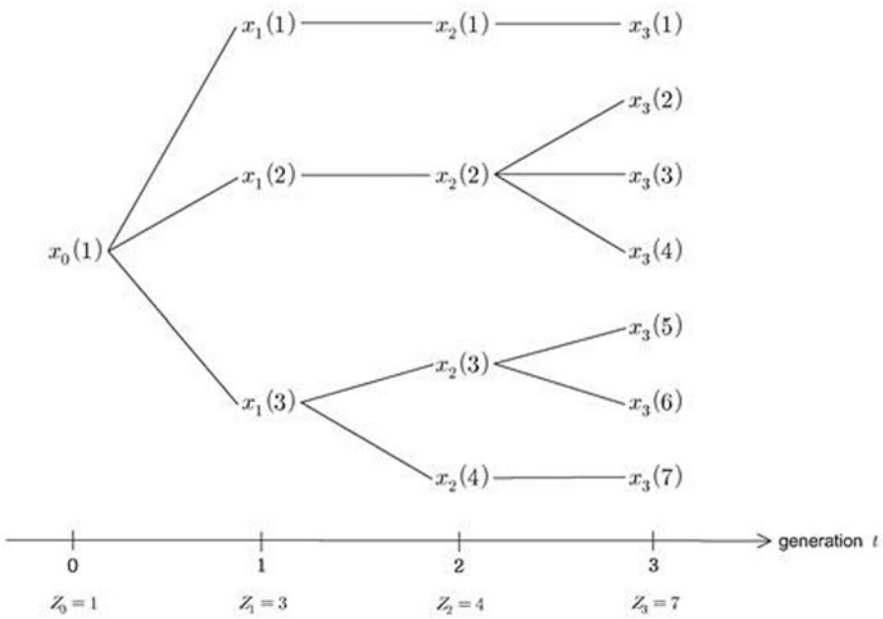


Fig. 2.1 A path of BMP model

where $\{\epsilon_t(j), t = 1, 2, \ldots$ and $j = 1, 2, \ldots\}$ are iid random variables with mean zero and variance σ_ϵ^2. The data is given as follows.

$$\{(z_t, x_t(j)); t = 1, 2, \ldots, n;\ j = 1, 2, \ldots, z_n\}$$

with initial observation $x_0(1)$ on $Z_0 = 1$.

Set $\theta = \big(\theta_0, \theta_1\big)^T$. If $\epsilon_t(j)$ is normal, the ML estimate $\hat{\theta}_{ML}$ is given by

$$\hat{\theta}_{ML} = \begin{pmatrix} \sum Z_t & \sum\sum X_{t-1}(t(j)) \\ \sum\sum X_{t-1}(t(j)) & \sum\sum X_{t-1}^2(t(j)) \end{pmatrix}^{-1} \begin{pmatrix} \sum\sum X_t(t(j)) \\ \sum\sum X_t(t(j)) X_{t-1}(t(j)) \end{pmatrix} \tag{2.66}$$

where $\sum = \sum_{t=1}^n$ and $\sum\sum = \sum_{t=1}^n \sum_{j=1}^{Z_t}$. We now define (2×2) non-random matrix

$$\eta = plim \begin{pmatrix} \sum Z_t & \sum\sum X_{t-1}(t(j)) \\ \sum\sum X_{t-1}(t(j)) & \sum\sum X_{t-1}^2(t(j)) \end{pmatrix} / \sum Z_t. \tag{2.67}$$

It can be verified that (2.59) is valid, viz.,

$$\eta_n^{1/2}\big(\hat{\theta}_{ML} - \theta\big) \xrightarrow{d} N\big(0, I_2\big) \tag{2.68}$$

if we choose the random norm η_n as $\eta_n = \sigma_\epsilon^{-2}\eta \sum Z_t$. It is interesting to note that $\hat{\theta}_{ML}$ is mixed-normal with the mixing random variable W when a non-random norm is used. Specifically, set

$$\delta_n = \frac{m^{n+1}}{m-1}, \ m > 1$$

where m is the offspring mean. Then, we have

$$\delta_n^{1/2}\big(\hat{\theta}_{ML} - \theta\big) \xrightarrow{d} \frac{1}{\sqrt{W}} \cdot N\big(0, \sigma_\epsilon^2 \eta^{-1}\big). \tag{2.69}$$

Note that $\hat{\theta}_{ML}$ is asymptotically optimal within Ψ in the sense of Theorem 4.2. We refer to Hwang and Basawa (2011a) and Hwang et al. (2013a) for asymptotic mixed normality arising from nonergodic MEFs.

2.4.2 Explosive AR(1) Processes

Consider the following zero mean explosive AR process defined by

$$X_t = \theta X_{t-1} + \epsilon_t, \ |\theta| > 1 \tag{2.70}$$

where $\{\epsilon_t\}$ is iid $N(0, \sigma_\epsilon^2)$ errors and the initial value $X_0 = 0$. The ML score function $S_n(\theta)$ is seen to be

$$S_n(\theta) = \sum_{t=1}^{n} l_t(\theta) \text{ with } l_t(\theta) = \sigma_\epsilon^{-2}\epsilon_t(\theta)X_{t-1}$$

where $\epsilon_t(\theta) = X_t - \theta X_{t-1}$ and it is obvious that $\hat{\theta}_{ML} = \sum X_t X_{t-1} / \sum X_{t-1}^2$ and the corresponding random norm is given by $\eta_n = \sigma_\epsilon^{-2} \sum X_{t-1}^2$. We conclude via (2.59)

$$\sigma_\epsilon^{-1}\sqrt{\sum X_{t-1}^2}\big(\hat{\theta}_{ML} - \theta\big) \xrightarrow{d} N(0,1). \tag{2.71}$$

We consider a case of mis-specification of the conditional variance h_t. Suppose that ϵ_t in (2.70) is misspecified as a ARCH(1) process. That is, we have

$$\epsilon_t = \sqrt{h_t} e_t$$

where $\{e_t\}$ is iid with mean zero and variance unity and

$$h_t = \alpha_0 + \alpha_1 \epsilon_{t-1}^2 \tag{2.72}$$

for which $\alpha_0 > 0$ and $0 \le \alpha_1 < 1$. Assume that α_0 and α_1 are known constants. Consider the following misspecified MEF $U_n(\theta)$ defined by

$$U_n(\theta) = \sum_{t=1}^{n} h_t^{-1}\epsilon_t(\theta)X_{t-1} \tag{2.73}$$

which in turn gives

$$\hat{\theta}_n = \sum X_t X_{t-1} h_t^{-1} / \sum X_{t-1}^2 h_t^{-1}. \tag{2.74}$$

To discuss asymptotics for $\hat{\theta}_n$, note that $\xi_n = \sum h_t^{-2} \sigma_\epsilon^2 X_{t-1}^2$. It can be shown that (2.58) holds, i.e.,

$$\xi_n^{1/2}(\hat{\theta}_n - \theta) \xrightarrow{d} N(0, G^{-2}) \tag{2.75}$$

where

$$G = plim \left[\sigma_\epsilon^{-2} \sum h_t^{-1} X_{t-1}^2 / \sum h_t^{-2} X_{t-1}^2 \right] = \sigma_\epsilon^{-2} \frac{E\left(\alpha_0 + \alpha_1 \epsilon_{t-1}^2\right)^{-1}}{E\left(\alpha_0 + \alpha_1 \epsilon_{t-1}^2\right)^{-2}}. \tag{2.76}$$

2.5 Concluding Remarks

This review paper presents asymptotic results on MEFs in stochastic processes. Standard estimation methods such as LS, QL, ML and PL can be unified via a single framework of MEFs. When the likelihood is known, ML score is shown to provide the "smallest" variance among the class U of regular MEFs. It is often the case in stochastic processes that the likelihood is unknown but only first few moment structures are given instead. The QL score is then verified to be asymptotically optimal within the restricted class $L \subset U$ of Godambe MEFs. Two convolution theorems are established to address optimality of ML score and QL score separately within appropriate classes of MEFs.

Both ergodic and non-ergodic cases are discussed. Applications to conditionally linear AR (CLAR) models, GARCH-type processes and bifurcating AR (BAR) models are presented to illustrate the ergodic case. A non-ergodic convolution theorem is established and in turn (BMP) and explosive AR models are discussed for non-ergodic applications. The results presented in the paper are mostly adapted from recent literature on MEF asymptotics as a unifying tool for estimation in stochastic processes.

We have not discussed testing problems in MEF asymptotics. When the likelihood is available, one may use the classical three tests (Rao's score, Wald, and LR statistics). If the likelihood is unknown, we look to appropriate MEFs, for instance, a QL score in constructing test statistics. Basawa (1991), Hwang and Basawa (2011b), and Hwang et al. (2013a) obtained some preliminary results on certain tests based on MEFs. However, a rigorous treatment on asymptotic power and efficiency of tests based on MEFs in a broad context has not yet been adequately addressed in the literature and this will be pursued elsewhere.

Acknowledgements We like to take this opportunity to acknowledge and celebrate Prof. Hira Koul's outstanding achievements in fundamental research and his service to the statistical community over several decades. He has inspired numerous researchers around the world and helped

generations of graduate students who themselves have become leaders in statistical research. We congratulate Hira for his life long achievements and contributions to the field of mathematical statistics. We thank the reviewer for careful reading of the paper. This work was supported by a grant from the National Research Foundation of Korea (NRF-2012012872).

References

Basawa IV (1983) Recent trends in asymptotic optimal inference for dependent observations. Austral Jour Statist 25:182–190

Basawa IV (1991) Generalized score tests for composite hypotheses. In V.P. Godambe (ed) Estimating functions. Oxford University Press, pp 121–132

Basawa IV (2001) Inference in stochastic processes. In: C.R. Rao, D.N. Shanbhag (eds) Handbook of statistics, Vol. 19. North Holland, pp 55–77

Basawa IV, Koul HL (1988) Large-sample statistics based on quadratic dispersion. Inter Statist Review 56:199–219

Basawa IV, Prakasa Rao BLS (1980a) Statistical inference for stochastic processes. Academic Press, London

Basawa IV, Prakasa Rao BLS (1980b) Asymptotic inference for stochastic processes. Stochastic Proc their Appl 10:221–254

Basawa IV, Scott DJ (1983) Asymptotic optimal inference for non-ergodic models. Springer, New York

Basawa IV, Feigin PD, Heyde CC (1976) Asymptotic properties of maximum likelihood estimators for stochastic processes. Sankhya Series A 38:259–270

Basawa IV, Godambe VP, Taylor RL (1997) Selected proceedings of the symposium on estimating equations. Lecture Notes, Vol. 32. IMS, Hayward, California

Bibby BM, Sorensen M (1995) Martingale estimating functions for discretely observed diffusion processes. Bernoulli 1:17–39

Choi MS, Park JA, Hwang SY (2012) Asymmetric GARCH processes featuring both threshold effect and bilinear structure. Stat Probabil Lett 82:419–426

Cowan R, Staudte RG (1986) The bifurcating autoregression model in cell lineage studies. Biometrics 42:769–783

Godambe VP (1985) The foundation of finite sample estimation in stochastic processes. Biometrika 72:419–428

Gourieroux C (1997) ARCH Models and Financial Applications. Springer, New York

Heyde CC (1997) Quasi-likelihood and Its Applications. Springer, New York

Hwang SY, Basawa IV (1993) Asymptotic optimal inference for a class of nonlinear time series models. Stochastic Proc their Appl 46:91–113

Hwang SY, Basawa IV (2009) Branching Markov processes and related asymptotics. J Multivariat Anal 100:1155–1167

Hwang SY, Basawa IV (2011a) Asymptotic optimal inference for multivariate branching-Markov processes via martingale estimating functions and mixed normality. Journal of Multivariate Analysis 102:1018–1031

Hwang SY, Basawa IV (2011b) Godambe estimating functions and asymptotic optimal inference. Statistics & Probability Letters 81:1121–1127

Hwang, S.Y. and Kang, Kee-Hoon (2012) Asymptotics for a class of generalized multicast autoregressive process. J Korean Statist Soc 41:543–554

Hwang SY, Basawa IV, Choi MS, Lee SD (2013a) Non-ergodic martingale estimating functions and related asymptotics. *Statistics*, online published, 1-21, doi: 10.1080/02331888.2012.748772

Hwang SY, Choi MS, Yeo IK (2013b) Quasilikelihood and quasi-maximum likelihood for GARCH-type processes: estimating function approach. under revision in Journal of the Korean Statistical Society

Klimko LA, Nelson PI (1979) On conditional least squares estimation for stochastic processes. Ann Statist 6:629–642

Pena VH, Lai TL, Shao Q-M (2009) Self-normalized processes: limit theory and statistical applications. Springer, Berlin

Tsay RS (2010) Analysis of financial time series, 3rd Ed. Wiley, New York

Wefelmeyer W (1996) Quasilikelihood models and optimal inference. Ann Statist 24:405–422

Chapter 3
Asymptotics of L_λ-Norms of ARCH(p) Innovation Density Estimators

Fuxia Cheng

3.1 Introduction

Let $X_{1-p}, \ldots, X_0, X_1, \ldots$ be random variables for some positive integer p. We assume they form an ARCH(p)-model:

$$X_i = \varepsilon_i \sqrt{\alpha_0 + \alpha_1 X_{i-1}^2 + \cdots + \alpha_p X_{i-p}^2}, \quad i = 1, 2, \ldots, \tag{3.1}$$

where the parameters $\alpha_0, \ldots, \alpha_p$ are positive and the innovations ε_i are independent and identically distributed random variables with mean 0, variance 1, unknown density function f and distribution function F and are independent of $X_{1-p}, \ldots, X_{i-1}$. It follows that the conditional variance of X_i satisfies

$$Var\{X_i | X_{1-p}, \ldots, X_{i-1}\} = \alpha_0 + \alpha_1 X_{i-1}^2 + \cdots + \alpha_p X_{i-p}^2, \quad i = 1, 2, \ldots \tag{3.2}$$

Property (3.2) is called conditional heteroscedasticity and explains, together with its autoregressive nature, the name of this model.

Model (3.1) has found much interest in financial econometrics. It was introduced by Engle (1982) in order to provide a framework in which so-called volatility clusters may occur, i.e., periods of high and low (conditional) variances depending on past values of the series. The model was later extended into various directions. See Gouriérous (1997) for details. In most of the work, the main focus has been on estimating the unknown parameters $\alpha_0, \ldots, \alpha_p$, see Weiss (1986), Horváth and Liese (2004) among others.

It is of interest and of practical importance to know the nature of the innovation distribution. Actually, if the distribution of the innovation is unspecified, the parametric component only partly determines the distribution behavior of (3.1). It is as important to investigate the distribution of the innovation as estimating α_j's. Stute (2001) uses the residual-based empirical distribution function (d.f.). F_n to estimate the distribution function of ε_i, and provides consistency and distributional

F. Cheng (✉)
Department of Mathematics, Illinois State University, Normal, IL 61790, USA
e-mail: fcheng@ilstu.edu

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_3,

convergence results for a class of statistics based on F_n. Cheng (2008a) considers the uniform strong consistency of the innovation distribution function estimation in autoregressive conditional heteroskedasticity (ARCH)(p)-time series, and obtains the extended Glivenko-Cantelli Theorem for the residual-based empirical d.f.. Cheng (2008b) develops the asymptotic distribution of the innovation density estimator at a fixed point and globally. Cheng and Wen (2011) obtain the strong consistency of the innovation density estimator under L_1-norm. Cheng, Sun, and Wen (2011) develop the asymptotic normality of the Bickel-Rosenblatt test statistic and show the strong consistency of the estimator for the true density in L_2-norm.

For generalized ARCH (GARCH) models, Koul and Mimoto (2012) prove asymptotic normality of a suitably standardized integrated square difference between a kernel type error density estimator based on residuals and the expected value of the error density estimator based on innovations of GARCH models.

Here, we will continue to develop the global property of the innovation density estimator in ARCH(p). Notice that central limit theorems for L_p-norms of density estimators (under independent and identically distributed (i.i.d.) set up) have been obtained in Csörgő and Horváth (1988); and the corresponding results have been derived for the L_p-norms of error density estimators in the first-order autoregressive models by Horváth and Zitikis (2004). For an autoregressive of order $p \geq 1$ (AR(p)) model, Yang, Fu, and Zhang (2011) have compared the kernel density estimator (based on residuals) with the theoretical kernel density estimator based on unobserved innovations, and they show that the L_r -norm of the difference is asymptotically negligible.

In this paper, we will consider asymptotic properties of residual-based kernel density estimators of the innovation density f in $L_\lambda(\lambda \geq 1)$-norms. The asymptotic result for L_λ-norms of density estimators (under i.i.d. set up) will be extended to L_λ-norms for the residual-based kernel density estimators in ARCH(p) time series. Our main result gives a rate for the L_λ-norm of the difference between the residual-based and the innovation-based kernel density estimators. This rate is faster than that for the L_λ-norm of the difference between innovation-based kernel density estimator and innovation density for the case $\lambda > 1$, and of the same order for the case $\lambda = 1$. Thus the known asymptotic behavior for the L_λ-norm of the latter difference carries over to that of the difference between residual-based kernel density estimator and innovation density for the case $\lambda > 1$, but not for the case $\lambda = 1$.

The paper is organized as follows. In Sect. 2 we introduce the residual-based kernel density estimator and state some basic assumptions. Section 3 presents the main result. Detailed proofs are provided in Sects. 4 and 5.

3.2 Estimators and Some Basic Assumptions

Assume that we observe $X_{1-p}, X_{2-p}, \cdots, X_n$ which obey the model (3.1). Let $\hat{\alpha}_n = (\hat{\alpha}_{0n}, \ldots, \hat{\alpha}_{pn})^\top$ denote an estimator of the parameter vector $\alpha = (\alpha_0, \ldots, \alpha_p)^\top$, based on these observations. Set

$$\hat{\varepsilon}_i = X_i \Big/ \sqrt{\hat{\alpha}_{0n} + \hat{\alpha}_{1n} X_{i-1}^2 + \cdots + \hat{\alpha}_{pn} X_{i-p}^2}, \quad 1 \leq i \leq n,$$

to be residuals. Using these residuals, we construct an estimator of the innovation density f as follows:

$$\hat{f}_n(t) := \frac{1}{n}\sum_{i=1}^{n} K_{h_n}(t - \hat{\varepsilon}_i), \quad t \in \mathbb{R},$$

with $K_{h_n}(t) = K(t/h_n)/h_n$ and h_n being positive numbers (usually called bandwidth) tending to zero as $n \to \infty$, and K is the kernel density function.

Define the kernel innovation density based on the true innovations (which we cannot observe) $\varepsilon_1, \varepsilon_2, \cdots, \varepsilon_n$:

$$f_n(t) := \frac{1}{n}\sum_{i=1}^{n} K_{h_n}(t - \varepsilon_i), \quad t \in \mathbb{R}.$$

The strong consistency of f_n for f under L_1-norm is given in Devroye (1983), i.e.,

$$\int |f_n(t) - f(t)|\, dt \to 0 \quad \text{almost surely (a.s.), as } n \to \infty.$$

In Cheng and Wen (2011), the above result is extended to $\hat{f}_n$, i.e.,

$$\int |\hat{f}_n(t) - f(t)| dt \to 0 \quad \text{a.s., as } n \to \infty.$$

For the integrated squared deviation of $\hat{f}_n$ from

$$E(f_n(t)) = \int K(x) f(t - h_n x) dx = K_{h_n} * f(t), \quad t \in \mathbb{R},$$

(where $K_{h_n} * f$ denotes the convolution of the two functions K_{h_n} and f) defined as

$$\int [\hat{f}_n(t) - K_{h_n} * f(t)]^2\, dt,$$

Cheng, Sun, and Wen (2011) develop its asymptotic normality which is the same as the one of the Bickel–Rosenblatt test statistic based on f_n in Bickel and Rosenblatt (1973). They also show that $\int [\hat{f}_n(t) - f(t)]^2 dt$ tends to zero almost surely.

For any (finite number) $\lambda \geq 1$, the L_λ-norm of a measurable function g is defined as follows:

$$\|g\|_\lambda := \left(\int |g(x)|^\lambda\, dx \right)^{1/\lambda}.$$

We mention that the convolution $q * r$ of an integrable function r with a function q of finite L_λ-norm has finite L_λ-norm and obeys the inequality

$$\|q * r\|_\lambda \leq \|q\|_\lambda \|r\|_1.$$

We should also point out that, for such a q, the map

$$s \mapsto \|q\,(\cdot - s) - q\|_\lambda$$

is bounded by $2\|g\|_\lambda$ and is uniformly continuous; see, e.g. Theorem 9.5 in Rudin (1974) for the later. The above inequality is a special case of the more general

inequality

$$\|V|q * r|^\lambda\|_1 \le \|V|q|^\lambda\|_1 \|Vr\|_1^\lambda$$

with $V(x) = (1 + |x|)^\beta$ for some $\beta \ge 0$. This follows from the inequalities

$$|q * r|^\lambda \le \|r\|_1^{\lambda-1} |q|^\lambda * |r| \le \|Vr\|_1^{\lambda-1} |q|^\lambda * |r|$$

and

$$\|Vu * r\|_1 \le \|Vu\|_1 \|Vr\|_1.$$

The former is a consequence of the Hölder inequality, while the latter is from Schick and Wefelmeyer (2007).

In this paper, we consider the asymptotic distribution of the L_λ-norm of the difference between the kernel innovation density estimators based on residuals and the true density function, i.e., $\|\hat{f}_n - f\|_\lambda$.

In order to show the main result, we need the following assumptions.

Assumption 1. The entries of α and $\hat{\alpha}_n$ are positive, and the estimator $\hat{\alpha}_n$ is root-n consistent: $n^{1/2}(\hat{\alpha}_n - \alpha) = O_p(1)$ as $n \to \infty$.

Assumption 2. The density f has mean zero and variance one and is absolutely continuous, and the function $x \mapsto (1 + x^2) f'(x)$ has finite L_1, L_2 and L_λ-norms.

Assumption 3. The kernel K is a three-times continuously differentiable symmetric density with compact support.

Remark 2.1 Let $\tilde{\alpha}_n = (\tilde{\alpha}_{0n}, \ldots, \tilde{\alpha}_{pn})^\top$ be a root-n consistent estimator of α. Then the estimator $\hat{\alpha}_n$ with entries $\hat{\alpha}_{jn} = \max(1/\mathrm{n}, \alpha_{\mathrm{jn}})$ meets the requirement of Assumption 1. A possible root-n consistent estimator is the least squares estimator.

For later use, we introduce functions ψ_1 and ψ_2 by

$$\psi_1(x) = xf(x) \quad \text{and} \quad \psi_2(x) = x^2 f(x), \quad x \in \mathbb{R}.$$

Remark 2.2 Assumption 2 implies the following. The density f is bounded. The function ψ_1 is absolutely continuous with almost everywhere derivative

$$\psi_1'(x) = f(x) + xf'(x), \quad x \in \mathbb{R},$$

which has finite L_1, L_2 and L_λ-norms. Thus ψ_1 is bounded. Similarly, the function ψ_2 is absolutely continuous with almost everywhere derivative

$$\psi_2'(x) = 2xf(x) + x^2 f'(x), \quad x \in \mathbb{R},$$

which has finite L_1, L_2 and L_λ-norms. Thus ψ_2 is bounded.

Remark 2.3 Assumption 3 guarantees that K and its first three derivatives are bounded and integrable. Hence these functions have finite L_λ-norms. So do K_{h_n} and its first three derivatives, and we have

$$\|K_{h_n}^{(\nu)}\|_\lambda = \frac{\|K^{(\nu)}\|_\lambda}{h_n^{\nu+1-1/\lambda}}, \quad \nu = 0, 1, 2, 3.$$

In the following sections, all limits are taken as the sample size n tends to ∞, unless specified otherwise.

3.3 Asymptotics of $\hat{f}_n$ Under L_λ-Norm

Throughout this section $\lambda \geq 1$ is a fixed finite number. We set

$$\lambda_* = \frac{\lambda - 1}{\lambda} = 1 - \frac{1}{\lambda}$$

and introduce the $p+1$-dimensional random vectors

$$W_i = (1, X_{i-1}^2, \ldots, X_{i-p}^2)^\top / [2(\alpha_0 + \alpha_1 X_{i-1}^2 + \cdots + \alpha_p X_{i-p}^2)], \quad i = 1, \ldots, n,$$

and their average

$$\bar{W}_n = \frac{1}{n} \sum_{i=1}^{n} W_i.$$

We are ready to state our main result.

Theorem 3.1 *Assume that Asumptions 1–3 hold, and the bandwidth h_n satisfies $h_n \to 0$ and $nh_n^{3+\lambda_*/2} \to \infty$. We also assume that*

1. for $1 \leq \lambda < 2$, $E(|\varepsilon_1|^3) < \infty$ and $\int (1+|x|)^{4\lambda+\beta} f(x)^\lambda \, dx < \infty$ for some $\beta > 1$.
2. for $\lambda \geq 2$, $E(|\varepsilon_1|^{2\lambda}) < \infty$.

Then we have

$$\sqrt{nh_n^{\lambda_*}} \| \hat{f}_n - f_n - (\hat{\alpha}_n - \alpha)^\top \bar{W}_n \psi_1'(x) \|_\lambda = o_p(1).$$

For $\lambda > 1$, this implies

$$\sqrt{nh_n^{\lambda_*}} \| \hat{f}_n - f_n \|_\lambda = o_p(1).$$

Remark 3.1 Let r_n denote the square root of $nh_n^{\lambda_*}$. The asymptotic distribution of $r_n \| f_n - f \|_\lambda$ has been developed in Csörgő and Horváth (1988). In fact, under some natural assumptions, for some positive constants σ and m,

$$(r_n \| f_n - f \|_\lambda)^\lambda - m/\sqrt{h_n} \longrightarrow N(0, \sigma^2).$$

Here we have shown that $r_n \| \hat{f}_n - f_n \|_\lambda = o_p(1)$ if $\lambda > 1$. Thus we can claim that (under appropriate conditions) $r_n \| \hat{f}_n - f_n \|_\lambda$ has the same asymptotic distribution as $r_n \| f_n - f_n \|_\lambda$ for such λ. For $\lambda = 1$, however, the asymptotic distributions of $r_n \| \hat{f}_n - f_n \|_\lambda$ and $r_n \| f_n - f_n \|_\lambda$ will differ.

Let us set

$$\varepsilon_i^* = \varepsilon_i - \varepsilon_i (\hat{\alpha}_n - \alpha)^\top W_i$$

and

$$\hat{f}_n^*(t) = \frac{1}{n} \sum_{i=1}^{n} K_{h_n}(t - \varepsilon_i^*), \quad t \in \mathbb{R}.$$

Then the theorem is a simple consequence of the Minkowski inequality and the following two lemmas.

Lemma 3.1 *Assume that ε_1 has finite mean, K' has finite L_λ-norm, Assumption 1 holds and $nh_n^{2+\lambda_*} \to \infty$. Then we have*

$$\sqrt{nh_n^{\lambda_*}} \| \hat{f}_n - \hat{f}_n^* \|_\lambda = o_p(1).$$

Lemma 3.2 *Under the assumptions of Theorem 3.1, we have*

$$\sqrt{nh_n^{\lambda_*}} \| \hat{f}_n^* - f_n - \left(\hat{\alpha}_n - \alpha\right)^\top \bar{W}_n \psi_1 \|_\lambda = o_p(1).$$

These lemmas are proved in the next sections.

3.4 Proof of Lemma 3.1

For $a = (a_0, \ldots, a_p)^\top \in \mathbb{R}^{p+1}$ and $i = 1, \ldots, n$, set

$$v_i(a) = a_0 + a_1 X_{i-1}^2 + \cdots + a_p X_{i-p}^2$$

and introduce

$$g_i(s) = \frac{X_i}{\sqrt{v_i(\alpha + s\Delta)}}, \quad 0 \le s \le 1,$$

with $\Delta = \hat{\alpha} - \alpha = (\hat{\alpha}_{0n} - \alpha_0, \cdots, \hat{\alpha}_{pn} - \alpha_p)^\top$. Then we have

$$\varepsilon_i = g_i(0) \quad \text{and} \quad \hat{\varepsilon}_i = g_i(1).$$

Since g_i is twice continuously differentiable, the identity

$$\hat{\varepsilon}_i = \varepsilon_i + g_i'(0) + \int_0^1 (1-s) g_i''(s) ds$$

holds. We calculate

$$g_i'(0) = -X_i \frac{v_i(\Delta)}{v_i^{3/2}(\alpha)} = -\varepsilon_i \frac{v_i(\Delta)}{2v_i(\alpha)} = -\varepsilon_i \Delta^\top W_i,$$

and

$$g_i''(s) = X_i \frac{3v_i^2(\Delta)}{4v_i^{5/2}(\alpha + s\Delta)} = \varepsilon_i \frac{3v_i^{1/2}(\alpha) v_i^2(\Delta)}{4v_i^{5/2}(\alpha + s\Delta)}.$$

It is easy to check that

$$\sup_{1 \le s \le 1} |g_i''(s)| \le |\varepsilon_i| T_n$$

with

$$T_n = \frac{3}{4} \left(\sum_{j=0}^p \frac{\alpha_j}{\min(\alpha_j, \hat{\alpha}_{jn})} \right)^{1/2} \left(\sum_{j=0}^p \frac{|\Delta_j|}{\min(\alpha_j, \hat{\alpha}_{jn})} \right)^2.$$

It follows from Assumption 1 that

$$T_n = O_p(n^{-1}).$$

In view of the identity

$$\hat{f}_n(t) - \hat{f}_n^*(t) = -\frac{1}{n}\sum_{i=1}^{n}\int_0^1 (\hat{\varepsilon}_i - \varepsilon_i^*)K_{h_n}^{'}(t - \varepsilon_i^* - u((\hat{\varepsilon}_i - \varepsilon_i^*))du,$$

the Minkowski inequality, and the Jensen inequality, we obtain the bound

$$\|\hat{f}_n - \hat{f}_n^*\|_\lambda \le \frac{1}{n}\sum_{i=1}^{n}|\hat{\varepsilon}_i - \varepsilon_i^*|\Big(\int\int_0^1 |K_{h_n}^{'}(t - \varepsilon_i^* - u((\hat{\varepsilon}_i - \varepsilon_i^*))|^\lambda\, du\, dt\Big)^{1/\lambda}.$$

Fubini's theorem and the substitution $x = t - \varepsilon_i^* - u(\hat{\varepsilon}_i - \varepsilon_i^*)$ now yield the inequality

$$\|\hat{f}_n - \hat{f}_n^*\|_\lambda \le \frac{1}{n}\sum_{i=1}^{n}|\hat{\varepsilon}_i - \varepsilon_i^*|\,\|K_{h_n}^{'}\|_\lambda.$$

In view of the identity $\|K_{h_n}^{'}\|_\lambda = \|K_{h_n}^{'}\|_\lambda / h_n^{1+\lambda_*}$ (see Remark 2.3) and the inequality

$$|\hat{\varepsilon}_i - \varepsilon_i^*| = \Big|\int_0^1 (1-s)g_i^{''}(s)\,ds\Big| \le |\varepsilon_i|T_n,$$

one derives the bound

$$\|\hat{f}_n - \hat{f}_n^*\|_\lambda \le \|K_{h_n}^{'}\|_\lambda \frac{T_n\sum_{i=1}^{n}|\varepsilon_i|}{nh_n^{1+\lambda_*}}.$$

Since ε_1 has finite mean, one has $\sum_{i=1}^{n}|\varepsilon_i| = O_p(n)$. Using this, $T_n = O_p(1/n)$, and $nh_n^{2+\lambda_*} \to \infty$, one obtains the rate

$$\sqrt{nh_n^{\lambda_*}}\|\hat{f}_n - \hat{f}_n^*\|_\lambda = O_p\left(\frac{\sqrt{nh_n^{\lambda_*}}}{nh_n^{1+\lambda_*}}\right) = O_p\left(\frac{1}{\sqrt{nh_n^{2+\lambda_*}}}\right) = o_p(1).$$

This is the desired result.

3.5 Proof of Lemma 3.2

We use the notation of the previous proof. It is easy to see that the j-th coordinate W_{ij} of the random vector W_i is bounded by $1/(2\alpha_j)$. From this we conclude that $|\Delta^\top W_i| \le S_n$ where

$$S_n = \sum_{j=0}^{p}\frac{|\Delta_j|}{2\alpha_j} = O_p(n^{-1/2})$$

Note that

$$\hat{f}_n^*(t) = \frac{1}{n}\sum_{i=1}^{n}K_{h_n}(t - \varepsilon_i + \varepsilon_i\Delta^\top W_i).$$

A Taylor expansion yields

$$\hat{f}_n^*(t) - f_n(t) = \frac{1}{n}\sum_{i=1}^{n}\left[\Delta^\top W_i \varepsilon_i K_{h_n}^{'}(t-\varepsilon_i) + \frac{1}{2}(\Delta^\top W_i)^2 \varepsilon_i^2 K_{h_n}^{''}(t-\varepsilon_i)\right] + R_n(t)$$

with

$$R_n(t) = \frac{1}{6n}\sum_{i=1}^{n}(\Delta^\top W_i)^3 \varepsilon_i^3 \int_0^1 K_{h_n}^{'''}(t-\varepsilon_i + s\Delta^\top W_i \varepsilon_i) 3(1-s)^2\, ds.$$

For $t \in \mathbb{R}$ and $j,k = 0,\ldots,p$, we set

$$A_j(t) = \frac{1}{n}\sum_{i=1}^{n} W_{ij}[\varepsilon_i K_{h_n}^{'}(t-\varepsilon_i) - \mu_{1n}(t)],$$

$$B_{jk}(t) = \frac{1}{n}\sum_{i=1}^{n} W_{ij} W_{ik}[\varepsilon_i^2 K_{h_n}^{''}(t-\varepsilon_i) - \mu_{2n}(t)],$$

with

$$\mu_{1n}(t) = E[\varepsilon_1 K_{h_n}^{'}(t-\varepsilon_1)]$$

and

$$\mu_{2n}(t) = E[\varepsilon_1^2 K_{h_n}^{''}(t-\varepsilon_1)].$$

Then we can rewrite the difference $\hat{f}_n^*(t) - f_n(t) - \Delta^\top \bar{W}_n \psi_1'(t)$ as

$$\sum_{j=0}^{p}\Delta_j A_j(t) + T_{1n}(\mu_{1n}(t) - \psi_1'(t)) + \frac{1}{2}\sum_{j=0}^{p}\sum_{k=0}^{p}\Delta_j\Delta_k B_{jk}(t) + \frac{1}{2}T_{2n}\mu_{2n}(t) + R_n(t)$$

where

$$T_{ln} = \frac{1}{n}\sum_{i=1}^{n}(\Delta^\top W_i)^l = O_p(n^{-l/2}), \quad l = 1,2.$$

Applications of the Minkowski inequality, the Jensen inequality, and the inequality $|\Delta^\top W_i| \le S_n$ yield

$$\|\hat{f}_n^* - f_n - \Delta^\top W_n \psi_1'\|_\lambda \le \sum_{j=0}^{p}|\Delta_j|\|A_j\|_\lambda + \sum_{j=0}^{p}\sum_{k=0}^{p}|\Delta_j||\Delta_k|\|B_{jk}\|_\lambda + Q_n$$

where

$$Q_n = S_n\|\mu_{1n} - \psi_1'\|_\lambda + S_n^2\|\mu_{2n}\|_\lambda + S_n^3\|K_{h_n}^{'''}\|_\lambda \frac{1}{n}\sum_{i=1}^{n}|\varepsilon_i|^3.$$

Next we show that $\|\mu_{1n} - \psi_1'\|_\lambda = o(1)$ and $\|\mu_{2n}\|_\lambda = O(1/h_n)$. In view of the identity

$$\mu_{1n}(t) = \int K_{h_n}'(t-x)xf(x)\,dx = \int K_{h_n}'(t-x)\psi_1(x)\,dx$$

we have $\mu_{1n} = \psi_1 * K'_{h_n} = \psi'_1 * K_{h_n}$ and

$$\mu_{1n}(t) - \psi'_1(t) = \int (\psi'_1(t - h_n u) - \psi'_1(t))K(u)\,du$$

and find with the help of the Hölder inequality

$$\|\mu_{1n} - \psi'_1\|_\lambda^\lambda \le \int\int |\psi'_1(t - h_n u) - \psi'_1(t)|^\lambda\,dt\,K(u)\,du \to 0.$$

The convergence follows from the Lebesgue-dominated convergence theorem and the fact that the map $s \mapsto \|\psi'_1(\cdot - s) - \psi'_1\|_\lambda$ is bounded and continuous. Similarly, $\mu_{2n} = \psi_2 * K''_{h_n} = \psi'_2 * K'_{h_n}$ and $\|\mu_{2n}\|_\lambda \le \|\psi'_2\|_\lambda \|K'_{h_n}\|_1 = \|\psi'_2\|_\lambda \|K'\|_1 / h_n$. Since ε_1 has a finite third moment and $nh_n^{3+\lambda_*/2} \to \infty$, we derive

$$\sqrt{nh_n^{\lambda_*}}\,Q_n = o_p(h_n^{\lambda_*/2}) + O_p(h_n^{\lambda_*/2}/(n^{1/2}h_n)) + O_p(1/(nh_n^{3+\lambda_*/2})) = o_p(1).$$

Since ε_i is independent of $\varepsilon_{i-1}, X_{i-1}, \varepsilon_{i-2}, X_{i-2}, \ldots$, we see that the summands of $A_j(t)$ are centered and uncorrelated and obtain

$$nE[A_j^2(t)] = \frac{1}{n}\sum_{i=1}^n E[W_{ij}^2(\varepsilon_i K'_{h_n}(t - \varepsilon_i) - \mu_{1n}(t))^2] \le \frac{1}{4\alpha_j^2}\gamma_n(t)$$

with

$$\gamma_n(t) = E[\varepsilon_1^2 (K'_{h_n})^2(t - \varepsilon_1)] = \int \psi_2(y)(K'_{h_n})^2(t - y)\,dy = \psi_2 * (K'_{h_n})^2(t).$$

Thus, for $1 \le \lambda < 2$, we have

$$\begin{aligned}\int E[|A_j(t)|^\lambda]\,dt &\le \int (E[A_j^2(t)])^{\lambda/2}\,dt \le n^{-\lambda/2}(2\alpha_j)^{-\lambda}\int (\gamma_n(t))^{\lambda/2}\,dt \\ &\le n^{-\lambda/2}(2\alpha_j)^{-\lambda}\big(\|V\gamma_n^\lambda\|_1 \|1/V\|_1\big)^{1/2} \\ &\le n^{-\lambda/2}(2\alpha_j)^{-\lambda}\big(\|V\psi_2^\lambda\|_1 \|V(K'_{k_n})^2\|_1^\lambda \|1/V\|_1\big)^{1/2}\end{aligned}$$

with $V(t) = (1 + |t|)^\beta$ and $\beta > 1$ and thus obtain $\|A_j\|_\lambda = O_P(n^{-1/2}h_n^{-3/2})$ provided $\int (1 + |x|)^{2\lambda+\beta} f(x)^\lambda\,dx$ is finite for some $\beta > 1$. Similarly one derives $\|B_{jk}\|_\lambda = O_p(n^{-1/2}h_n^{-5/2})$ provided $\int (1+|x|)^{4\lambda+\beta} f(x)^\lambda\,dx$ is finite for some $\beta > 1$. Thus, for $1 \le \lambda < 2$, we find

$$\sqrt{nh_n^{\lambda_*}}\sum_{j=0}^p |\Delta_j| \|A_j\|_\lambda = O_p(n^{-1/2}h_n^{\lambda_*/2-3/2}) = o_p(1)$$

and

$$\sqrt{nh_n^{\lambda_*}}\sum_{j=0}^p\sum_{k=0}^p |\Delta_j||\Delta_k| \|B_{jk}\|_\lambda = O_p(n^{-1}h_n^{\lambda_*/2-5/2}) = o_p(1).$$

For the case $\lambda \ge 2$, we use the following lemma which gives bounds on the moments of martingales.

Lemma 5.1 *Let $\{S_n, n \geq 1\}$ be a martingale, $S_0 = 0, X_n = S_n - S_{n-1}$. Then for all $\lambda \geq 2$ and $n = 1, 2, \cdots$*

$$E(|S_n)|^\lambda) \leq C_\lambda n^{\lambda/2-1} \sum_{i=1}^{n} E(|X_i|^\lambda),$$

where $C_\lambda = [8(\lambda - 1)\max(1, 2^{\lambda-3})]^\lambda$.

See Dharmadhikari, Fabian, and Jogdeo (1968) for its proof.

For the remainder of this section we assume that $\lambda \geq 2$. Note that

$$nA_j(t) = \sum_{i=1}^{n} W_{ij} Y_{ni}(t), \quad Y_{ni}(t) = \varepsilon_i K'_{h_n}(t - \varepsilon_i) - \mu_{n1}(t).$$

Then $\{\sum_{i=1}^{k} W_{ij} Y_{ni}(t), k = 1, \ldots, n\}$ is a martingale with respect to the filtration $\{\sigma(X_{1-p}, \ldots, X_0, \varepsilon_1, \ldots, \varepsilon_k), k = 0, \ldots, n\}$. Using Lemma 5.1, we calculate

$$\begin{aligned}
E[\|nA_j\|_\lambda{}^\lambda] &= \int E\Big[\Big|\sum_{i=1}^{n} W_{ij} Y_{ni}(t)\Big|^\lambda\Big]\, dt \\
&\leq C_\lambda n^{\lambda/2-1} \int \sum_{i=1}^{n} E[|W_{ij} Y_{ni}(t)|^\lambda]\, dt \\
&\leq \frac{C_\lambda n^{\lambda/2-1}}{2^\lambda \alpha_j^\lambda} \int \sum_{i=1}^{n} E[|Y_{ni}(t)|^\lambda]\, dt \\
&\leq \frac{C_\lambda n^{\lambda/2-1}}{\alpha_j^\lambda} \int \sum_{i=1}^{n} E[|\varepsilon_i K'_{h_n}(t - \varepsilon_i)|^\lambda]\, dt.
\end{aligned}$$

In the last step we used the fact that $E[|Y - E[Y]|^\lambda] \leq 2^\lambda E[Y|^\lambda]$ holds for every random variable Y with finite λ-moment. With $\psi(x) = |x|^\lambda f(x)$, we can write

$$E[|\varepsilon_i K'_{h_n}(t - \varepsilon_i)|^\lambda] = \psi * |K'_{h_n}|^\lambda(t)$$

and find

$$E[\|nA_j\|_\lambda{}^\lambda] \leq \frac{C_\lambda n^{\lambda/2}}{\alpha_j^\lambda} \|\psi\|_1 \|K'_{h_n}\|_\lambda{}^\lambda.$$

Thus, if ε_1 has a finite λ-moment, then $\|A_j\|_\lambda = O_p(n^{-1/2} h^{-1-\lambda_*})$ and we obtain

$$\sqrt{nh_n^{\lambda_*}} \sum_{j=0}^{p} |\Delta_j| \|A_j\|_\lambda = O_p(n^{-1/2} h^{-1-\lambda_*/2}) = o_p(1).$$

In a similar fashion one derives

$$E[\|nB_{jk}\|_\lambda^\lambda] \leq \frac{C_\lambda n^{\lambda/2}}{\alpha_j^\lambda \alpha_k^\lambda} E[|\varepsilon_1|^{2\lambda}] \|K''_{h_n}\|_\lambda^\lambda.$$

Thus, if ε_1 has a finite 2λ-moment, then we have the rate $\|B_{jk}\|_\lambda = O_p(n^{-1/2}h^{-2-\lambda_*})$ and obtain

$$\sqrt{nh_n^{\lambda_*}} \sum_{j=0}^{p} \sum_{k=0}^{p} |\Delta_j||\Delta_k| \|B_{jk}\|_\lambda = O_p(n^{-1}h^{-2-\lambda_*/2}) = o_p(1).$$

This completes the proof.

Acknowledgements This chapter is written in honor of Prof. Hira Koul (my Ph.D. thesis adviser) to celebrate his 70th birthday. I am very grateful to coeditor Anton Schick for helpful comments and suggestions which greatly improved the presentation of this article.

References

Cheng F, Sun S, Wen M (2011) Asymptotics for L2-norm of ARCH innovation density estimator. J Statist Plann Inference 141:3771–3779

Cheng F, Wen M (2011) The L_1 strong consistency of ARCH innovation density estimator. Statist Probab Lett 81:548-551

Cheng F (2008) Extended Glivenko–Cantelli theorem in ARCH(p)-time series. Statist Probab Lett 78:1434–1439

Cheng F (2008) Asymptotic properties in ARCH(p)-time series. J Nonparametr Stat 20:47–60

Csörgő M, Horváth L (1988) Central limit theorems for L_p-norms of density estimators. Probab Theory Relat Fields 80:269–291

Dharmadhikari SW, Fabian V, Jogdeo K (1968) Bounds on the moments of martingales. Ann Math Statist 39:1719–1723

Engle RF (1982) Autoregressive conditional heteroscedasticity with estimates of variance of U.K. inflation. Econometrica 50:987–1008

Gouriérous C (1997) ARCH models and financial applications. Springer, Berlin

Horváth L, Liese F (2004) L_p-estimators in ARCH models. J Statist Plann Inference 119:277–309

Horváth L, Zitikis R (2004) Asymptotics of the Lp-norms of density estimators in the first-order autoregressive models. Statist Probab Lett 66:91–103

Koul H, Mimoto N (2012) A goodness-of-fit test for GARCH innovation density. Metrika 75:127–149

Rudin W (1974) Real and complex analysis, 2nd edn. McGraw–Hill, New York

Schick A, Wefelmeyer W (2007) Root-n consistent density estimators of convolutions in weighted L_1-norms. J Statist Plann Inference 137:1765–1774

Stute W (2001) Residual analysis for ARCH(p)-time series. Test 10:393–403

Weiss AA (1986) Asymptotic theory for ARCH models: estimation and testing. Econometric Theor 2:107–131

Yang X, Fu K, Zhang L (2011) Asymptotic of the L_r-norm of density estimators in the autoregressive time series. Statistics 45:163–178

Chapter 4
Asymptotic Risk and Bayes Risk of Thresholding and Superefficient Estimates and Optimal Thresholding

Anirban DasGupta and Iain M. Johnstone

4.1 Introduction

The classic Hodges' estimator (Hodges, 1951, unpublished) of a one dimensional normal mean demolished the statistical folklore that maximum likelihood estimates are asymptotically uniformly optimal, provided the family of underlying densities satisfies enough regularity conditions. Hodges' original estimate is

$$T_n(X_1, \cdots, X_n) = \begin{cases} \bar{X}_n & \text{if} \quad |\bar{X}_n| > n^{-1/4} \\ 0 & \text{if} \quad |\bar{X}_n| \le n^{-1/4} \end{cases} \tag{4.1}$$

A more general version is

$$S_n(X_1, \ldots, X_n) = \begin{cases} \bar{X}_n & \text{if} \quad |\bar{X}_n| > c_n \\ a_n \bar{X}_n & \text{if} \quad |\bar{X}_n| \le c_n \end{cases} \tag{4.2}$$

Here, c_n, for the moment, is a general positive sequence and $0 \le a_n \le 1$. With squared error as the loss function, the risk of $\bar{X}_n$, the unique MLE, satisfies $nR(\theta, \bar{X}_n) \equiv 1$, and Hodges' original estimate T_n satisfies

$$\lim_{n\to\infty} n^{\beta} R(0, T_n) = 0 \ \forall \beta > 0,$$

while

$$\lim_{n\to\infty} \sup_{\theta} \ nR(\theta, T_n) = \infty.$$

Thus, at $\theta = 0$, Hodges' estimate is asymptotically infinitely superior to the MLE, while globally its peak risk is infinitely more relative to that of the MLE. *Superefficiency at $\theta = 0$ is purchased at a price of infinite asymptotic inflation in risk away from zero*. Hodges' example showed that the claim of the uniform asymptotic optimality of the MLE is false even in the normal case, and it seeded the development

A. DasGupta (✉) · I. M. Johnstone
Purdue University and Stanford University, USA
e-mail: dasgupta@purdue.edu

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_4,

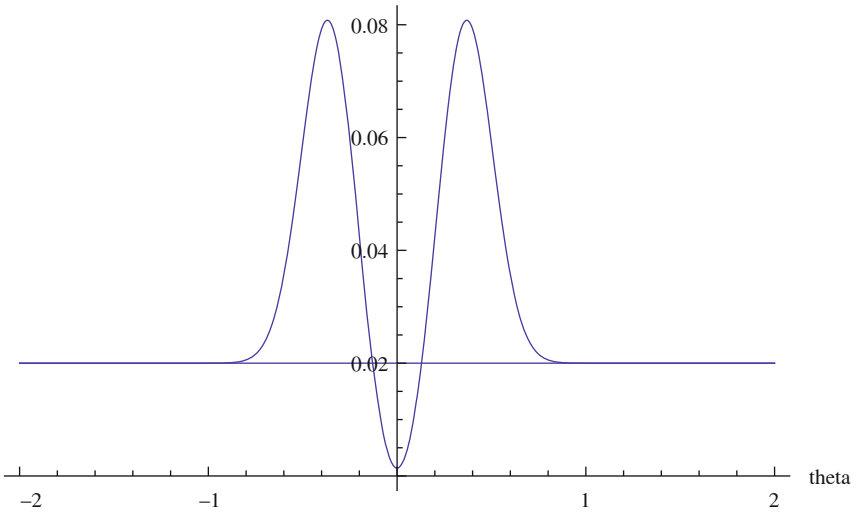


Fig. 4.1 Risk of Hodges' Estimate for n=50

of such fundamental concepts as regular estimates. It culminated in the celebrated *Hájek-Le Cam convolution theorem*. It probably, also had some indirect impact on the development and study of the now common *thresholding estimates* in *large p small n* problems, the most well known among them being the Donoho-Johnstone estimates (Donoho and Johnstone (1994)), although while the classic Hodges' estimate uses a small threshold ($n^{-1/4}$), the new thresholding estimates use a large threshold (Fig 4.1).

It is of course already well understood that the risk inflation of Hodges' estimate occurs *close to zero*, and that the worst inflation occurs in a neighborhood of small size. This was explicitly pointed out in Le Cam (1953):

$$\lim_{n\to\infty} \sup_{U_n} \sup_{\theta\in U_n} nR(\theta, T_n) = \infty,$$

where U_n denotes a general sequence of open neighborhoods of zero such that $\lambda(U_n)$, the Lebesgue measure of U_n, goes to zero; *we cannot have asymptotic superefficiency in nonvanishing neighborhoods*. Provided only that a competitor estimate sequence T_n has a limit distribution under every θ, i.e., $\sqrt{n}(T_n - \theta)$ has some limiting distribution L_θ, it must have an asymptotic pointwise risk at least as large as that of $\bar{X}$ at almost all θ:

$$\text{For almost all } \theta, \limsup_{n\to\infty} nR(\theta, T_n) \geq 1.$$

Indeed, a plot of the risk function of Hodges' estimate nicely illustrates these three distinct phenomena, *superefficiency at zero, inflation close to zero, worst inflation in a shrinking neighborhood*: Similar in spirit are the contemporary thresholding

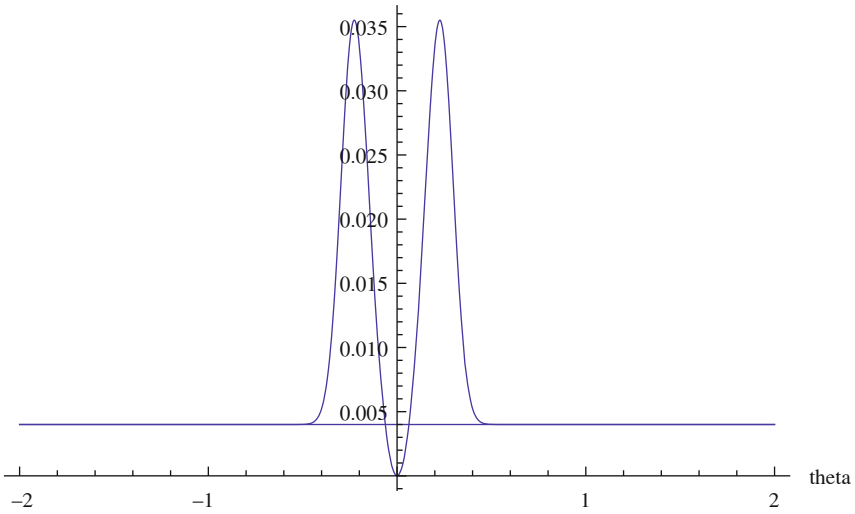


Fig. 4.2 Risk of Hodges' Estimate for n = 250

estimates of Gaussian means. Formally, given $X \sim N(\theta, 1)$, and $\lambda > 0$, the *hard thresholding estimate* is defined as

$$\begin{aligned} \hat{\theta}_\lambda &= X \quad \text{if } |X| > \lambda \\ &= 0 \quad \text{if } |X| \leq \lambda \end{aligned}$$

Implicit in this construction is an underlying Gaussian sequence model

$$X_i \overset{indep.}{\sim} N(\theta_i, 1), i = 1, 2, \cdots, n,$$

and

$$\hat{\theta}_i = X_i I_{|X_i| > \lambda(n)}, \tag{4.3}$$

and $\lambda(n)$ often being asymptotic to $\sqrt{2 \log n}$, which is a first order asymptotic approximation (although not very accurate practically) to the expectation of the maximum of n iid $N(0, 1)$ observations. The idea behind this construction is that we expect nearly all the means to be zero (i.e., the observed responses are instigated by pure noise), and we estimate a specific θ_i to be equal to the observed signal only if the observation stands out among a crowd of roughly n pure Gaussian white noises. See Johnstone (2012) for extensive discussion and motivation (Fig 4.2).

The similarity between Hodges' estimate and the above hard thresholding estimate is clear. We would expect the hard thresholding estimate to manifest risk phenomena similar to that of Hodges' estimate: better risk than the naive estimate X_i itself if the true θ_i is zero, risk inflation if the true θ_i is adequately away from zero, and we expect that the finer details will depend on the choice of the threshold level λ. One may ask what is the *optimal* λ that suitably balances the risk gain at zero with the risk inflation away from zero.

Another commonality in the behavior of Hodges' estimate and the hard thresholding estimate is that if we take a prior distribution on the true mean that is very tightly concentrated near zero, then they ought to have smaller Bayes risks than the MLE, and the contrary is expected if we take an adequately diffuse prior.

It is meaningful and also interesting to ask if these various anticipated phenomena can be pinned down with some mathematical precision. The main contributions of this article are the following:

a) For the one dimensional Gaussian mean and superefficient estimates of the general form as in (4.2), we precisely quantify the behavior of the risk at zero (Eq. (4.10), Corollary 1.2.5).
b) We precisely quantify the risk at $\frac{k}{\sqrt{n}}$ for fixed positive k (Eq. (4.22)), and we show that the risk at $\frac{1}{\sqrt{n}}$ (which is exactly one standard deviation away from zero) is for all practical purposes equal to $\frac{1}{n}$, which is the risk of the MLE (Theorem 1.2.4, Corollary 1.2.5).
c) We show that in the very close vicinity of zero, the risk of superefficient estimates increases at an increasing rate, i.e., the risk is locally convex (Theorem 1.2.2).
d) We show that the global peak of the risk is *not* attained within $n^{-1/2}$ neighborhoods. In fact, we show that at $\theta = c_n$, the risk is much higher (Theorem 1.2.5, Eq. (4.26)), and that *immediately below* $\theta = c_n$, the risk is even higher. Precisely, we exhibit explicit and parsimonious shrinking neighborhoods U_n of $\theta = c_n$, such that

$$\liminf c_n^{-2} \sup_{\theta \in U_n} R(\theta, S_n) \geq 1. \tag{4.4}$$

(Theorem 1.2.6, Eq. (4.28)). Note that we can obtain the lower bound in (4.4) with an lim inf, rather than lim sup.

Specifically, our calculations indicate that $\text{argmax}_\theta R(\theta, S_n) \approx c_n - \sqrt{\frac{\log(nc_n^2)}{n}}$, and $\sup_\theta R(\theta, S_n) \approx c_n^2 - 2c_n\sqrt{\frac{\log n}{n}}$ (Eq. (4.35)).
e) For normal priors $\pi_n = N(0, \sigma_n^2)$, we obtain exact closed form expressions for the Bayes risk $B_n(\pi_n, S_n)$ of S_n (Theorem 1.2.7, Eq. (4.45)), and characterize those priors for which $B_n(\pi_n, S_n) \leq \frac{1}{n}$ for all large n. Specifically, we show that $\sigma^2 = \frac{1}{n}$ acts in a very meaningful way as the boundary between $B_n(\pi_n, S_n) < \frac{1}{n}$ and $B_n(\pi_n, S_n) > \frac{1}{n}$ (Theorem 1.2.8).
More generally, we use the theory of regular variation to show the quite remarkable fact that for *general smooth* prior densities $\pi_n(\theta) = \sqrt{n}h(\theta\sqrt{n})$, *all Hodges type estimates are approximately equivalent in Bayes risk to the MLE* $\bar{X}$ and that the exact rate of convergence of the difference in Bayes risks is determined by whether or not $\text{Var}_h(\theta) = 1$ (Theorem 1.2.10, Eq. (4.64)). This theorem, in turn, follows from a general convolution representation for the difference in Bayes risks under general π_n (Theorem 1.2.9, Eq. (4.48)).
f) For the Gaussian sequence model, we obtain appropriate corresponding versions of a)-e) for hard thresholding estimates of the form (4.3).

g) We identify the specific estimate in the class (4.2) that minimizes an approximation to the global maximum of the risk subject to a guaranteed specified improvement at zero; this is usually called a *restricted minimax problem*. More precisely, we show that subject to the constraint that the percentage risk improvement at zero is at least $100(1-\epsilon_n)\%$, the global maximum risk is approximately minimized when $c_n = \sqrt{2\log\frac{1}{\epsilon_n}}$ (Eq. (4.38)).

h) We illustrate the various results with plots, examples, and summary tables.

Several excellent sources where variants of a few of our problems have been addressed include Hájek (1970), Johnstone (2012), Le Cam (1953, 1973), Lehmann and Romano (2005), van der Vaart (1997, 1998), and Wasserman (2005). Also, see DasGupta (2008) and lecture notes written by Jon Wellner and Moulinath Banerjee. Superefficiency has also been studied in some problems that do not have the LAN (locally asymptotically normal) structure; one reference is Jeganathan (1983).

If the variance σ^2 of the observations was unknown, estimates similar to Hodges' are easily constructed by hard thresholding the MLE whenever $\frac{|\bar{X}|}{s} \le c_n$, where s is the sample standard deviation. Some of its risk properties can be derived along the lines of this article. However, the optimal thresholding and global maximum risk problems are likely to be even more difficult.

4.2 Risk Function of Generalized Hodges Estimates

Consider generalized Hodges estimates of the form (4.2). We first derive an expression for the risk function of the estimate $S_n(X_1, \cdots, X_n)$. This formula will be repeatedly used for many of the subsequent results. This formula for the risk function then leads to formulas for its successive derivatives, which are useful to pin down finer properties of S_n.

4.2.1 Global Formulas

Theorem 1.2.1 *Let $n \ge 1$ and $X_1, \cdots, X_n$ iid $N(\theta, 1)$. Let $0 \le a_n \le 1$ and $c_n > 0$. For the estimate $S_n(X_1, \cdots, X_n)$ as in* (4.2)*, the risk function under squared error loss is given by*

$$R(\theta, S_n) = \frac{1}{n} + e_n(\theta),$$

where

$$e_n(\theta) = \Big[\frac{a_n^2 - 1}{n} + (1-a_n)^2\theta^2\Big]\Big(\Phi(\sqrt{n}(c_n-\theta)) + \Phi(\sqrt{n}(c_n+\theta)) - 1\Big)$$

$$+\frac{2a_n(a_n-1)\theta}{\sqrt{n}}\Big(\phi(\sqrt{n}(c_n+\theta)) - \phi(\sqrt{n}(c_n-\theta))\Big)$$

$$+\frac{1-a_n^2}{\sqrt{n}}\Big((c_n+\theta)\phi(\sqrt{n}(c_n+\theta))+(c_n-\theta)\phi(\sqrt{n}(c_n-\theta))\Big), \tag{4.5}$$

where ϕ and Φ denote the density and the CDF of the standard normal distribution.

Proof Write $R(\theta, S_n)$ as

$$R(\theta,S_n)=E[(\bar{X}-\theta)^2 I_{|\bar{X}|>c_n}]+E[(a_n\bar{X}-\theta)^2 I_{|\bar{X}|\le c_n}]$$

$$=E[(\bar{X}-\theta)^2]+E[(a_n\bar{X}-\theta)^2 I_{|\bar{X}|\le c_n}]-E[(\bar{X}-\theta)^2 I_{|\bar{X}|\le c_n}]$$

$$=\frac{1}{n}+\int_{-\sqrt{n}(c_n+\theta)}^{\sqrt{n}(c_n-\theta)}[a_n(\theta+\frac{z}{\sqrt{n}})-\theta]^2\phi(z)dz-\frac{1}{n}\int_{-\sqrt{n}(c_n+\theta)}^{\sqrt{n}(c_n-\theta)}z^2\phi(z)dz$$

$$=\frac{1}{n}+T_1+T_2 \qquad \text{(say)} \tag{4.6}$$

On calculation, we get

$$T_1=\Big[\frac{a_n^2}{n}+(1-a_n)^2\theta^2\Big]\Big(\Phi(\sqrt{n}(c_n-\theta))+\Phi(\sqrt{n}(c_n+\theta))-1\Big)$$

$$-\frac{a_n^2}{\sqrt{n}}\Big((c_n+\theta)\phi(\sqrt{n}(c_n+\theta))+(c_n-\theta)\phi(\sqrt{n}(c_n-\theta))\Big)$$

$$+\frac{2a_n(a_n-1)\theta}{\sqrt{n}}\Big(\phi(\sqrt{n}(c_n+\theta))-\phi(\sqrt{n}(c_n-\theta))\Big), \tag{4.7}$$

and

$$T_2=\frac{1}{n}\Big(\Phi(\sqrt{n}(c_n-\theta))+\Phi(\sqrt{n}(c_n+\theta))-1\Big)$$

$$-\frac{1}{\sqrt{n}}\Big((c_n+\theta)\phi(\sqrt{n}(c_n+\theta))+(c_n-\theta)\phi(\sqrt{n}(c_n-\theta))\Big) \tag{4.8}$$

On combining (4.6), (4.7), and (4.8), and further algebraic simplification, the stated expression in (4.5) follows.

4.2.1.1 Behavior at Zero

Specializing the global formula (4.5) to $\theta = 0$, we can accurately pin down the improvement at zero.

Corollary 1.2.1 *The risk improvement of S_n over $\bar{X}$ at $\theta = 0$ satisfies*

$$e_n(0) = \frac{1}{n} - R(0, S_n) = \frac{2(1 - a_n^2)}{n}\phi(\sqrt{n}c_n)\Big[\frac{\Phi(\sqrt{n}c_n) - \frac{1}{2}}{\phi(\sqrt{n}c_n)} - \sqrt{n}c_n\Big] \tag{4.9}$$

Furthermore, provided that $\limsup_n |a_n| \leq 1$, and $\gamma_n = \sqrt{n}c_n \to \infty$,

$$R(0, S_n) = \frac{a_n^2}{n} + \sqrt{\frac{2}{\pi}}\,\frac{1 - a_n^2}{n}\gamma_n\, e^{-\gamma_n^2/2} + o(\frac{\gamma_n e^{-\gamma_n^2/2}}{n}) \tag{4.10}$$

Corollary 1.2.1 *can be proved by using* (4.5) *and standard facts about the $N(0, 1)$ CDF; we will omit these details.*

An important special case of Corollary 1.2.1 is the original Hodges' estimate, for which $c_n = n^{-1/4}$ and $a_n \equiv 0$. In this case, an application of Corollary 1.2.1 gives the following asymptotic expansion; it is possible to make this into a higher order asymptotic expansion, although it is not done here.

Corollary 1.2.2 *For Hodges' estimate T_n as in* (4.1),

$$R(0, T_n) = \sqrt{\frac{2}{\pi}}n^{-3/4}e^{-\frac{\sqrt{n}}{2}} + o(n^{-3/4}e^{-\frac{\sqrt{n}}{2}}) \tag{4.11}$$

In particular,

$$\lim_{n\to\infty} \frac{\log(nR(0, T_n))}{\sqrt{n}} = -\frac{1}{2} \tag{4.12}$$

We record the following corollary for completeness. Note that $\sqrt{n}c_n$ need not go to ∞ for superefficiency to occur, as shrinkage will automatically take care of it.

Corollary 1.2.3 *Suppose $\gamma_n = \sqrt{n}c_n \to \gamma, 0 < \gamma \leq \infty$. Then, S_n is superefficient at zero, i.e.,* $\limsup_n nR(0, S_n) < 1$ *iff* $\limsup_n |a_n| < 1$.

4.2.1.2 Local Convexity and Behavior in Ultrasmall Neighborhoods

For understanding the local shape properties of the risk function of S_n, it is necessary to understand the behavior of its derivatives. This is the content of the next result, which says in particular that the risk function of all generalized Hodges estimates is locally convex near zero. For these results, we need the following notation:

$$f_n(\theta) = (1 - a_n)^2\theta\left[2\Phi(\sqrt{n}(c_n + \theta)) - 1\right] \tag{4.13}$$

$$g_n(\theta) = (a_n - 1)\left[(1 + a_n)\sqrt{n}c_n^2 + \frac{2a_n}{\sqrt{n}} + 2\sqrt{n}c_n\theta\right]\phi(\sqrt{n}(c_n + \theta)) \tag{4.14}$$

Theorem 1.2.2 *For all n and θ,*

$$\frac{d}{d\theta}R(\theta, S_n) = f_n(\theta) - f_n(-\theta) + g_n(\theta) - g_n(-\theta) \tag{4.15}$$

In particular, $\frac{d}{d\theta}R(\theta, S_n)|_{\theta=0} = 0$, and provided that $|a_n| < 1$, $\frac{d^2}{d\theta^2}R(\theta, S_n) > 0$ in a neighborhood of $\theta = 0$. Hence, under the hypothesis that $|a_n| < 1$, $R(\theta, S_n)$ is locally convex near zero, and $\theta = 0$ is a local minima of $R(\theta, S_n)$.

Proof Proof of (4.15) is a direct calculation followed by rearranging the various terms. The calculation is not presented.

That the derivative of $R(\theta, S_n)$ at $\theta = 0$ is zero follows from symmetry of $R(\theta, S_n)$, or, also immediately from (4.15). We now sketch a proof of the local convexity property. Differentiating (4.15),

$$\frac{d^2}{d\theta^2}R(\theta, S_n) = f_n'(\theta) + f_n'(-\theta) + g_n'(\theta) + g_n'(-\theta). \tag{4.16}$$

Now, on algebra,

$$f_n'(\theta) = (1 - a_n)^2\left[2\Phi(\sqrt{n}(c_n + \theta)) - 1\right] + 2\theta(1 - a_n)^2\sqrt{n}\phi(\sqrt{n}(c_n + \theta))$$

and $g_n'(\theta) = 2(a_n - 1)\sqrt{n}c_n\phi(\sqrt{n}(c_n + \theta)) - n(c_n + \theta)\phi(\sqrt{n}(c_n + \theta))$

$$\times\left[2(a_n - 1)\sqrt{n}c_n\theta + \frac{2a_n(a_n - 1)}{\sqrt{n}} + (a_n^2 - 1)\sqrt{n}c_n^2\right] \tag{4.17}$$

On substituting (4.17) into (4.16), and then setting $\theta = 0$, we get after further algebraic simplification,

$$\begin{aligned}\frac{d^2}{d\theta^2}R(\theta, S_n)|_{\theta=0} =& 4(1 - a_n)^2\left[\Phi(\sqrt{n}c_n) - \frac{1}{2} - \sqrt{n}c_n\phi(\sqrt{n}c_n)\right] \\ &+ 2(1 - a_n^2)c_n^3 n^{3/2}\phi(\sqrt{n}c_n)\end{aligned} \tag{4.18}$$

By simple calculus, $\Phi(x) - \frac{1}{2} - x\phi(x) > 0$ for all positive x. Therefore, on using our hypothesis that $|a_n| < 1$, from (4.18), $\frac{d^2}{d\theta^2}R(\theta, S_n)|_{\theta=0} > 0$. It follows from the continuity of $\frac{d^2}{d\theta^2}R(\theta, S_n)$ that it remains strictly positive in a neighborhood of $\theta = 0$, which gives the local convexity property.

Remark Consider now the case of original Hodges' estimate, for which $a_n = 0$ and $c_n = n^{-1/4}$. In this case, (4.18) gives us $\lim_{n\to\infty}\frac{d^2}{d\theta^2}R(\theta, T_n)|_{\theta=0} = 2$. Together with (4.11), we then have the approximation

$$R(\theta, T_n) \approx \sqrt{\frac{2}{\pi}}n^{-3/4}e^{-\frac{\sqrt{n}}{2}} + \theta^2 \tag{4.19}$$

for θ very close to zero. Of course, we know that this approximation cannot depict the subtleties of the shape of $R(\theta, T_n)$, because $R(\theta, T_n)$ is known to have turning points, which the approximation in (4.19) fails to recognize. *We will momentarily see that $R(\theta, T_n)$ rises and turns so steeply that (4.19) is starkly inaccurate in even $n^{-1/2}$ neighborhoods of zero.*

4.2.2 Behavior in $n^{-1/2}$ Neighborhoods

We know that the superefficient estimates T_n, or S_n have a much smaller risk than the MLE at zero, and that subsequently their risks reach a peak that is much higher than that of the MLE. Therefore, these risk functions must again equal the risk of the MLE, namely $\frac{1}{n}$ at some point in the vicinity of zero. We will now first see that reversal to the $\frac{1}{n}$ level happens within $n^{-1/2}$ neighborhoods of zero. A general risk lower bound for generalized Hodges estimates S_n would play a useful role for this purpose, and also for a number of the later results. This is presented first.

Theorem 1.2.3 *Consider the generalized Hodges estimate S_n.*

(i) Suppose $0 \le a_n \le 1$. Then, for every n and $0 \le \theta \le c_n$,

$$R(\theta, S_n) \ge \frac{a_n^2}{n} + (1-a_n)^2\theta^2\Big[\Phi(\sqrt{n}(c_n+\theta)) + \Phi(\sqrt{n}(c_n-\theta)) - 1\Big] \tag{4.20}$$

(ii) Suppose $\sqrt{n}c_n \to \infty$, and that $a, 0 \le a < 1$ is a limit point of the sequence a_n. Let $\theta_n = \frac{1}{(1-a)^2\sqrt{n}}$. Then, $\limsup_n nR(\theta_n, S_n) \ge a^2 + 1$.

Proof In expression (4.5) for $e_n(\theta)$, observe the following:

$$0 \le \Phi(\sqrt{n}(c_n+\theta)) + \Phi(\sqrt{n}(c_n-\theta)) - 1 \le 1;$$

$$\text{For } 0 \le \theta \le c_n,\ \phi(\sqrt{n}(c_n+\theta)) - \phi(\sqrt{n}(c_n-\theta)) \le 0;$$

$$\text{For } 0 \le \theta \le c_n,\ (c_n+\theta)\phi(\sqrt{n}(c_n+\theta)) + (c_n-\theta)\phi(\sqrt{n}(c_n-\theta)) \ge 0.$$

Therefore, by virtue of the hypothesis $0 \le a_n \le 1$, from (4.5),

$$R(\theta, S_n) \ge \frac{1}{n} + \frac{a_n^2 - 1}{n} + (1-a_n)^2\theta^2\Big[\Phi(\sqrt{n}(c_n+\theta)) + \Phi(\sqrt{n}(c_n-\theta)) - 1\Big]$$

$$= \frac{a_n^2}{n} + (1-a_n)^2\theta^2\Big[\Phi(\sqrt{n}(c_n+\theta)) + \Phi(\sqrt{n}(c_n-\theta)) - 1\Big],$$

as claimed in (4.20).

For the second part of the theorem, choose a subsequence $\{a_{n_k}\}$ of $\{a_n\}$ converging to a. For notational brevity, we denote the subsequence as a_n itself. Then, (along this subsequence), and with $\theta_n = \frac{1}{(1-a)^2\sqrt{n}}$,

$$a_n^2 + (1-a_n)^2\theta_n^2\Big[\Phi(\sqrt{n}(c_n+\theta_n)) + \Phi(\sqrt{n}(c_n-\theta_n)) - 1\Big] \to a^2 + 1 \tag{4.21}$$

Since we assume for the second part of the theorem that $\sqrt{n}c_n \to \infty$, we have that $\theta_n \le c_n$ for all large n, and hence the lower bound in (4.20) applies. Putting together

(4.20) and (4.21), and the Bolzano-Weierstrass theorem, we have one subsequence for which the limit of $nR(\theta_n, S_n)$ is $\geq a^2+1$, and hence, $\limsup_n nR(\theta_n, S_n) \geq a^2+1$.

We will now see that if we strengthen our control on the sequence $\{a_n\}$ to require it to have a limit, and likewise require $\sqrt{n}c_n$ also to have a limit, then the (normalized) risk of S_n at $\frac{k}{\sqrt{n}}$ will also have a limit for any given k. *Furthermore, if the limit of a_n is zero and the limit of $\sqrt{n}c_n$ is ∞, which, for instance, is the case for Hodges' original estimate, then the risk of S_n at $\frac{1}{\sqrt{n}}$ is exactly asymptotic to the risk of the MLE, namely $\frac{1}{n}$. So, reversal to the risk of the MLE occurs, more or less, at $\theta = \frac{1}{\sqrt{n}}$.* The next result says that, but in a more general form.

Theorem 1.2.4 *Consider the generalized Hodges estimate S_n.*

(a) If $a_n \to a, -\infty < a < \infty$, and $\sqrt{n}c_n \to \gamma, 0 \leq \gamma \leq \infty$, then for any fixed $k \geq 0$,

$$\lim_{n\to\infty} nR(\frac{k}{\sqrt{n}}, S_n) = 1 + \Big[a^2 - 1 + k^2(1-a)^2\Big]\Big[\Phi(k+\gamma) - \Phi(k-\gamma)\Big]$$
$$+ 2a(a-1)k\Big[\phi(k+\gamma) - \phi(k-\gamma)\Big]$$
$$+ (1-a^2)\Big[(k+\gamma)\phi(k+\gamma) - (k-\gamma)\phi(k-\gamma)\Big], \tag{4.22}$$

with (4.22) being interpreted as a limit as $\gamma \to \infty$ if $\sqrt{n}c_n \to \infty$.

(b) In particular, if $a_n \to 0$ and $\sqrt{n}c_n \to \infty$, then, $\lim_{n\to\infty} nR(\frac{k}{\sqrt{n}}, S_n) = k^2$.

(c) If $a_n = 0$ for all n and $\sqrt{n}c_n \to \infty$, then for any positive k, we have the asymptotic expansion

$$nR(\frac{k}{\sqrt{n}}, S_n) = k^2 + \frac{1}{\sqrt{2\pi}} e^{-\gamma_n^2/2 - k^2/2}$$
$$\times \Big[(\gamma_n - k)e^{k\gamma_n} + (\gamma_n + k)e^{-k\gamma_n} - (k^2-1)\frac{e^{k\gamma_n}}{\gamma_n} - (k^2-1)\frac{e^{-k\gamma_n}}{\gamma_n}\Big]$$
$$+ O(\frac{e^{-\gamma_n^2/2 + k\gamma_n}}{\gamma_n^2}) \tag{4.23}$$

(d) If $a_n = 0$ for all n and $\sqrt{n}c_n \to \infty$, then for $k = 0$, we have the asymptotic expansion

$$nR(0, S_n) = \sqrt{\frac{2}{\pi}} e^{-\gamma_n^2/2}\Big[\gamma_n + \frac{2}{\gamma_n}\Big] + O(\frac{e^{-\gamma_n^2/2}}{\gamma_n^3}) \tag{4.24}$$

The plot below nicely exemplifies the limit result in part (b) of Theorem 1.2.4 Fig. 4.3.

The proofs of the various parts of Theorem 1.2.4 involve use of standard facts about the standard normal tail and rearrangement of terms. We omit these calculations. It follows from part (b) of this theorem, by letting $k \to \infty$ that for the original Hodges' estimate T_n, $\sup_\theta R(\theta, T_n) >> \frac{1}{n}$ for large n, in the following sense.

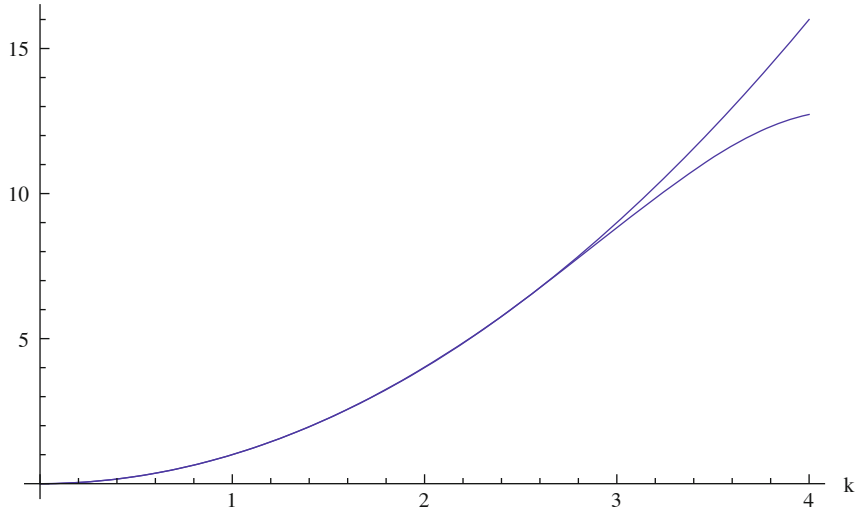


Fig. 4.3 Plot of n* Risk of Hodges' Estimate at k/sqrt(n) and k^2 for n = 500

Corollary 1.2.4 *If* $a_n \to 0$ *and* $\sqrt{nc_n} \to \infty$, *then* $\lim_n \left[\sup_\theta nR(\theta, S_n)\right] = \infty$. *On the other hand, part (c) and part (d) of the above theorem together lead to the following asymptotic expansions for the risk of Hodges' original estimate* T_n *at* $\theta = 0$ *and* $\theta = \frac{1}{\sqrt{n}}$. *We can see how close to* $\frac{1}{n}$ *the risk at* $\frac{1}{\sqrt{n}}$ *is, and the rapid relative growth of the risk near* $\theta = 0$ *by comparing the two expansions in the corollary below, which is also a strengthening of Corollary* 1.2.2.

Corollary 1.2.5 *For Hodges' estimate* T_n *as in* (4.1),

$$R(0, T_n) = \sqrt{\frac{2}{\pi}} e^{-\frac{\sqrt{n}}{2}} n^{-3/4}\left[1 + \frac{2}{\sqrt{n}}\right] + O(\frac{e^{-\frac{\sqrt{n}}{2}}}{n^{7/4}}); R(\frac{1}{\sqrt{n}}, T_n)$$
$$= \frac{1}{n} + \frac{1}{\sqrt{2\pi}} n^{-3/4} e^{-\frac{1}{2}(n^{1/4}-1)^2}\left[1 - n^{-1/4}\right] + O(\frac{e^{-\frac{1}{2}(n^{1/4}-1)^2}}{n^{3/2}}) \quad (4.25)$$

4.2.3 Behavior in c_n Neighborhoods

We saw in the previous section that reversal to the risk of the MLE occurs in $n^{-1/2}$ neighborhoods of zero. However, $n^{-1/2}$ neighborhoods are still too short for the risk to begin to approach its peak value. If $c_n >> \frac{1}{\sqrt{n}}$ and we expand the neighborhood of $\theta = 0$ to c_n neighborhoods, then the risk of S_n increases by factors of magnitude, and captures the peak value. We start with the risk of S_n at $\theta = c_n$ and analyze its asymptotic behavior.

Theorem 1.2.5 *Consider the generalized Hodges estimate S_n.*

(a) Suppose $0 \le a_n \le 1$ and that $\sqrt{n}c_n \to \infty$. Then, $\limsup_n c_n^{-2} R(c_n, S_n) \ge \frac{(1-\liminf_n a_n)^2}{2}$, and $\liminf_n c_n^{-2} R(c_n, S_n) \ge \frac{(1-\limsup_n a_n)^2}{2}$.

(b) If $a_n \to a, -\infty < a < \infty$, and $\sqrt{n}c_n \to \gamma, 0 \le \gamma \le \infty$, then

$$\lim_{n\to\infty} c_n^{-2} R(c_n, S_n) = \frac{1}{\gamma^2} + \Big[\frac{a^2-1}{\gamma^2} + (1-a)^2\Big]\Big[\Phi(2\gamma) - \frac{1}{2}\Big]$$

$$+ \frac{2a(a-1}{\gamma}\Big[\phi(2\gamma) - \phi(0)\Big] + 2(1-a^2)\frac{\phi(2\gamma)}{\gamma}, \tag{4.26}$$

with (4.26) being interpreted as a limit as $\gamma \to \infty$ if $\sqrt{n}c_n \to \infty$.

Proof By (4.20),

$$R(c_n, S_n) \ge \frac{a_n^2}{n} + c_n^2(1-a_n)^2\Big[\Phi(2\sqrt{n}c_n) - \frac{1}{2}\Big]$$

$$\Rightarrow c_n^{-2} R(c_n, S_n) \ge (1-a_n)^2\Big[\Phi(2\sqrt{n}c_n) - \frac{1}{2}\Big] \tag{4.27}$$

Since $\sqrt{n}c_n \to \infty$, (4.24) implies that given $\epsilon > 0$, for all large enough n,

$$c_n^{-2} R(c_n, S_n) \ge (\frac{1}{2} - \epsilon)(1-a_n)^2$$

$$\Rightarrow \limsup_n c_n^{-2} R(c_n, S_n) \ge \limsup_n (\frac{1}{2} - \epsilon)(1-a_n)^2 = (\frac{1}{2} - \epsilon)(1 - \liminf_n a_n)^2.$$

Since $\epsilon > 0$ is arbitrary, this means $\limsup_n c_n^{-2} R(c_n, S_n) \ge \frac{(1-\liminf_n a_n)^2}{2}$; the lim inf inequality follows similarly.

4.2.3.1 Behavior Near c_n and Approach to the Peak

Theorem 1.2.6 *Consider the generalized Hodges estimate S_n. Suppose $a_n = 0$ for all n and $\gamma_n = \sqrt{n}c_n \to \infty$. Then, for any fixed $\alpha, 0 < \alpha \le 1$, we have the asymptotic expansion*

$$c_n^{-2} R((1-\alpha)c_n, S_n) = (1-\alpha)^2 + \frac{\phi(\alpha\gamma_n)}{\alpha\gamma_n}(2\alpha - 1) + \frac{\phi((2-\alpha)\gamma_n)}{(2-\alpha)\gamma_n}(3 - 2\alpha)$$
$$+ O(\frac{\phi(\alpha\gamma_n)}{\gamma_n^3}) \tag{4.28}$$

Proof: Fix $0 < \alpha < 1$, and denote $\theta_n = (1-\alpha)c_n$. Using (4.5),

$$
\begin{aligned}
R(\theta_n, S_n) &= \frac{1}{n} + \Big[(1-\alpha)^2 c_n^2 - \frac{1}{n}\Big]\Big[\Phi((2-\alpha)\gamma_n) - \Phi(-\alpha\gamma_n)\Big] \\
&\quad + \frac{1}{\sqrt{n}}\Big[(2-\alpha)c_n\phi((2-\alpha)\gamma_n) + \alpha c_n \phi(\alpha\gamma_n)\Big] \\
&\Rightarrow c_n^{-2} R(\theta_n, S_n) = \frac{1}{\gamma_n^2} + \Big[(1-\alpha)^2 - \frac{1}{\gamma_n^2}\Big]\Big[\Phi((2-\alpha)\gamma_n) - \Phi(-\alpha\gamma_n)\Big] \\
&\quad + \frac{1}{\gamma_n}\Big[(2-\alpha)\phi((2-\alpha)\gamma_n) + \alpha\phi(\alpha\gamma_n)\Big] = \frac{1}{\gamma_n^2} + \Big[(1-\alpha)^2 - \frac{1}{\gamma_n^2}\Big] \\
&\quad \Big[1 - \frac{\phi((2-\alpha)\gamma_n)}{(2-\alpha)\gamma_n}(1 + O(\gamma_n^{-2})) - \frac{\phi(\alpha\gamma_n)}{\alpha\gamma_n}(1 + O(\gamma_n^{-2}))\Big] \\
&\quad + \frac{(2-\alpha)\phi((2-\alpha)\gamma_n)}{\gamma_n} + \frac{\alpha\phi(\alpha\gamma_n)}{\gamma_n} \\
&= (1-\alpha)^2 + \frac{\phi((2-\alpha)\gamma_n)}{\gamma_n}\Big[(2-\alpha) - \frac{(1-\alpha)^2}{2-\alpha}\Big] \\
&\quad + \frac{\phi(\alpha\gamma_n)}{\gamma_n}\Big[\alpha - \frac{(1-\alpha)^2}{\alpha}\Big] + O(\frac{\phi(\alpha\gamma_n)}{\gamma_n^3}). \qquad (4.29)
\end{aligned}
$$

The theorem now follows from (4.29).

By scrutinizing the proof of Theorem 1.2.6, we notice that the constant α can be generalized to suitable sequences α_n, and this gives us a useful and more general corollary. Note that, indeed, the remainder term in the corollary below is $O(\frac{\phi(\alpha_n\gamma_n)}{\gamma_n})$, rather than $O(\frac{\phi(\alpha_n\gamma_n)}{\gamma_n^3})$.

Corollary 1.2.6 *Consider the generalized Hodges estimate S_n. Suppose $a_n = 0$ for all n and $\gamma_n = \sqrt{n}c_n \to \infty$. Let α_n be a positive sequence such that $\alpha_n \to 0, \alpha_n\gamma_n \to \infty$. Let $\theta_n = (1-\alpha_n)c_n$. Then we have the asymptotic expansion*

$$
c_n^{-2} R(\theta_n, S_n) = (1-\alpha_n)^2 - \frac{\phi(\alpha_n\gamma_n)}{\alpha_n\gamma_n} + O(\frac{\phi(\alpha_n\gamma_n)}{\gamma_n}) \qquad (4.30)
$$

Remark Together, Theorem 1.2.5 and Corollary 1.2.6 enable us to make the following conclusion: *at $\theta = c_n$, $R(\theta, S_n) \approx \frac{c_n^2}{2} >> \frac{1}{n}$, which is the risk of the MLE, provided $\gamma_n = \sqrt{n}c_n \to \infty$. If we move slightly to the left of $\theta = c_n$, then the risk increases even more. Precisely, if we take $\theta = (1-\alpha_n)c_n$ with a very small α_n, then $R(\theta, S_n) \approx c_n^2$. We believe that this is the exact rate of convergence of the global maximum of the risk,* i.e.,

$$
\lim_{n\to\infty} c_n^{-2} \sup_{-\infty<\theta<\infty} R(\theta, S_n) = 1. \qquad (4.31)
$$

4.2.3.2 Global Maximum of the Risk and Point of Maxima

Corollary 1.2.6 suggests a pathway to addressing the two related questions: what is an approximation to the point at which the global maximum of the risk is attained, and what is a higher order approximation to the value of the global maximum. In Eq. (4.31), if we use the two leading terms $(1-\alpha_n)^2 - \frac{\phi(\alpha_n\gamma_n)}{\alpha_n\gamma_n}$, we notice that $(1-\alpha)^2$ and $\frac{\phi(\alpha\gamma_n)}{\alpha\gamma_n}$ are both decreasing in α. Therefore, if we maximize $(1-\alpha)^2 - \frac{\phi(\alpha\gamma_n)}{\alpha\gamma_n}$ over α (in $(0,1)$), it will give us an approximation to the global maximum of $R(\theta, S_n)$ and at the same time, an approximation to the point $\theta_n = (1-\alpha_n)c_n$ where the maximum is attained. *It must be understood that these two approximations are heuristic, because we do not have a proof that* $\sup_\theta R(\theta, S_n)$ *is attained at a point of the form* $(1-\alpha_n)c_n$ *with* α_n *as in Corollary 1.2.6.*

To maximize $(1-\alpha)^2 - \frac{\phi(\alpha\gamma_n)}{\alpha\gamma_n}$, we want to find the root of

$$0 = \frac{d}{d\alpha}\Big[(1-\alpha)^2 - \frac{\phi(\alpha\gamma_n)}{\alpha\gamma_n}\Big]$$

$$= 2(\alpha-1) + \phi(\alpha\gamma_n)\Big[\gamma_n + \frac{1}{\alpha\gamma_n}\Big] = 2(\alpha-1) + \gamma_n\phi(\alpha\gamma_n) + 0(\gamma_n\phi(\alpha\gamma_n))$$

$$\Rightarrow (1-\alpha) = \frac{\gamma_n}{2}\phi(\alpha\gamma_n)(1+0(1))$$

$$\Rightarrow -\alpha = \log\gamma_n - \frac{\alpha^2\gamma_n^2}{2} + O(1)$$

$$\Rightarrow \alpha^2\gamma_n^2 - 2\alpha - 2\log\gamma_n + O(1) = 0 \tag{4.32}$$

An approximation to the root of the quadratic Eq. (4.32) is

$$\alpha = \frac{\sqrt{2\log\gamma_n}}{\gamma_n}, \tag{4.33}$$

which results in the following two heuristic approximations:

Conjecture *In the class of estimates*

$$S_n(X_1,\ldots,X_n) = \begin{array}{lll} \bar{X}_n & \text{if} & |\bar{X}_n| > c_n \\ 0 & \text{if} & |\bar{X}_n| \le c_n \end{array}, \tag{4.34}$$

one has,

$$\text{argmax}_{-\infty<\theta<\infty} R(\theta, S_n) \approx c_n - \sqrt{\frac{\log(nc_n^2)}{n}}; \quad \sup_{-\infty<\theta<\infty} R(\theta, S_n) \approx c_n^2 - \frac{2c_n\sqrt{\log n}}{\sqrt{n}}. \tag{4.35}$$

Example 1.2.1 We look at the credibility of (4.35) for Hodges' original estimate T_n, for which $c_n = n^{-1/4}$. In that case, (4.35) says that the global maximum of $R(\theta, T_n)$ should be approximately $\frac{1}{\sqrt{n}} - \frac{2\sqrt{\log n}}{n^{3/4}}$ and it should be attained at $\theta_n \approx n^{-1/4} - \sqrt{\frac{\log n}{2n}}$. We show in the following table the exact global maximum (computed numerically), the risk at c_n and at θ_n and the approximation to the maximum risk as claimed in (4.35). For very large n, our conjecture appears to work out almost exactly. Otherwise, it does not.

n	Exact Maximum	$R(c_n, T_n)$	$R(\theta_n, T_n)$	Approx. (1.35)
100	0.0558	0.0550	0.0112	0.0357
2500	0.0126	0.0102	0.0073	0.0042
100000	0.0025	0.0016	0.0021	0.0020
250000	0.0016	0.0010	0.0014	0.0014
10^6	0.0008	0.0005	0.0008	0.0008

4.2.3.3 Optimal Thresholding

The approximation laid out in (4.35) enables us to pose and give a solution to another relevant question: what is an optimal choice of the thresholding parameter (sequence) c_n? Obviously, this calls for a definition of optimal thresholding. We adopt the definition of controlled minimaxity. Here is an explanation, and then a formal mathematical definition.

It is clear that the choice of the thresholding parameter affects two key quantities in the problem, the risk at zero, and the maximum risk. For instance, as an extreme, if we choose $c_n = 0$, then the risk at zero is zero, but the maximum risk is infinity. Thus, there is a trade-off between $R(0, S_n)$ and $\sup_\theta R(\theta, S_n)$, and the thresholding parameter c_n influences both of them, but in opposite directions. It seems reasonable to ask for the sequence c_n that minimizes $\sup_\theta R(\theta, S_n)$ subject to a guaranteed percentage improvement in risk over the MLE at $\theta = 0$. More precisely, the question is: which sequence c_n minimizes $\sup_\theta R(\theta, S_n)$ subject to the constraint $n|e_n(0)| \geq 1 - \epsilon_n$, where, $e_n(\theta) = R(\theta, S_n) - \frac{1}{n}$. Thus, in this formulation we seek the thresholding estimate that is minimax subject to a risk gain of at least $100(1 - \epsilon_n)\%$ at zero; ϵ_n is supposed to be user provided. Such restricted minimax formulations have been proposed and studied in other problems before; one reference is Bickel (1983).

From (4.9) and (4.35), we wish to

$$\text{minimize } \gamma_n^2 - 2\gamma_n\sqrt{\log n} \text{ subject to } H(\gamma_n) = \Phi(\gamma_n) - \frac{1}{2} - \gamma_n\phi(\gamma_n) \geq \frac{1-\epsilon_n}{2}$$

The unconstrained minimum of $\gamma_n^2 - 2\gamma_n\sqrt{\log n}$ is $\gamma_n = \sqrt{\log n}$. If $H(\sqrt{\log n}) \geq \frac{1-\epsilon_n}{2}$ (which approximately corresponds to $\epsilon_n \geq \frac{1}{\sqrt{n}}$, then the solution to our problem is $\gamma_n = \sqrt{\log n}$. Otherwise, since $H(x)$ is increasing in x for positive x, i.e.,

increasing in x for $x > 0$, it follows that the sequence γ_n that solves the constrained minimum problem is the root of the equation

$$\Phi(\gamma_n) - \frac{1}{2} - \gamma_n \phi(\gamma_n) = \frac{1-\epsilon_n}{2} \tag{4.36}$$

$$\Leftrightarrow 1 - \Phi(\gamma_n) + \gamma_n \phi(\gamma_n) = \frac{\epsilon_n}{2}$$

$$\Leftrightarrow \phi(\gamma_n)\Big[\gamma_n + O(\frac{1}{\gamma_n})\Big] = \frac{\epsilon_n}{2}$$

$$\Leftrightarrow \sqrt{\frac{\pi}{2}} e^{\frac{\gamma_n^2}{2}} \frac{\gamma_n}{\gamma_n^2 + O(1)} = \frac{1}{\epsilon_n} \tag{4.37}$$

A first approximation to the root of (4.36) is $\gamma_n = \sqrt{2 \log \frac{1}{\epsilon_n}}$. Plugging the first approximation back into (4.36), a higher order approximation is

$$\gamma_n^2 = 2 \log \frac{1}{\epsilon_n} + 2 \log \left(\sqrt{2 \log \frac{1}{\epsilon_n}} \right) = 2 \log \frac{1}{\epsilon_n} + \log \log \frac{1}{\epsilon_n} + O(1),$$

which gives

$$\gamma_n = \sqrt{2 \log \frac{1}{\epsilon_n} + \log \log \frac{1}{\epsilon_n} + O(1)} = \sqrt{2 \log \frac{1}{\epsilon_n}} \Big[1 + \frac{\log \log \frac{1}{\epsilon_n}}{4 \log \frac{1}{\epsilon_n}} + o\left(\frac{\log \log \frac{1}{\epsilon_n}}{\log \frac{1}{\epsilon_n}} \right) \Big]$$

$$= \sqrt{2 \log \frac{1}{\epsilon_n}} + \frac{\log \log \frac{1}{\epsilon_n}}{2\sqrt{2 \log \frac{1}{\epsilon_n}}} + o\left(\frac{\log \log \frac{1}{\epsilon_n}}{\sqrt{\log \frac{1}{\epsilon_n}}} \right)$$

We propose finally the following thresholding sequence:

$$\gamma_n = \sqrt{n} c_n = \sqrt{\log n}, \quad \text{if } \epsilon_n \geq \frac{1}{\sqrt{n}},$$

$$\gamma_n = \sqrt{n} c_n = \sqrt{2 \log \frac{1}{\epsilon_n}} + \frac{\log \log \frac{1}{\epsilon_n}}{2\sqrt{2 \log \frac{1}{\epsilon_n}}}, \quad \text{if } \epsilon_n < \frac{1}{\sqrt{n}} \tag{4.38}$$

Example 1.2.2 The recommended thresholding sequence in (4.38) depends on the specification of ϵ_n. We work out the form of c_n for four choices of ϵ_n. Suppose,

independent of n, we want a fixed percentage risk improvement $100(1-\epsilon)\%$ at zero. Then, $\epsilon_n \equiv \epsilon$, which, by (4.38), leads to

$$c_n = \sqrt{\frac{\log n}{n}}$$

Thus, a fixed percentage risk improvement at zero leads to $c_n \sim \sqrt{\frac{\log n}{n}}$.

Suppose we want the percentage risk improvement at zero to increase with n at a polynomial rate, $\epsilon_n = n^{-\beta}, \beta > \frac{1}{2}$. Then, (4.38) leads to

$$c_n = \frac{\sqrt{2\beta \log n}}{\sqrt{n}} + \frac{\log\log n}{2\sqrt{2\beta n \log n}} + O(\frac{1}{\sqrt{n \log n}}).$$

Thus, for polynomial growth in the percentage risk improvement at zero, still, the recommended thresholding sequence $c_n \sim \sqrt{\frac{\log n}{n}}$, but with a constant in front that is > 1.

Next, suppose we want the percentage risk improvement at zero to increase at a subexponential rate, namely, $\epsilon_n = e^{-\beta\sqrt{n}}, \beta > 0$. Then, (4.38) leads to

$$c_n = \sqrt{2\beta} n^{-1/4} + \frac{\log n}{4\sqrt{2\beta} n^{3/4}}.$$

Thus, for subexponential growth in the percentage risk improvement at zero,we get $c_n \sim n^{-1/4}$. *Compare this with Eq. (4.11) which describes the percentage risk improvement at zero of Hodges' original estimate T_n. Interestingly, his choice of $c_n = n^{-1/4}$ matches to the first order the recommended sequence we just derived above.*

Finally, suppose we want the percentage risk improvement at zero to increase at the fully exponential rate, namely, $\epsilon_n = e^{-\beta n}, \beta > 0$. Then, (4.38) leads to

$$c_n = \sqrt{2\beta} + \frac{\log n}{2\sqrt{2\beta} n}.$$

Thus, for exponential growth in the percentage risk improvement, we get $c_n \sim c$, a constant.

4.2.4 Comparison of Bayes Risks and Regular Variation

Since the risk functions of the MLE and thresholding estimates S_n cross, it is meaningful to seek a comparison between them by using Bayes risks. Because of the intrinsic specialty of the point $\theta = 0$ in this entire problem, it is sensible to consider priors that are symmetric about zero. Purely for technical convenience, we only consider normal priors here, $N(0, \sigma_n^2)$, and we ask the following question: how should σ_n behave for the thresholding estimate to have (asymptotically) a smaller Bayes risk

than the MLE? It turns out that certain interesting stories emerge in answering the question, and we have a fairly complete answer to the question we have posed.

We start with some notation. Let $\pi = \pi_n$ denote a prior density and $B_n(S_n, \pi)$ the Bayes risk of S_n under π. Let also $B_n(\pi)$ denote the Bayes risk of the Bayes rule under π. Then,

$$B_n(S_n, \pi) = \int R(\theta, S_n)\pi(\theta)d\theta = \frac{1}{n} + \int e_n(\theta)\pi(\theta)d\theta \tag{4.39}$$

and

$$B_n(\pi) = \frac{1}{n} - \frac{1}{n^2}\int \frac{(m'(x))^2}{m(x)}dx, \tag{4.40}$$

where $m(x) = m_n(x)$ denotes the marginal density of $\bar{X}$ under π. In the case where $\pi = \pi_n$ is the $N(0, \sigma_n^2)$ density, $B_n(\pi) = \frac{\sigma_n^2}{n\sigma_n^2+1}$.

4.2.4.1 Normal Priors

We use (4.5) to write a closed form formula for $B_n(S_n, \pi)$; it is assumed until we specifically mention otherwise that henceforth $\pi = N(0, \sigma_n^2)$, and for brevity, we drop the subscript and write σ^2 for σ_n^2.

Toward this agenda, the following formulas are used; for reasons of space, we will not provide their derivations.

$$\int \Phi(\sqrt{n}(c_n \pm \theta))\frac{1}{\sigma}\phi(\frac{\theta}{\sigma})d\theta = \Phi(\frac{\sqrt{n}c_n}{\sqrt{1+n\sigma^2}}) \tag{4.41}$$

$$\int \phi(\sqrt{n}(c_n \pm \theta))\frac{1}{\sigma}\phi(\frac{\theta}{\sigma})d\theta = \frac{\sigma e^{-nc_n^2/(2(1+n\sigma^2))}}{\sqrt{2\pi}\sqrt{1+n\sigma^2}} \tag{4.42}$$

$$\int \theta\phi(\sqrt{n}(c_n \pm \theta))\frac{1}{\sigma}\phi(\frac{\theta}{\sigma})d\theta = \mp\frac{\sigma^2 nc_n e^{-nc_n^2/(2(1+n\sigma^2))}}{\sqrt{2\pi}(1+n\sigma^2)^{3/2}} \tag{4.43}$$

$$\int \theta^2\Phi(\sqrt{n}(c_n \pm \theta))\frac{1}{\sigma}\phi(\frac{\theta}{\sigma})d\theta = \sigma^2\Big[\Phi(\frac{\sqrt{n}c_n}{\sqrt{1+n\sigma^2}}) - \frac{\sigma^2 n^{3/2}c_n e^{-nc_n^2/(2(1+n\sigma^2))}}{\sqrt{2\pi}(1+n\sigma^2)^{3/2}}\Big] \tag{4.44}$$

By plugging (4.41), (4.42), (4.43), (4.44) into $\int e_n(\theta)\frac{1}{\sigma}\phi(\frac{\theta}{\sigma})d\theta$, where the expression for $e_n(\theta)$ is taken from (4.5), additional algebraic simplification gives us the following closed form expression.

Theorem 1.2.7

$$B_n(S_n, \pi) = \frac{1}{n} + \int e_n(\theta)\pi(\theta)d\theta,$$

with

$$\begin{aligned}\int e_n(\theta)\pi(\theta)d\theta = \frac{1-a_n^2}{n} - (1-a_n)^2\sigma^2 \\ + \left[2(1-a_n)^2\sigma^2 - \frac{2(1-a_n^2)}{n}\right]\Phi\left(\frac{\sqrt{n}c_n}{\sqrt{1+n\sigma^2}}\right) \\ - \frac{\sqrt{n}c_n}{\sqrt{1+n\sigma^2}}\phi\left(\frac{\sqrt{n}c_n}{\sqrt{1+n\sigma^2}}\right)\left[\frac{2n(1-a_n)^2\sigma^4}{1+n\sigma^2}\right. \\ \left. + \frac{2(1-a_n)^2\sigma^2}{1+n\sigma^2} - \frac{2(1-a_n^2)}{n}\right] \qquad (4.45)\end{aligned}$$

Theorem 1.2.7 leads to the following more transparent corollary.

Corollary 1.2.7 Consider the generalized Hodges estimate S_n with $a_n \equiv 0$. Then

$$\int e_n(\theta)\pi(\theta)d\theta = \frac{n\sigma^2-1}{n}\left[2\Phi\left(\frac{\gamma_n}{\sqrt{1+n\sigma^2}}\right) - \frac{1+n\sigma^2}{1+\sigma^2}\frac{\gamma_n}{\sqrt{1+n\sigma^2}}\phi\left(\frac{\gamma_n}{\sqrt{1+n\sigma^2}}\right)\right] \qquad (4.46)$$

In particular, if $\sigma^2 = \frac{1}{n}$, then whatever be the thresholding sequence c_n, $B_n(S_n, \pi) = \frac{1}{n}$, i.e., S_n and the MLE $\bar{X}$ have the same Bayes risk if $\theta \sim N(0, \frac{1}{n})$. By inspecting (4.46), we can make more general comparisons between $B_n(S_n, \pi)$ and $\frac{1}{n} = B_n(\bar{X}, \pi)$ when $\sigma^2 \neq \frac{1}{n}$. It turns out that $\sigma^2 = \frac{1}{n}$ acts in a very meaningful sense as a boundary between $B_n(S_n, \pi) < B_n(\bar{X}, \pi)$ and $B_n(S_n, \pi) > B_n(\bar{X}, \pi)$. We will now make it precise. In this analysis, it will be useful to note that once we know whether $\sigma^2 >$ or $< \frac{1}{n}$, by virtue of formula (4.46), the algebraic sign of $\triangle_n(\pi) = B_n(S_n, \pi) - B_n(\bar{X}, \pi)$ is determined by the algebraic sign of $\eta_n = 2\Phi(\frac{\gamma_n}{\sqrt{1+n\sigma^2}}) - \frac{1+n\sigma^2}{1+\sigma^2}\frac{\gamma_n}{\sqrt{1+n\sigma^2}}\phi(\frac{\gamma_n}{\sqrt{1+n\sigma^2}})$.

Theorem 1.2.8 *Provided the thresholding sequence c_n satisfies $c_n \to 0, \gamma_n = \sqrt{n}c_n \to \infty$,*

(a) $\triangle_n(\pi) < 0$ for all large n if $\sigma^2 = \frac{c}{n} + o(\frac{1}{n})$ for some $c, 0 \leq c < 1$.
(b) $\triangle_n(\pi) > 0$ for all large n if $\sigma^2 = \frac{c}{n} + o(\frac{1}{n})$ for some $c, c > 1$.
(c) $\triangle_n(\pi) = 0$ for all n if $\sigma^2 = \frac{1}{n}$.
(d) If $n\sigma^2 \to 1$, then in general $\triangle_n(\pi)$ oscillates around zero.
(e) If $n\sigma^2 \to \infty$, then $\triangle_n(\pi) < 0$ for all large n.

Proof We indicate the proof of part (a). In this case, $n\sigma^2 - 1 < 0$ for all large n. On the other hand,

$$\Phi(\frac{\gamma_n}{\sqrt{1+n\sigma^2}}) \to 1;\ \frac{1+n\sigma^2}{1+\sigma^2} \to 1+c;\ \frac{\gamma_n}{\sqrt{1+n\sigma^2}}\phi(\frac{\gamma_n}{\sqrt{1+n\sigma^2}}) \to 0.$$

Therefore, $\eta_n \to 1 > 0$, and hence, for all all large n, $\triangle_n(\pi) < 0$. The other parts use the same line of argument and so we do not mention them.

4.2.4.2 General Smooth Priors

We now give an asymptotic expansion for $\triangle_n = B_n(S_n, \pi) - B_n(\bar{X}, \pi)$ for general smooth prior densities of the form $\pi(\theta) = \pi_n(\theta) = \sqrt{n}h(\theta\sqrt{n})$, where h is a fixed sufficiently smooth density function on $(-\infty, \infty)$. It will be seen below that scaling by $\sqrt{n}$ is the right scaling to do in π_n, similar to our finding that in the normal case, $\sqrt{n}\,\theta \sim N(0,1)$ acts as a boundary between $\triangle_n < 0$ and $\triangle_n > 0$. We introduce the following notation

$$q(z) = \int_0^z (t^2 - 1)h(t)dt - h'(z), -\infty < z < \infty;\ \ w(z) = -\frac{d}{dz}\log q(z). \quad (4.47)$$

The functions $q(z)$ and $\log q(z)$ will play a pivotal role in the three main results below, Theorem 1.2.9, Proposition 1.2.1, and Theorem 1.2.10. Note that $q(z) \equiv 0$ if $h = \phi$, the standard normal density. For general h, q can take both positive and negative values, and this will complicate matters in the analysis that follows.

We will need the following assumptions on h and q. *Not all of the assumptions are needed for every result below*. But we find it convenient to list all the assumptions together, at the expense of some generality.

Assumptions on h

(1) $h(z) < \infty \forall z$.
(2) $h(-z) = h(z) \forall z$.
(3) $\int_{-\infty}^{\infty} z^2\, h(z)dz < \infty$.
(4) h is twice continuously differentiable, and $h'(z) \to 0$ as $z \to \infty$.
(5) q is ultimately decreasing and positive.
(6) $\log q$ is absolutely continuous, ultimately negative, and ultimately concave or convex.
(7) $\liminf_{z\to\infty} \frac{d}{dz}\log q(z) > -\infty$.

The first result below, Theorem 1.2.9, is on a unified convolution representation and some simple asymptotic order results for the Bayes risk difference $\triangle_n = B_n(S_n, \pi) - B_n(\bar{X}, \pi)$. A finer result on the asymptotic order of $\triangle_n$ is the content of Theorem 1.2.10. *In the result below, (4.49) and (4.50) together say that the first order behavior of $\triangle_n$ is determined by whether or not $Var_h(\theta) = 1$. If $Var_h(\theta) \neq 1$, then $\triangle_n$ converges at the rate $\frac{1}{n}$; but if $Var_h(\theta) = 1$, then $\triangle_n$ converges at a rate faster than $\frac{1}{n}$. This provides greater insight into the result of part (c) of Theorem 1.2.8.*

Theorem 1.2.9 *Consider generalized Hodges estimates S_n of the form (1.2) with $a_n \equiv 0$. Let h be a fixed density function satisfying the assumptions (1)-(4) above and let $\pi(\theta) = \pi_n(\theta) = \sqrt{n}h(\theta\sqrt{n}), -\infty < \theta < \infty$. Then we have the identity*

$$\begin{aligned}\triangle_n = \frac{2}{n}(q * \phi)(\gamma_n) &= \frac{2}{n}\int_{-\infty}^{\infty} q(z)\phi(\gamma_n - z)dz \\ &= \frac{2}{n}\int_0^{\infty} q(z)\Big[\phi(\gamma_n - z) - \phi(\gamma_n + z)\Big]dz \qquad (4.48)\end{aligned}$$

In particular, if $q \in \mathcal{L}_1$, then

$$n\triangle_n \to 0, \text{ i.e., } \triangle_n = o(\frac{1}{n}), \qquad (4.49)$$

and if $q(z) \to c \neq 0$ as, $z \to \infty$, then

$$n\triangle_n \to 2c, \text{ i.e., } \triangle_n = \frac{2c}{n} + o(\frac{1}{n}). \qquad (4.50)$$

In any case, if $\mathrm{Var}_h(\theta) < \infty$, and $h' \in \mathcal{L}_\infty$, then, for every fixed n,

$$|n\triangle_n| \leq 1 + \mathrm{Var}_h(\theta) + ||h'||_\infty. \qquad (4.51)$$

Proof Using (4.5) and the definition of $\pi(\theta)$,

$$\begin{aligned}\triangle_n &= \int_{-\infty}^{\infty} e_n(\theta)\pi_n(\theta)d\theta \\ &= \int_{-\infty}^{\infty} (\theta^2 - \frac{1}{n})\Big[\Phi(\gamma_n + \theta\sqrt{n}) + \Phi(\gamma_n - \theta\sqrt{n}) - 1\Big]\sqrt{n}h(\theta\sqrt{n})d\theta \\ &\quad + \frac{1}{\sqrt{n}}\int_{-\infty}^{\infty}\Big[(c_n + \theta)\phi(\gamma_n + \theta\sqrt{n}) + (c_n - \theta)\phi(\gamma_n - \theta\sqrt{n})\Big]\sqrt{n}h(\theta\sqrt{n})d\theta \\ &= \frac{1}{n}\Big(\int_{-\infty}^{\infty}(z^2 - 1)\Big[\Phi(\gamma_n + z) + \Phi(\gamma_n - z) - 1\Big]h(z)dz \\ &\quad + \int_{-\infty}^{\infty}\Big[(\gamma_n + z)\phi(\gamma_n + z) + (\gamma_n - z)\phi(\gamma_n - z)\Big]h(z)dz\Big) \\ &= \frac{1}{n}\Big(\int_{-\infty}^{\infty}(z^2 - 1)\Big[2\Phi(\gamma_n + z) - 1\Big]h(z)dz + 2\int_{-\infty}^{\infty}(\gamma_n + z)\phi(\gamma_n + z)h(z)dz\Big) \\ &= \frac{2}{n}\Big(\int_{-\infty}^{\infty}(z^2 - 1)h(z)\Phi(\gamma_n + z)dz + \int_{-\infty}^{\infty}(\gamma_n + z)\phi(\gamma_n + z)h(z)dz\Big) \\ &\quad - \frac{1}{n}\int_{-\infty}^{\infty}(z^2 - 1)h(z)dz \\ &= \frac{2}{n}\Big(\int_{-\infty}^{\infty}(z^2 - 1)h(z)\Phi(\gamma_n + z)dz - \int_{-\infty}^{\infty}\Phi(\gamma_n + z)h''(z)dz\Big) \\ &\quad - \frac{1}{n}\int_{-\infty}^{\infty}(z^2 - 1)h(z)dz\end{aligned}$$

(by twice integrating by parts the integral $\int_{-\infty}^{\infty}(\gamma_n+z)\phi(\gamma_n+z)h(z)dz$)

$$
\begin{aligned}
&= \frac{2}{n}\int_{-\infty}^{\infty}\Big[(z^2-1)h(z)-h''(z)\Big]\Phi(\gamma_n+z)dz - \frac{1}{n}\int_{-\infty}^{\infty}(z^2-1)h(z)dz\\
&= \frac{2}{n}\int_{-\infty}^{\infty}q'(z)\Phi(\gamma_n+z)dz - \frac{1}{n}\int_{-\infty}^{\infty}(z^2-1)h(z)dz\\
&= \frac{2}{n}\Big(q(z)\Phi(\gamma_n+z)|_{-\infty}^{\infty} - \int_{-\infty}^{\infty}q(z)\phi(\gamma_n+z)dz\Big) - \frac{1}{n}\int_{-\infty}^{\infty}(z^2-1)h(z)dz\\
&= \frac{2}{n}\int_{0}^{\infty}(z^2-1)h(z)dz - \frac{2}{n}\int_{-\infty}^{\infty}q(z)\phi(\gamma_n+z)dz - \frac{1}{n}\int_{-\infty}^{\infty}(z^2-1)h(z)dz
\end{aligned}
$$

(refer to (4.47))

$$
\begin{aligned}
&= -\frac{2}{n}\int_{-\infty}^{\infty}q(z)\phi(\gamma_n+z)dz\\
\Big(\text{since } 2\int_{0}^{\infty}(z^2-1)h(z)dz &= \int_{-\infty}^{\infty}(z^2-1)h(z)dz\Big)\\
&= \frac{2}{n}\int_{-\infty}^{\infty}q(z)\phi(\gamma_n-z)dz\\
&= \frac{2}{n}\int_{0}^{\infty}q(z)\Big[\phi(\gamma_n-z)-\phi(\gamma_n+z)\Big]dz \qquad (4.52)
\end{aligned}
$$

(since $q(-z)=-q(z)$ for all z), and this gives (4.48). (4.49), (4.50), and (4.51) follow on application of the dominated convergence theorem and the triangular inequality, and this establishes the theorem.

Remark Eq. (4.48) is a pleasant general expression for the Bayes risk difference $\triangle_n$ and what is more, has the formal look of a convolution density. One might hope that techniques from the theory of convolutions can be used to assert useful things about the asymptotic behavior of $\triangle_n$, via (4.48). We will see that indeed this is the case.

Before embarking on further analysis of $\triangle_n$, we need to keep two things in mind. First, the function $q(z)$ is usually a signed function and, therefore, we are not dealing with convolutions of probability measures in (4.48). This adds a bit of additional complexity into the analysis. Second, it does not take too much to fundamentally change the asymptotic behavior of $\triangle_n$. In the two pictures below, we have plotted $\int_0^{\infty}q[z]\Big[\phi(\gamma-z)-\phi(\gamma+z)\Big]dz$, for two different choices of the (probability density) function h. In the first picture, h is a standard Laplace (double exponential) density, while in the second picture, h is a Laplace density scaled to have variance exactly equal to 1. We can see that just a scale change changes both the asymptotic (in γ) sign and shape of $\triangle_n$ (refer to (4.49) and (4.50) as well). Thus, in our further analysis of $\triangle_n$ by exploiting the formula in (4.48), we must remain mindful of small changes in h that can make big changes in (4.48).

For future reference, we record the following formula.

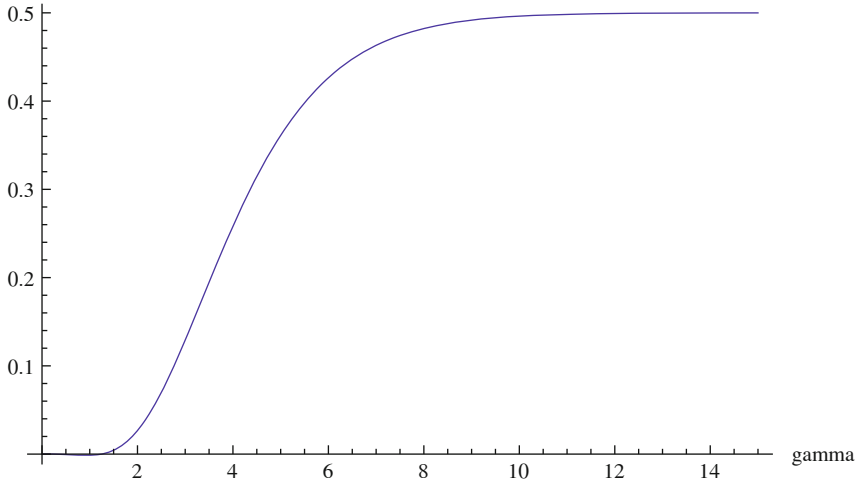


Fig. 4.4 Plot of (48) for a Standard Double Exponential h

If $h(t) = \frac{1}{2\sigma} e^{-|t|/\sigma}$, then (for $z > 0$),

$$q(z) = \sigma^2 - \frac{1}{2} + (\alpha_0 + \alpha_1 z + \alpha_2 z^2) e^{-z/\sigma}, \tag{4.53}$$

where

$$\alpha_0 = \frac{1}{2} + \frac{1}{2\sigma^2} - \sigma^2, \quad \alpha_1 = -\sigma, \quad \alpha_2 = -\frac{1}{2}$$

Thus, if $2\sigma^2 \neq 1$, then q acts asymptotically like a nonzero constant; but if $2\sigma^2 = 1$, then asymptotically q dies. This affects the asymptotic sign and shape of the convolution expression (4.48), and explains why the two pictures below look so different. Fig. 4.4 and 4.5

The next technical proposition will be useful for our subsequent analysis of (4.48) and $\triangle_n$. For this proposition, we need two special functions.

For $-\infty < p < \infty$, by $D_p(z)$ we denote the *parabolic cylinder function* which solves the differential equation $u'' + (p + \frac{1}{2} - \frac{z^2}{4})u = 0$. For $-\infty < a < \infty$ and $c \neq 0, -1, -2, \cdots$, $M(a, c, z)$ (also often written as ${}_1F_1(a, c, z)$) denotes the *confluent hypergeometric function* $\sum_{k=0}^{\infty} \frac{(a)_k}{(c)_k} \frac{z^k}{k!}$. We have the following proposition.

Proposition 1.2.1 *Let $k \geq 0$ be an integer and a a nonnegative real number. Then, for any real number μ,*

$$\int_0^\infty z^k e^{-az} \phi(\mu - z) dz = \frac{k! e^{-\mu^2/2}}{2^{k/2+1}} \Big[\frac{M(\frac{k+1}{2}, \frac{1}{2}, \frac{(\mu-a)^2}{2})}{\Gamma(\frac{k+2}{2})} + \sqrt{2}(\mu - a) \frac{M(\frac{k+2}{2}, \frac{3}{2}, \frac{(\mu-a)^2}{2})}{\Gamma(\frac{k+1}{2})} \Big] \tag{4.54}$$

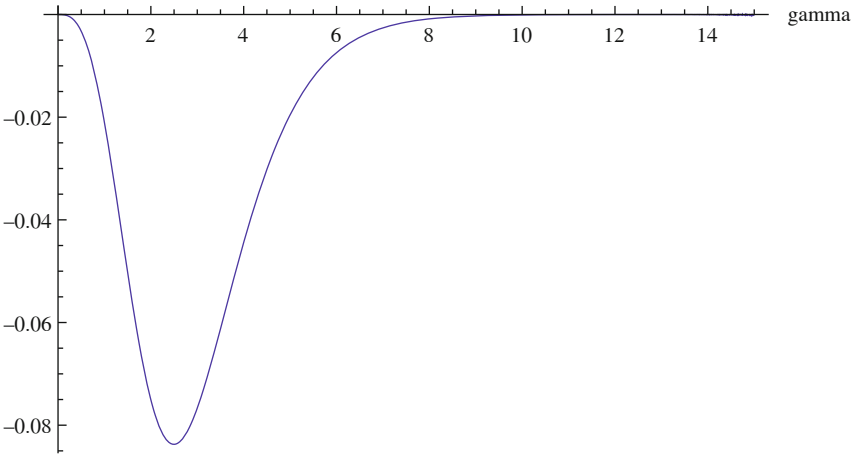


Fig. 4.5 Plot of (48) for a Scaled Double Exponential h

and, as $\gamma \to \infty$,

$$\int_0^\infty z^k e^{-az}\Big[\phi(\gamma - z) - \phi(\gamma + z)\Big]dz \sim e^{a^2/2}e^{-a\gamma}\gamma^k, \tag{4.55}$$

(in the sense that the ratio of the two sides converges to 1 as $\gamma \to \infty$)

Proof To obtain (4.54), write for any real number μ,

$$\int_0^\infty z^k e^{-az}\phi(\mu - z)dz = \frac{e^{-\mu^2/2}}{\sqrt{2\pi}}\int_0^\infty z^k e^{(\mu-a)z-z^2/2}dz, \tag{4.56}$$

and first, use the integration formula

$$\int_0^\infty z^k e^{-bz-z^2/2}dz = k!e^{b^2/4}D_{-k-1}(b) \tag{4.57}$$

(pp 360, Gradshteyn and Ryzhik (1980)) Next, use the functional identity

$$D_p(z) = 2^{p/2}e^{-z^2/4}\Big[\frac{\sqrt{\pi}}{\Gamma(\frac{1-p}{2})}M(-\frac{p}{2},\frac{1}{2},\frac{z^2}{2}) - \frac{\sqrt{2\pi}z}{\Gamma(-\frac{p}{2})}M(\frac{1-p}{2},\frac{3}{2},\frac{z^2}{2})\Big] \tag{4.58}$$

(pp 1018, Gradshteyn and Ryzhik (1980))

Substituting (4.57) and (4.58) into (4.56), we get (4.54), on careful algebra.

For (4.55), we use the asymptotic order result

$$M(\alpha,\beta,z) \sim e^z z^{\alpha-\beta}\frac{\Gamma(\beta)}{\Gamma(\alpha)}, \quad z \to \infty \tag{4.59}$$

(see, for example, pp 255-259 in Olver (1997))

Use of (4.59) in (4.54) with $\mu = \mp\gamma$, and then subtraction, leads to the asymptotic order result that as $\gamma \to \infty$,

$$\int_0^\infty z^k e^{-az}\Big[\phi(\gamma - z) - \phi(\gamma + z)\Big]dz = \frac{k!e^{a^2/2}}{2^{k/2+1}}$$
$$\times\Big\{e^{-a\gamma}(\frac{(\gamma-a)^2}{2})^{k/2}\frac{\sqrt{\pi}}{\Gamma(\frac{k+1}{2})\Gamma(\frac{k+2}{2})}$$
$$+\sqrt{2}(\gamma-a)e^{-a\gamma}(\frac{(\gamma-a)^2}{2})^{k/2-1/2}\frac{\frac{1}{2}\sqrt{\pi}}{\Gamma(\frac{k+1}{2})\Gamma(\frac{k+2}{2})}\Big\}\times(1+o(1))$$
$$+\Big\{\sqrt{2}(\gamma+a)e^{a\gamma}(\frac{(\gamma+a)^2}{2})^{k/2-1/2}\frac{\frac{1}{2}\sqrt{\pi}}{\Gamma(\frac{k+1}{2})\Gamma(\frac{k+2}{2})}$$
$$-e^{a\gamma}(\frac{(\gamma+a)^2}{2})^{k/2}\frac{\sqrt{\pi}}{\Gamma(\frac{k+1}{2})\Gamma(\frac{k+2}{2})}\Big\}\times(1+o(1)) = \frac{k!e^{a^2/2}\sqrt{\pi}}{2^{k+1/2}\Gamma(\frac{k+1}{2})\Gamma(\frac{k+2}{2})}$$
$$\Big[e^{-a\gamma}\frac{(\gamma-a)^k}{\sqrt{2}} + e^{-a\gamma}\frac{(\gamma-a)^k}{\sqrt{2}} - e^{a\gamma}\frac{(\gamma+a)^k}{\sqrt{2}} + e^{a\gamma}\frac{(\gamma+a)^k}{\sqrt{2}}\Big]\times(1+o(1))$$
$$= \frac{k!e^{a^2/2}\sqrt{\pi}}{2^k\Gamma(\frac{k+1}{2})\Gamma(\frac{k+2}{2})}e^{-a\gamma}(\gamma-a)^k\times(1+o(1)) \tag{4.60}$$

In (4.60), by using the *Gamma duplication formula*

$$\Gamma(z+1/2) = \sqrt{\pi}2^{1-2z}\frac{\Gamma(2z)}{\Gamma(z)},$$

we get

$$\int_0^\infty z^k e^{-az}\Big[\phi(\gamma - z) - \phi(\gamma + z)\Big]dz$$
$$= e^{a^2/2}e^{-a\gamma}(\gamma-a)^k\times(1+o(1)) = e^{a^2/2}e^{-a\gamma}\gamma^k\times(1+o(1)), \tag{4.61}$$

as claimed in (4.55).

Remark The real use of Proposition 1.2.1 is that by using (4.54), we get an *exact analytical formula* for $\triangle_n$ in terms of the confluent hypergeometric function. If all we care for is the asymptotic order result (4.55), then we may obtain it in a less complex way. Indeed, by using techniques in Feller (1971, pp 442-446) and Theorem 3.1 in Berman (1992), we can conclude that $\int_0^\infty z^k e^{-az}\phi(\gamma - z)dz = \gamma^k e^{-a\gamma}\int_{-\infty}^\infty e^{(a-\frac{k}{\gamma})t}\phi(t)dt \times(1+o(1))$, and (4.55) follows from this.

Corollary 1.2.8 *Consider generalized Hodges estimates of the form (4.2) with* $a_n \equiv 0$. *Let* $h(\theta) = \frac{1}{2\sigma} e^{-|\theta|/\sigma}$ *and* $\pi(\theta) = \pi_n(\theta) = \sqrt{n}h(\theta\sqrt{n})$. *Then,*

$$\triangle_n = \frac{2\sigma^2 - 1}{n}(1 + o(1)), \quad \text{if } \mathrm{Var}_h(\theta) \neq 1 \Leftrightarrow 2\sigma^2 - 1 \neq 0, \tag{4.62}$$

and,

$$\triangle_n = -e\frac{\gamma_n^2 e^{-\gamma_n\sqrt{2}}}{n}(1 + o(1)), \quad \text{if } \mathrm{Var}_h(\theta) = 1 \Leftrightarrow 2\sigma^2 - 1 = 0 \tag{4.63}$$

This corollary follows by using the formula in (4.53) and the result in (4.55). *Notice that the critical issue in determining the rate of convergence of* $\triangle_n$ *to zero is whether or not* $Var_h(\theta) = 1$.

As indicated previously, we can generalize the result on the asymptotic order of the Bayes risk difference $\triangle_n$ to more general priors. The important thing to understand is that Theorem 1.2.9 (more precisely, (4.48)) gives a representation of $\triangle_n$ in a convolution form. Hence, we need to appeal to results on orders of the tails of convolutions. The right structure needed for such results is that of *regular variation*. We state two known results to be used in the proof of Theorem 1.2.10 as lemmas.

Lemma 1.2.1 *(Landau's Theorem) Let* U *be a nonnegative absolutely continuous function with derivative* u. *Suppose* U *is of regular variation of exponent* $\rho \neq 0$ *at* ∞, *and that* u *is ultimately monotone and has a finite number of sign-changes. Then* u *is of regular variation of exponent* $\rho - 1$ *at* ∞.

Lemma 1.2.2 (Berman (1992)) *Suppose* $p(z)$ *is a probability density function on the real line, and* $q(z)$ *is ultimately nonnegative, and that* $w(z) = -\frac{d}{dz}\log q(z)$, $v(z) = -\frac{d}{dz}\log p(z)$ *exist and are functions of regular oscillation, i.e., if* $z, z' \to \infty$, $\frac{z}{z'} \to 1$, *then* $\frac{f(z)}{f(z')} \to 1$ *if* $f = w$ *or* v. *If, moreover,* $\liminf_{z\to\infty}\frac{d}{dz}\log q(z) > \liminf_{z\to\infty}\frac{d}{dz}\log p(z)$, *then,* $\int_{-\infty}^{\infty} q(z)p(\gamma - z)dz = q(\gamma)\int_{-\infty}^{\infty} e^{-zw(\gamma)}p(z)dz\,(1 + o(1))$, *as* $\gamma \to \infty$.

We now present the following general result.

Theorem 1.2.10 *Suppose assumptions (1)-(7) hold true and if* $-\log q(z)$ *is a function of regular variation of some exponent* $\rho \neq 0$ *at* $z = \infty$. *Then,*

$$\triangle_n = \frac{2q(\gamma_n)e^{\frac{1}{2}\left[w(\gamma_n)\right]^2}}{n}(1 + o(1)), \quad \text{as } n \to \infty. \tag{4.64}$$

Proof By assumption (6), $w(z)$ is ultimately monotone, and by assumption (5), $w(z)$ is ultimately positive. By hypothesis, $-\log q(z)$ is a function of regular variation. Therefore, all the conditions of *Landau's theorem* (Lemma 1.2.1) are satisfied, and hence it follows that $w(z)$ is also a function of regular variation at ∞. This will imply, by well known local uniformity of convergence for functions of regular variation that if $z, z' \to \infty$, and $\frac{z}{z'} \to 1$, then $\frac{w(z)}{w(z')} \to 1$. By assumption (7), we have

$\limsup_{z\to\infty} w(z) < \infty = \limsup_{z\to\infty} \frac{d}{dz} - \log\phi(z)$. Hence, we can now appeal to Lemma 1.2.2 to coclude that

$$\int_{-\infty}^{\infty} q(z)\phi(\gamma_n - z)dz = q(\gamma_n)\int_{-\infty}^{\infty} e^{-zw(\gamma_n)}\phi(z)dz\,(1+o(1))$$

$$= q(\gamma_n)e^{\frac{1}{2}\left[w(\gamma_n)\right]^2}(1+o(1))$$

(by completing the squares), and hence, by (4.48),

$$\triangle_n = \frac{2}{n}\int_{-\infty}^{\infty} q(z)\Big[\phi(\gamma_n - z) - \phi(\gamma_n + z)\Big]dz$$

$$= \frac{2}{n}\int_{-\infty}^{\infty} q(z)\phi(\gamma_n - z)dz\,(1+o(1))$$

$$= \frac{2q(\gamma_n)e^{\frac{1}{2}\left[w(\gamma_n)\right]^2}}{n}(1+o(1)), \tag{4.65}$$

as claimed.

Acknowledgements It is a pleasure to thank Jyotishka Datta for his very gracious help with formatting the paper.

References

Banerjee M (2006) Superefficiency, contiguity, LAN University Michigan Lecture Notes
Bickel PJ (1983) Minimax estimation of the mean of a normal distribution subject to doing well at a point. In: Recent Advances in Statistics, A Festschrift for H. Chernoff, M. Rizvi, J. Rustagi and D. Siegmund eds., 511-528, Academic Press, NY
DasGupta A (2008) Asymptotic theory of statistics and probability. Springer, New York
DasGupta A (2011) Probability for statistics and machine learning: fundamentals and advanced topics. Springer, New York
Donoho D, Johnstone IM (1994) Ideal spatial adaptation by wavelet shrinkage. Biom 81: 425-455
Hájek J (1970) A characterization of limiting distributions of regular estimates, Z. Wahr verw Geb 14: 323-330
Jeganathan P (1983) Some asymptotic properties of risk functions when the limit of the experiment is mixed normal. Sankhyā Ser A 45:66-87
Johnstone IM (2012) Function estimation and Gaussian sequence models, Cambridge University Press, Forthcomiing
Le Cam L (1953) On some asymptotic properties of maximum likelihood estimates and related Bayes estimates. University California Publications 1: 277-330
Le Cam L (1973) Sur les contraintes imposees par les passages a la limite usuels en statistique, Proc. 39th session International Statistical Institute, XLV, 169-177
Lehmann EL, Romano J (2005) Testing statistical hypotheses, 3rd edn. Springer, New York
van der Vaart A (1997) Superefficiency, in Festschrift for Lucien Le Cam, D. Pollard, E. Torgersen, G. Yang, Eds. Springer, New York
van der Vaart A (1998) Asymptotic statistics. Cambridge University Press, Cambridge
Wasserman L (2005) All of Nonparametric Statistics. Springer, New York
Wellner J (2012) University Washington Lecture Notes

Chapter 5
A Note on Nonparametric Estimation of a Bivariate Survival Function Under Right Censoring

Haitao Zheng, Guiping Yang and Somnath Datta

5.1 Introduction

Bivariate survival or event time data are frequently encountered in biomedical research. Examples include data obtained from twin studies, data collected on eyes, ears, legs, breasts or kidneys from the same person, event times of two related diseases happening in one patient, etc. Like the univariate survival data, the bivaraite survival times are not always observed due to right censoring. Generally, each of the two component survival times T_j is subject to right censoring by a corresponding censoring time C_j, $j = 1, 2$. In some applications, it may be assumed that $C_1 = C_2$ and this type of censoring is sometimes referred to as univariate censoring. However, in general, the two censoring times are distinct and the data consists of independent and identically distributed (i.i.d.) copies of following four tuples: $D = (X_1, X_2, \delta_1, \delta_2)$, where $X_j = T_j \wedge C_j$, and $\delta_j = I(T_j \leq C_j)$, $j = 1, 2$, and the problem at hand is that of estimation of the bivariate survival function of $\boldsymbol{T} = (T_1, T_2)$, $S(t_1, t_2) \doteq Pr\{T_1 > t_1, T_2 > t_2\}$ on the basis of $D_1, \cdots, D_n$.

This problem has received considerable attention over the years. However, unlike the case of univariate survival data where the celebrated Kaplan–Meier estimator is "the" solution, this problem has led to many solutions each with its pros and cons. In the remainder of this section, we briefly review a number of these estimators and discuss the main issues associated with them. In the next section, we introduce a class of novel bivariate function estimators. Chapter 3 reports the results of a simulation study which shows superior performance of our estimators over existing estimators. In Sect. 4, we apply our estimators to a real life bivariate data set for illustration where

S. Datta (✉)
Department of Biostatistics and Bioinformatics,
University of Louisville, Louisville, Kentucky
e-mail: somnath.datta@louisville.edu

H. Zheng
Department of Statistics, South-west Jiaotong University, Chengdu, China

G. Yang
Teva Pharmaceuticals, Malvern, USA

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_5,

we also discuss an appropriate resampling scheme for construction of a confidence interval. The paper ends with a discussion section (Sect. 5).

As mentioned before, unlike the univariate case, there are several nonparametric estimators of the bivariate survival function that were proposed over the years. Hanley and Parnes (1983) and van der Laan (1996, 1997) studied the nonparametric maximum likelihood estimation (NPMLE) of bivariate survival function. A maximum likelihood approach with imputed observations was undertaken by Pruitt (1991). Lin and Ying (1993), Wang and Wells (1997) and Tsai and Crowley (1998) proposed some methods for estimation problem under some special censoring mechanisms, for instance, univariate censoring. Dabrowska (1988) proposed a bivariate product limit estimator using product integration. Prentice and Cai (1992) used marginal survival functions and their covariance function to estimate a bivariate survival function. Akritas and Van Keilegom (2003) utilized a marginal distribution function estimation combined with a conditional distribution function to estimate a bivariate survival function. Dabrowska (1988), Pruitt (1991), van der Laan (1996) and Akritas and Van Keilegom (2003) seem to have received the most attention in the literature.

As pointed out by Akritas and Van Keilegom (2003), many of these estimators are not proper probability distributions (even after normalization) or have non-explicit formulae, and some do not behave well in practice or depend heavily on the choice of smoothing parameters. Among other things, many such estimators (e.g., Dabrowska, 1988) may assign negative masses to certain rectangles. Recently, Shen (2010) developed three new nonparametric estimators and studied the performance of the proposed methods using simulation. Dai and Fu (2012) proposed a novel estimator based on a polar coordinate transformation.

Extensive comparative studies for some of these estimators have been performed by various researchers. For instance, van der Laan (1997) compared the Dabrowska estimator, the Prentice-Cai estimator and the NPMLE of van der Laan (1996). Akritas and Keilegom (2003) compared their estimators with those of Pruitt (1991) and van der Laan (1996). They demonstrated through simulation studies that their estimator is more efficient and easy to calculate. However, their estimator does not reduce to the empirical survival function when the data has no right censored observations. Very recently, Wang and Zafra (2009) computed a Volterra estimator with dynamic programming and compared it with Dabrowska estimator. They have shown that the new method improved the computational efficiency and produced an estimator with reasonable performance.

5.2 The Estimators

We propose a class of nonparametric estimators of a bivariate survival function under full bivariate censoring starting with a basic estimator constructed using the principle of inverse probability of censoring weighting (IPCW). This technique has its root in sample survey (Horvitz and Thompson 1952). In the survival analysis context, this was first used by Koul et al. (1981) and later on popularized by Robins and his

co-authors (Rotnitzky and Robins 2005; Satten et al. 2001; Satten and Datta 2001). Although it is not immediately obvious, it turns out that this basic rewighted estimator is equivalent to the estimator proposed by Akritas and van Keilegom (2003).

5.2.1 A Basic IPCW Estimator

First we consider estimating $g_1(t_1;s) \doteq P(T_1 > t_1|T_2 = s)$ for $s, t_1 \geq 0$. Let $0 < h < 1$ be a bandwidth tending to zero with the sample size. Note that

$$\begin{aligned} g_1(t_1;s) &\approx \prod_{t\leq t_1}\left(1 - \frac{P\{T_1 \in [t, t+dt), T_2 \in [s-h, s+h]\}}{P\{T_1 \geq t, T_2 \in [s-h, s+h]\}}\right) \\ &\approx \prod_{t\leq t_1}\left(1 - \frac{P\{T_1 \in [t, t+dt), \delta_1 = 1, T_2 \in [s-h, s+h], \delta_2 = 1\}}{P\{T_1 \geq t, C_1 \geq t, T_2 \in [s-h, s+h], \delta_2 = 1\}}\right). \end{aligned}$$

Therefore an estimator of $P(T_1 > t_1|T_2 = s)$ is given by the product limit

$$\widehat{g}_1(t_1;s) \doteq \widehat{P}(T_1 > t_1|T_2 = s) = \prod_{t\leq t_1}\left(1 - \frac{dN(t;s,h)}{Y(t;s,h)}\right), \tag{5.1}$$

where

$$N(t;s,h) = \sum_{i=1}^{n} I(T_{1i} \leq t, \delta_{1i} = 1, T_{2i} \in [s-h, s+h], \delta_{2i} = 1)$$

and

$$Y(t;s,h) = \sum_{i=1}^{n} I(T_{1i} \wedge C_{1i} \geq t, T_{2i} \in [s-h, s+h], \delta_{2i} = 1).$$

Basically, the above estimator is a conditional Kaplan estimator with the uniform kernel; it is a slight generalization of the Beran's estimator (Beran, 1981) since the conditioning variable T_2 is also subject to censoring. It is possible to use general kernel based weights (Meira-Machado et al. 2013) in defining the above conditional counting and number at risk processes.

Next, note that the bivariate survival function can be expressed as a mean $S(t_1,t_2) = E\{g_1(t_1;T_2)\, I(T_2 > t_2)\}$. Therefore, using the IPCW principles to estimate means (Datta 2005), we can estimate the joint survival function estimator by

$$\widehat{S}_1(t_1,t_2) \doteq \frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{g}_1(t_1;T_{2i})\,\delta_{2i}}{\widehat{K}_2(T_{2i}-)} I(T_{2i} > t_2) = \frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{g}_1(t_1;X_{2i})\,\delta_{2i}}{\widehat{K}_2(X_{2i}-)} I(X_{2i} > t_2),$$

where $\widehat{K}_2$ is a Kaplan–Meier estimator of the survival function of C_2 (right censored by T_2); note that K_2 can be computed by the standard Kaplan-Meier product limit formula with the data $\{X_{2i}, 1-\delta_{2i};\ 1 \le i \le n\}$. We can exchange the roles of T_1 and T_2 in the above estimator to obtain yet another bivariate survival function estimator

$$\widehat{S}_2(t_1,t_2) \leftarrows \frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{g}_2(t_2;T_{1i})\,\delta_{1i}}{\widehat{K}_1(T_{1i}-)} I(T_{1i} > t_1) = \frac{1}{n}\sum_{i=1}^{n} \frac{\widehat{g}_2(t_2;X_{1i})\,\delta_{1i}}{\widehat{K}_1(X_{1i}-)} I(X_{1i} > t_1),$$

and more generally, we can take a convex combination of these two estimators to obtain a class of IPCW estimators

$$\widehat{S}_W(t_1,t_2) = \alpha(t_1,t_2)\widehat{S}_1(t_1,t_2) + \{1-\alpha(t_1,t_2)\}\widehat{S}_2(t_1,t_2), \tag{5.2}$$

where $0 \le \alpha(t_1,t_2) \le 1$ is a known (user selectable) function.

It turns out that this IPCW estimator is well connected to the Akritas and Van Keilegom (2003) (AK, hereafter) estimator. They obtained their estimator by averaging (integrating) a Beran's estimator $\widehat{g}_j$ with respect to the Kaplan–Meier estimator of $T_{j'}$, $j' = 3-j$. Since Satten and Datta (2001) showed that the Kaplan–Meier estimator has a IPCW representation, we immediately obtain the following result.

Proposition 1. *The IPCW estimator (1.2) is the same as Akritas and Van Keilegom estimator.*

This IPCW estimator was proposed by Yang (2005), where the connection with AK was not established. Note that $\widehat{S}_W(t_1,t_2)$ may not be a proper bivariate distribution function unless $\alpha(t_1,t_2)$ is free from t_1,t_2. Furthermore, it does not reduce to the empirical survival function for uncensored data.

5.2.2 A Class of Modified Estimators

We now introduce a new bivariate survival estimator that uses the above estimator in its construction and is expected to be more efficient. One nice feature of the new estimator is that it reduces to the empirical survival function in the case of complete (e.g., uncensored) data.

The idea behind the new estimator is as follows. Consider the following representation of the empirical survival function

$$\begin{aligned} S_n(t_1,t_2) &= n^{-1}\sum_{i=1}^{n} \delta_{1i}\delta_{2i} I(T_{1i} > t_1, T_{2i} > t_2) \\ &\quad + n^{-1}\sum_{i=1}^{n} \delta_{1i}\overline{\delta}_{2i} I(T_{1i} > t_1, T_{2i} > t_2) \\ &\quad + n^{-1}\sum_{i=1}^{n} \overline{\delta}_{1i}\delta_{2i} I(T_{1i} > t_1, T_{2i} > t_2) \end{aligned}$$

$$+n^{-1}\sum_{i=1}^{n}\overline{\delta}_{1i}\overline{\delta}_{2i}I\left(T_{1i}>t_1,T_{2i}>t_2\right) \tag{5.3}$$

where $\overline{\delta}_{ji}$, $j=1,2$, indicates that the jth component is censored. Note, however, that only the terms in the first sum can be evaluated in the presence of censoring. Therefore, we replace the terms of the other sums by their conditional expectations given the available data in a suitable way. These conditional expectations can be estimated using our previously defined estimators of the conditional and bivariate survival functions.

The second summand in (5.3) is replaced by

$$E\left(\delta_{1i}\overline{\delta}_{2i}I\left(T_{1i}>t_1,T_{2i}>t_2\right)\left|X_{1i},X_{2i},\delta_{1i},\delta_{2i}\right.\right)$$

$$=\delta_{1i}\overline{\delta}_{2i}I\left(X_{1i}>t_1\right)\frac{P\left(T_{2i}>t_2\vee C_{2i}|T_{1i}\right)}{P\left(T_{2i}>C_{2i}|T_{1i}\right)}.$$

We can use our earlier estimator $\widehat{g}_2$ to estimate the above conditional probabilities. The third term of (5.3) can be handled in a similar way. The summand corresponding to the last term of (5.3) is replaced by

$$E\left(\overline{\delta}_{1i}\overline{\delta}_{2i}I\left(T_{1i}>t_1,T_{2i}>t_2\right)\left|X_{1i},X_{2i},\delta_{1i},\delta_{2i}\right.\right)=\overline{\delta}_{1i}\overline{\delta}_{2i}\frac{S\left(t_1\vee C_{1i},t_2\vee C_{2i}\right)}{S\left(C_{1i},C_{2i}\right)}.$$

Note that we could use our preliminary IPCW estimator $\widehat{S}_W$ in estimating this term. Finally combining the terms together we get second estimator of the bivariate survival function that appears to be novel in the literature of bivariate survival estimator. A penultimate form of this work appears in an unpublished thesis by Yang (2005); we call this estimator described below a 1-step modified estimator

$$\begin{aligned}\widehat{S}_{1,M}\left(t_1,t_2\right) &= n^{-1}\sum_{i=1}^{n}\delta_{1i}\delta_{2i}I\left(X_{1i}>t_1,X_{2i}>t_2\right)\\ &\quad+n^{-1}\sum_{i=1}^{n}\delta_{1i}\overline{\delta}_{2i}I\left(X_{1i}>t_1\right)\frac{\widehat{g}_2\left(t_2\vee X_{2i};X_{1i}\right)}{\widehat{g}_2\left(X_{2i};X_{1i}\right)}\\ &\quad+n^{-1}\sum_{i=1}^{n}\overline{\delta}_{1i}\delta_{2i}I\left(X_{2i}>t_2\right)\frac{\widehat{g}_1\left(t_1\vee X_{1i};X_{2i}\right)}{\widehat{g}_1\left(X_{1i};X_{2i}\right)}\\ &\quad+\overline{\delta}_{1i}\overline{\delta}_{2i}\frac{\widehat{S}_W\left(t_1\vee X_{1i},t_2\vee X_{2i}\right)}{\widehat{S}_W\left(X_{1i},X_{2i}\right)}.\end{aligned} \tag{5.4}$$

Note that we can use (5.4) iteratively to obtain a sequence of bivariate survival function estimators

$$\begin{aligned}\widehat{S}_{k,M}(t_1,t_2) &= n^{-1}\sum_{i=1}^{n}\delta_{1i}\delta_{2i}I(X_{1i}>t_1, X_{2i}>t_2)\\ &+n^{-1}\sum_{i=1}^{n}\delta_{1i}\overline{\delta}_{2i}I(X_{1i}>t_1)\frac{\widehat{g}_2(t_2\vee X_{2i};X_{1i})}{\widehat{g}_2(X_{2i};X_{1i})}\\ &+n^{-1}\sum_{i=1}^{n}\overline{\delta}_{1i}\delta_{2i}I(X_{2i}>t_2)\frac{\widehat{g}_1(t_1\vee X_{1i};X_{2i})}{\widehat{g}_1(X_{1i};X_{2i})}\\ &+\overline{\delta}_{1i}\overline{\delta}_{2i}\frac{\widehat{S}_{k-1,M}(t_1\vee X_{1i},t_2\vee X_{2i})}{\widehat{S}_{k-1,M}(X_{1i},X_{2i})},\end{aligned}$$

for $k \geq 1$, where $\widehat{S}_{1,M} = \widehat{S}_W$. We call this a k-step modified estimator (MB-k).

5.2.3 Bandwidth Selection

Akritas and Van Keilegom (2003) suggested a resampling based bandwidth selector. Here we propose a cross-validation based bandwidth selector which could be computationally less intensive. Let $\widehat{S}(t_1,t_2) = \widehat{S}(t_1,t_2;h)$ be a bivariate survival function estimate where h is a smoothing parameter. Let us attempt to minimize the following (weighted) integrated mean squared error:

$$\text{IMSE} = E\int \left|\widehat{S}(t_1,t_2) - S(t_1,t_2)\right|^2 \{H(dt_1,dt_2)\},$$

where H is the bivariate cumulative hazard function,

$$= E\int \frac{\widehat{S}^2(t_1,t_2)}{S(t_1,t_2)}\{-S(dt_1,dt_2)\} - 2E\widehat{S}(T_1^*,T_2^*) + \int S(t_1,t_2)\{-S(dt_1,dt_2)\}$$

where (T_1^*,T_2^*) is an independent (of the original sample) realization of the true bivariate failure time.

Since, the third term is a constant it can be dropped from the minimization process. Furthermore, replacing the first two terms by their estimates we get the following CV criterion function to be minimized:

$$CV(h) = \int \widehat{S}^2(t_1,t_2)\{\widehat{H}(dt_1,dt_2)\} - \frac{2}{n}\sum_{i=1}^{n}\frac{\widehat{S}_{-i}(X_{1i},X_{2i})\delta_{1i}\delta_{2i}}{\widehat{S}_n^C(X_{1i}-,X_{2i}-)};$$

here $\widehat{S}_n^C$ is a bivariate survival function estimator of the censoring times (that may be based on an auxiliary bandwidth that is expected to have little effect on the entire process) and $\widehat{S}_{-i}$ is the bivariate survival function estimator based on the sample with the ith pair deleted. Finally, h can be selected to minimize CV over a grid of values.

We have not studied the performance of this bandwidth selector in this paper. It may be pursued elsewhere.

Table 5.1 Design choices for the simulation

Parameters	Values
Correlation Coefficients ρ	0, 0.6
Failure Time Quartiles t_j	0.2, 0.5, 0.8
Sample Size n (bandwidth $h(n)$)	50 (0.5), 100 (0.3), 200 (0.2)
Censoring proportion	0.2, 0.5

5.3 Simulations

We conducted a simulation study to compare the performances of the bivariate survival estimators described in the previous section. In particular, since the IPCW estimator is the same as the one proposed by Akritas and Van Keilegom (2003), this provides a comparison between that and the novel modified estimators. It is perhaps worth noting that in their paper, Akritas and Van Keilegom already established superiority of their estimator over earlier estimators developed by Dabrowska (1988), Pruitt (1991), Prentice and Cai (1992), Van der Laan (1996), and Wang and Wells (1997). Therefore, this also provides a basis for an indirect comparison with those estimators.

For the IPCW or AK estimator, we considered both uniform and normal kernels. However, in order to minimize the computational burden, we only use the IPCW estimator with a uniform kernel to compute our modified estimators. Also, throughout the weight function α was taken to be 0.5.

The true survival pairs were generated from a bivariate log-normal distribution; i.e., $\log(T_1, T_2) \sim N_2(0, 0, 1, 1, \rho)$. The censoring distribution function is bivariate Gamma with independent components; that is $C_1, C_2 \overset{iid}{\sim} G(\alpha, 1)$. We adjust the values of α to control the censoring rate.

The scope of this simulation study was fairly extensive. This included comparing performances under two choices of the correlation between the log-survival times: zero and moderate. Three sample sizes 50, 100, and 200 were considered in this study. The estimators were computed over a grid of quantile pairs. We also considered different censoring proportions (rates). For computational ease, a non-random bandwidth sequence decreasing with the sample size was used. These are listed in Table 5.1 below.

For each simulation setting, we compute the bias and mean squared error of the estimators based on the Monte Carlo technique with 500 trials each. These values are reported within parenthesis in Tables 5.2–5.5. In these tables, the format ($10^3 \times$BIAS, $10^3 \times$MSE) is used; for example, $(-3.56, 4.48)$ means that bias is -3.45×10^{-3} and MSE is 4.48×10^{-3}. We have considered the following five estimators: IPCW/AK with uniform kernel, IPCW/AK with normal kernel, MB-1, MB-2 and MB-3.

In Tables 5.3 and 5.5, we report the simulation results for $\rho = 0.6$. We compare the results under different censoring proportion. With censoring proportion 0.5, we find that the modified estimators give much better result than the IPCW/AK estimators in terms of smaller bias and smaller MSE for all sample sizes under consideration.

Table 5.2 Values of Bias and MSE of five bivariate survival function estimators when $\rho = 0$ and censoring rate is 20 %

Sample size	Times (t_1, t_2)	IPCW/AK (uniform kernel)	IPCW/AK (normal kernel)	MB-1	MB-2	MB-3
	(3.8, 3.8)	(−2.98 4.64)	(−23.00 4.43)	(−3.56 4.48)	(−3.57 4.49)	(−3.57 4.49)
	(3.8, 4.5)	(−1.73 4.63)	(−17.29 4.20)	(−1.60 4.67)	(−1.61 4.68)	(−1.61 4.68)
	(3.8, 5.3)	(0.24 2.77)	(−11.29 2.38)	(1.71 2.66)	(1.76 2.66)	(1.77 2.66)
	(4.5, 3.8)	(−0.52 4.71)	(−16.80 4.15)	(−1.81 4.71)	(−1.75 4.71)	(−1.74 4.71)
$n = 50$	(4.5, 4.5)	(−0.93 3.38)	(−12.88 2.71)	(−1.87 3.69)	(−1.82 3.69)	(−1.80 3.70)
	(4.5, 5.3)	(0.76 1.61)	(−8.03 1.11)	(1.02 1.72)	(1.10 1.73)	(1.12 1.73)
	(5.3, 3.8)	(1.97 2.90)	(−11.25 2.51)	(1.01 2.87)	(0.95 2.88)	(0.93 2.88)
	(5.3, 4.5)	(1.60 1.54)	(−8.01 1.11)	(0.69 1.75)	(0.63 1.76)	(0.62 1.76)
	(5.3, 5.3)	(2.55 0.62)	(−3.88 0.32)	(3.02 0.78)	(3.02 0.79)	(3.02 0.79)
	(3.78 3.79)	(1.84 2.42)	(−10.60 2.23)	(0.49 2.26)	(0.48 2.26)	(0.48 2.26)
	(3.78 4.48)	(0.39 2.33)	(−9.63 2.13)	(−1.26 2.34)	(−1.26 2.34)	(−1.25 2.34)
	(3.78 5.27)	(−1.69 1.33)	(−9.54 1.17)	(−2.17 1.22)	(−2.16 1.23)	(−2.16 1.23)
	(4.49 3.79)	(1.12 2.46)	(−9.22 2.11)	(0.30 2.30)	(0.28 2.30)	(0.28 2.30)
$n = 100$	(4.49 4.48)	(−0.21 1.72)	(−8.11 1.46)	(−1.84 1.86)	(−1.84 1.86)	(−1.83 1.86)
	(4.49 5.27)	(−1.15 0.83)	(−6.81 0.64)	(−1.54 0.97)	(−1.53 0.98)	(−1.53 0.98)
	(5.30 3.79)	(−2.73 1.56)	(−9.97 1.31)	(−2.64 1.36)	(−2.66 1.36)	(−2.66 1.36)
	(5.30 4.48)	(−2.42 0.86)	(−7.51 0.66)	(−2.93 0.94)	(−2.95 0.94)	(−2.94 0.94)
	(5.30 5.27)	(−0.82 0.33)	(−4.34 0.20)	(−0.76 0.40)	(−0.78 0.40)	(−0.78 0.40)
	(3.78 3.79)	(1.10 1.22)	(−5.19 1.18)	(0.68 1.23)	(0.69 1.23)	(0.69 1.23)
	(3.78 4.48)	(1.57 1.40)	(−3.69 1.26)	(1.10 1.38)	(1.11 1.38)	(1.11 1.38)
	(3.78 5.27)	(0.17 0.83)	(−3.73 0.73)	(−0.06 0.81)	(−0.06 0.81)	(−0.06 0.81)
	(4.49 3.79)	(0.12 1.27)	(−4.99 1.21)	(−0.20 1.28)	(−0.19 1.29)	(−0.19 1.29)
$n = 200$	(4.49 4.48)	(1.06 0.92)	(−3.27 0.79)	(1.18 0.99)	(1.19 0.99)	(1.19 0.99)
	(4.49 5.27)	(0.12 0.47)	(−2.76 0.37)	(0.04 0.50)	(0.05 0.50)	(0.04 0.50)
	(5.30 3.79)	(1.42 0.80)	(−2.61 0.70)	(1.27 0.77)	(1.28 0.77)	(1.28 0.77)
	(5.30 4.48)	(1.38 0.45)	(−1.75 0.35)	(1.14 0.51)	(1.15 0.51)	(1.15 0.51)
	(5.30 5.27)	(0.78 0.17)	(−1.19 0.12)	(0.64 0.20)	(0.64 0.20)	(0.64 0.20)

In each cell, the numbers denote $10^3 \times$bias and $10^3 \times$MSE; for example, (-3.56, 4.48) means that bias is -3.56×10^{-3} and MSE is 4.48×10^{-3}

Table 5.3 Values of Bias and MSE of five bivariate survival function estimators when $\rho = 0.6$ and censoring rate is 20 %.

Sample size	Times (t_1, t_2)	IPCW/AK (uniform kernel)	IPCW/AK (normal kernel)	MB-1	MB-2	MB-3
	(3.78 3.79)	(−15.85 5.43)	(−49.46 7.30)	(−3.58 4.43)	(−3.44 4.42)	(−3.43 4.42)
	(3.78 4.48)	(−15.49 5.88)	(−49.05 7.49)	(−0.65 5.04)	(−0.39 5.03)	(−0.37 5.03)
	(3.78 5.27)	(−5.27 3.68)	(−27.48 3.65)	(3.47 3.18)	(3.75 3.18)	(3.79 3.18)
	(4.49 3.79)	(−16.28 5.93)	(−50.30 7.67)	(−3.02 5.05)	(−2.82 5.05)	(−2.81 5.05)
$n = 50$	(4.49 4.48)	(−23.55 5.23)	(−65.02 8.23)	(−2.76 4.43)	(−2.44 4.43)	(−2.41 4.43)
	(4.49 5.27)	(−12.39 3.12)	(−42.36 3.86)	(1.10 2.92)	(1.41 2.93)	(1.46 2.93)
	(5.30 3.79)	(−5.47 4.15)	(−28.97 4.27)	(1.04 3.85)	(1.17 3.87)	(1.17 3.87)
	(5.30 4.48)	(−12.05 3.32)	(−43.14 4.20)	(0.41 3.31)	(0.63 3.33)	(0.65 3.34)
	(5.30 5.27)	(−10.52 1.94)	(−36.87 2.40)	(0.44 2.04)	(0.76 2.06)	(0.80 2.06)
	(3.78 3.79)	(−8.92 2.85)	(−29.16 3.59)	(−0.47 2.59)	(−0.41 2.59)	(−0.41 2.59)
	(3.78 4.48)	(−15.55 3.07)	(−36.25 4.00)	(−5.92 2.67)	(−5.80 2.67)	(−5.79 2.67)
	(3.78 5.27)	(−7.00 2.07)	(−20.32 2.20)	(−1.26 1.90)	(−1.11 1.91)	(−1.10 1.91)
	(4.49 3.79)	(−14.12 2.97)	(−35.08 3.96)	(−4.18 2.73)	(−4.06 2.73)	(−4.06 2.73)
$n = 100$	(4.49 4.48)	(−23.01 2.77)	(−52.06 4.82)	(−7.97 2.38)	(−7.77 2.38)	(−7.76 2.38)
	(4.49 5.27)	(−12.49 1.65)	(−33.92 2.37)	(−2.30 1.50)	(−2.10 1.51)	(−2.08 1.51)
	(5.30 3.79)	(−8.34 2.17)	(−22.31 2.32)	(−3.36 1.95)	(−3.24 1.95)	(−3.23 1.95)
	(5.30 4.48)	(−13.71 1.81)	(−35.56 2.53)	(−4.39 1.67)	(−4.21 1.67)	(−4.19 1.67)
	(5.30 5.27)	(−10.79 1.01)	(−30.85 1.55)	(−2.62 1.01)	(−2.40 1.01)	(−2.38 1.02)
	(3.78 3.79)	(−6.22 1.24)	(−18.57 1.54)	(0.10 1.10)	(0.15 1.10)	(0.15 1.10)
	(3.78 4.48)	(−8.29 1.32)	(−20.98 1.68)	(−0.93 1.23)	(−0.84 1.23)	(−0.84 1.23)
	(3.78 5.27)	(−3.68 0.99)	(−11.14 1.07)	(−0.69 0.97)	(−0.59 0.97)	(−0.59 0.97)
	(4.49 3.79)	(−7.23 1.45)	(−19.93 1.79)	(−0.51 1.36)	(−0.42 1.36)	(−0.42 1.36)
$n = 200$	(4.49 4.48)	(−12.83 1.28)	(−33.41 2.20)	(−1.35 1.13)	(−1.21 1.13)	(−1.21 1.13)
	(4.49 5.27)	(−6.96 0.82)	(−21.57 1.17)	(−0.84 0.82)	(−0.70 0.82)	(−0.70 0.82)
	(5.30 3.79)	(−6.92 0.98)	(−13.55 1.05)	(−3.31 0.91)	(−3.22 0.91)	(−3.22 0.91)
	(5.30 4.48)	(−10.40 0.84)	(−24.12 1.21)	(−3.96 0.76)	(−3.83 0.76)	(−3.82 0.76)
	(5.30 5.27)	(−8.41 0.48)	(−21.98 0.80)	(−3.31 0.46)	(−3.16 0.47)	(−3.15 0.47)

Table 5.4 Values of Bias and MSE of five bivariate survival function estimators when $\rho = 0$ and censoring rate is 50 %.

Sample size	Times (t_1, t_2)	IPCW/AK (uniform kernel)	IPCW/AK (normal kernel)	MB-1	MB-2	MB-3
	(3.78 3.79)	(0.88 6.61)	(−27.52 5.33)	(0.01 4.96)	(0.18 4.97)	(0.19 4.97)
	(3.78 4.48)	(−2.23 7.58)	(−23.34 5.87)	(0.36 6.18)	(0.91 6.18)	(0.97 6.18)
	(3.78 5.27)	(−1.37 4.71)	(−16.30 3.25)	(2.66 3.92)	(3.21 3.97)	(3.29 3.98)
	(4.49 3.79)	(3.81 7.31)	(−20.49 5.43)	(2.74 5.89)	(2.88 5.88)	(2.90 5.88)
$n = 50$	(4.49 4.48)	(2.06 5.38)	(−15.78 3.72)	(4.85 5.10)	(5.37 5.18)	(5.43 5.19)
	(4.49 5.27)	(1.88 2.67)	(−10.29 1.54)	(4.81 2.67)	(5.24 2.76)	(5.30 2.79)
	(5.30 3.79)	(1.80 4.48)	(−17.65 3.15)	(2.64 3.76)	(2.99 3.78)	(3.06 3.79)
	(5.30 4.48)	(2.81 2.69)	(−12.11 1.50)	(5.18 2.55)	(5.74 2.63)	(5.84 2.65)
	(5.30 5.27)	(3.84 1.13)	(−6.31 0.43)	(6.07 1.24)	(6.56 1.33)	(6.66 1.36)
	(3.78 3.79)	(3.70 3.67)	(−16.40 3.04)	(−1.52 2.79)	(−1.57 2.79)	(−1.56 2.79)
	(3.78 4.48)	(1.45 4.06)	(−13.64 3.13)	(−1.80 3.27)	(−1.79 3.27)	(−1.77 3.28)
	(3.78 5.27)	(0.17 2.59)	(−10.65 1.96)	(−0.55 2.10)	(−0.36 2.09)	(−0.31 2.09)
	(4.49 3.79)	(5.95 4.00)	(−10.96 2.88)	(2.50 3.00)	(2.43 2.99)	(2.43 2.99)
$n = 100$	(4.49 4.48)	(1.93 2.87)	(−9.84 2.00)	(−0.83 2.41)	(−0.92 2.42)	(−0.92 2.42)
	(4.49 5.27)	(0.38 1.49)	(−7.51 1.03)	(−0.36 1.38)	(−0.27 1.41)	(−0.22 1.41)
	(5.30 3.79)	(1.44 2.76)	(−10.92 1.89)	(−0.09 2.06)	(−0.04 2.06)	(0.00 2.07)
	(5.30 4.48)	(0.00 1.56)	(−8.36 0.98)	(−0.90 1.33)	(−0.84 1.35)	(−0.81 1.36)
	(5.30 5.27)	(1.28 0.66)	(−4.75 0.34)	(1.55 0.64)	(1.73 0.67)	(1.81 0.67)
	(3.78 3.79)	(1.09 1.58)	(−11.65 1.81)	(−0.66 1.22)	(−0.65 1.22)	(−0.65 1.22)
	(3.78 4.48)	(0.62 1.79)	(−10.48 1.62)	(−1.41 1.47)	(−1.40 1.47)	(−1.40 1.47)
	(3.78 5.27)	(2.48 1.22)	(−7.74 1.52)	(1.21 0.88)	(1.11 0.87)	(1.11 0.87)
	(4.49 3.79)	(0.61 1.72)	(−7.32 2.39)	(−0.24 1.36)	(−0.24 1.35)	(−0.24 1.35)
$n = 200$	(4.49 4.48)	(0.40 1.31)	(−6.50 1.61)	(−0.82 1.15)	(−0.82 1.15)	(−0.82 1.15)
	(4.49 5.27)	(3.80 0.88)	(−4.55 0.87)	(2.93 0.69)	(2.93 0.69)	(2.87 0.69)
	(5.30 3.79)	(1.57 1.26)	(−5.60 0.60)	(2.13 1.07)	(2.10 1.07)	(1.98 1.07)
	(5.30 4.48)	(2.59 0.88)	(−4.84 0.66)	(2.75 0.79)	(2.77 0.80)	(2.78 0.80)
	(5.30 5.27)	(4.59 0.44)	(−2.72 0.30)	(4.58 0.39)	(4.68 0.40)	(4.67 0.40)

Table 5.5 Values of Bias and MSE of five bivariate survival function estimators when $\rho = 0.6$ and censoring rate is 50 %

Sample size	Times (t_1, t_2)	IPCW/AK (uniform kernel)	IPCW/AK (normal kernel)	MB-1	MB-2	MB-3
	(3.78 3.79)	(−4.93 9.08)	(−46.08 7.41)	(−1.22 4.73)	(0.10 4.71)	(0.20 4.71)
	(3.78 4.48)	(−16.13 10.02)	(−51.88 8.41)	(−9.04 5.91)	(−7.17 5.88)	(−6.94 5.87)
	(3.78 5.27)	(−10.40 7.74)	(−33.32 5.13)	(−5.19 4.81)	(−3.63 4.75)	(−3.37 4.76)
	(4.49 3.79)	(−15.39 10.81)	(−53.57 9.47)	(−8.71 6.55)	(−6.98 6.52)	(−6.78 6.52)
$n = 50$	(4.49 4.48)	(−28.11 9.53)	(−73.50 9.34)	(−15.43 5.91)	(−13.07 5.90)	(−12.74 5.90)
	(4.49 5.27)	(−19.09 6.35)	(−51.91 5.53)	(−11.37 4.23)	(−9.57 4.22)	(−9.26 4.23)
	(5.30 3.79)	(−6.84 8.60)	(−32.75 5.88)	(−3.78 5.31)	(−2.33 5.27)	(−2.11 5.26)
	(5.30 4.48)	(−14.71 6.58)	(−49.89 5.49)	(−7.44 4.25)	(−5.54 4.24)	(−5.23 4.24)
	(5.30 5.27)	(−14.32 4.04)	(−44.45 3.33)	(−7.64 2.81)	(−5.89 2.86)	(−5.56 2.88)
	(3.78 3.79)	(−10.87 3.10)	(−34.50 3.76)	(−5.21 2.24)	(−4.41 2.21)	(−4.34 2.21)
	(3.78 4.48)	(−7.65 3.83)	(−30.91 4.14)	(−2.43 3.13)	(−1.40 3.12)	(−1.26 3.11)
	(3.78 5.27)	(−3.82 2.52)	(−17.26 2.34)	(−0.79 2.27)	(0.25 2.28)	(0.47 2.28)
	(4.49 3.79)	(−10.44 3.91)	(−31.46 4.20)	(−3.11 2.99)	(−1.96 2.98)	(−1.82 2.98)
$n = 100$	(4.49 4.48)	(−15.22 3.49)	(−47.55 3.88)	(−5.55 2.84)	(−4.06 2.84)	(−3.84 2.84)
	(4.49 5.27)	(−9.78 2.03)	(−33.01 2.50)	(−4.03 1.88)	(−2.69 1.91)	(−2.42 1.92)
	(5.30 3.79)	(−2.99 2.72)	(−17.51 2.17)	(0.80 2.00)	(1.78 1.98)	(1.99 1.98)
	(5.30 4.48)	(−6.42 2.17)	(−32.93 2.36)	(−0.39 1.58)	(0.92 1.57)	(1.18 1.57)
	(5.30 5.27)	(−7.08 1.19)	(−31.03 1.53)	(−1.60 1.09)	(−0.23 1.13)	(0.10 1.14)
	(3.78 3.79)	(−7.50 1.35)	(−20.65 1.81)	(−3.54 1.31)	(−3.49 1.29)	(−3.46 1.29)
	(3.78 4.48)	(−8.01 1.23)	(−22.48 1.62)	(−1.05 0.95)	(−1.90 0.90)	(−1.85 0.90)
	(3.78 5.27)	(−3.52 1.69)	(−11.74 1.52)	(−1.90 1.11)	(−0.90 1.04)	(−0.79 1.04)
	(4.49 3.79)	(−8.77 3.00)	(−30.32 2.89)	(−2.41 2.27)	(−2.21 2.21)	(−2.14 2.20)
$n = 200$	(4.49 4.48)	(−8.62 2.04)	(−40.50 2.61)	(−3.52 1.23)	(−1.22 1.15)	(−1.13 1.15)
	(4.49 5.27)	(−5.70 1.67)	(−19.55 1.87)	(−3.62 1.09)	(−2.43 1.03)	(−2.30 1.02)
	(5.30 3.79)	(−6.65 1.25)	(−12.60 1.60)	(−0.44 1.06)	(−0.60 0.99)	(−0.48 0.98)
	(5.30 4.48)	(−0.76 1.28)	(−18.84 1.66)	(−0.71 1.06)	(−0.78 0.97)	(−0.65 0.97)
	(5.30 5.27)	(−3.15 1.12)	(−20.72 1.34)	(−0.85 0.91)	(−0.76 0.86)	(−0.76 0.86)

Table 5.6 Range of kidney infection times (in days).

Event times	X_1	Percent censored	X_2	Percent censored
Range	[2,562]	28.9	[5,511]	18.4

IPCW/AK with the uniform kernel has better performance than that with normal kernel. We observe the same pattern of performance amongst the five estimators at 20 % censoring as well. All four estimation methods have better performance at larger sample size and/or small censoring proportion. The modified estimators MB-2 and MB-3 have only slightly better performance than the one step modified estimator MB-1 in most cases.

When the two components of bivariate survival data are independent, the relative performances of IPCW/AK and the MB estimators are mixed (Tables 5.2 and 5.4). However, for moderate (50 %) censoring, the MB estimators have better performance than the IPCW/AK estimator in terms of MSE. If censoring proportion decreases all methods perform better in terms of MSE.

5.4 An Application to Real Data

We use our MB-1 bivariate survival estimator on a data set from McGilchrist and Aisbett (1991). This data set contains recurrence times to infection at the point of insertion of a catheter for 38 kidney patients using portable dialysis equipment. Two times to recurrence of an infection (days since catheter placement for each episode) were recorded as T_1 and T_2 for each patient; δ_1 and δ_2 were also recorded as the event (infection or censoring) indicators.

We present the range of the event times X_j and the corresponding censoring rates in Table 5.6. The overall censoring rate, where at least one of T_1 or T_2 in a pair was right censored, was 39.5 %.

We construct our estimator MB-1 on a bivariate grid of 30×30 pairs of time points that are evenly spaced between the observed marginal ranges in the data set. For reference to the IPCW/AK estimator that was constructed in Akritas and van Keilegom (2003), we choose the same bandwidth of $h = 80$. The result is displayed in Fig. 5.1.

We also report (Fig. 5.2) the corresponding marginal estimators of the survival functions of T_j, $j = 1, 2$. The corresponding pointwise confidence intervals were obtained by a smoothed bootstrap. Let $\widehat{S}$ be our MB-1 estimator of the joint survival function of $\boldsymbol{T} = (T_1, T_2)$ and by switching the roles of $\boldsymbol{T}$ and $\boldsymbol{C}$, let $\widehat{S}^C$ be the MB-1 survival function estimator of the pairs of censoring times $\boldsymbol{C} = (C_1, C_2)$. A smoothed bootstrap sample of size n, $(T_{11}^*, T_{21}^*) \cdots, (T_{1n}^*, T_{2n}^*)$, is generated from $\widehat{S}(\cdot, \cdot; \widetilde{h})$, and independently, the corresponding censoring pairs $(C_{11}^*, C_{21}^*) \cdots, (C_{1n}^*, C_{2n}^*)$ are generated from $\widehat{S}^C(\cdot, \cdot; \widetilde{h})$ to produce $X_{ji}^* = T_{ji}^* \wedge C_{ji}^*$, $\delta_{ji}^* = I(T_{ji}^* \le C_{ji}^*)$, $j = 1, 2$; $1 \le i \le n$. Note that a larger bandwidth $\widetilde{h} = h^{1.2}$ is needed to generate the

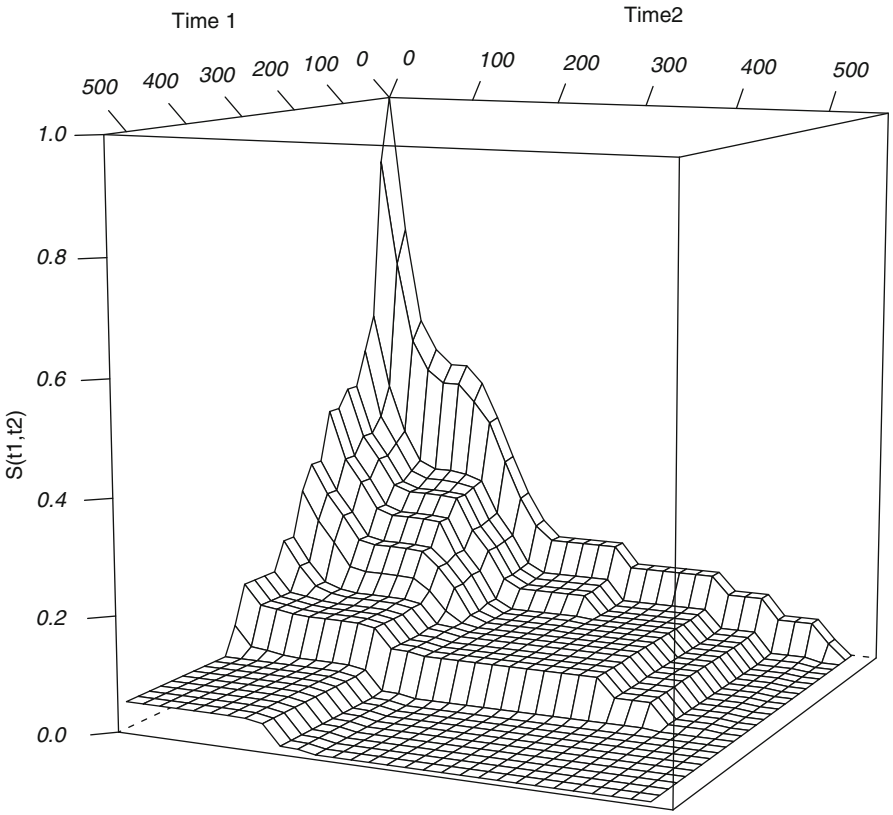


Fig. 5.1 Estimated bivariate survival function for two kidney infection times (in days)

bootstrap version of the data in order to capture the bias term in the bootstrap world (Li and Datta 2001).

Let $\widehat{S}_1(t_1;h) = \widehat{S}(t_1,0;h)$ denote the marginal survival function estimator of T_1. For $0 < \alpha < 1$, let $\widehat{\Delta}_{1-\alpha/2}(t_1)$ be the $(1-\alpha/2) \times 100^{\text{th}}$ percentile of the bootstrap distribution of

$$\Delta^* = \left| \sin^{-1}\left\{\sqrt{\widehat{S}_1^*(t_1;h)}\right\} - \sin^{-1}\left\{\sqrt{\widehat{S}_1(t_1;\widetilde{h})}\right\} \right|$$

where $\widehat{S}_1^*(t_1;h)$ uses the same bandwidth as in the original but is based on the bootstrap sample; however, $\widehat{S}_1(t_1;\widetilde{h})$ for centering is recomputed from the original sample but using the new bandwidth $\widetilde{h} = h^{1.2}$. Then the pointwise confidence interval for the marginal survival function of T_1 at a time t_1 is given by $[L,U]$, with

$$L = \sin^2\left\{\max\left(0, \sin^{-1}\left\{\sqrt{\widehat{S}_1(t_1;h)}\right\} - \widehat{\Delta}_{1-\alpha/2}\right)\right\},$$

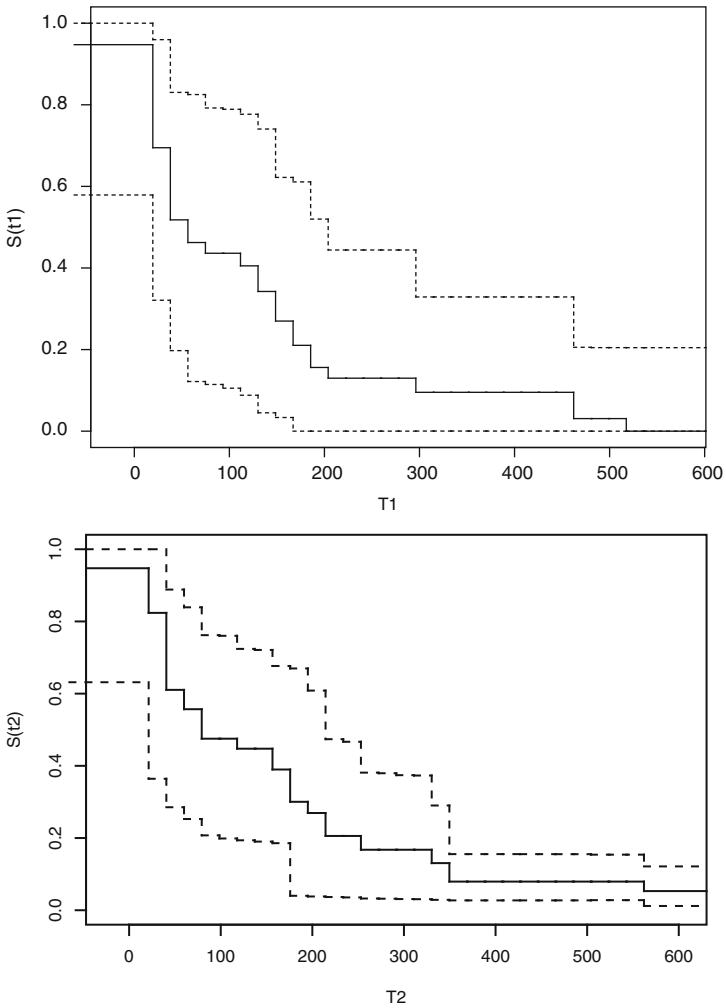


Fig. 5.2 Estimated marginal survival functions with 95% confidence intervals for kidney infection times (in days)

and

$$U = \sin^2 \left\{ \min \left(\frac{\pi}{2}, \sin^{-1} \left\{ \sqrt{\widehat{S}_1 \left(t_1; h \right)} \right\} + \widehat{\Delta}_{1-\alpha}/2 \right) \right\};$$

the confidence interval for the T_2 can be calculated in the same way.

The two marginal distributions look largely similar (Fig. 5.2); however the second infection time appears to be stochastically larger.

5.5 Discussion

In this paper, we propose a class of novel non-parametric estimators MB-k of a bivariate survival function under general independent right censoring. The proposed estimators were investigated via an extensive simulation study and compared with the IPCW estimators which were shown to be equivalent to earlier estimators proposed by Akritas and Van Keilegom (2003). From our simulation study, we find that correlation between paired failure times may play an important role on the behavior of our bivariate survival function estimators. The novel estimators may indeed perform better if the association between the paired survival times is moderate/strong.

It is fairly easy to extend these estimators to a general dependent censoring setup due to their IPCW forms. Basically, we can handle any censoring mechanism that can express the hazard of censoring C_j in terms of an observed, possibly time dependent, predictable covariate $Z_j(t_j)$. Once we fit the appropriate model we would replace the $\widehat{K}_j$ by the following formula

$$\widehat{K}_j(t_j) = exp\left\{-\int_0^{t_j} \widehat{\lambda}^{C_j}(s|Z_j(u), 0 \le u \le s)ds\right\}.$$

A flexible model suitable for this is the additive hazard model by Aalen (1989). See Satten et al. (2001) for further details of the construction of this general $\widehat{K}_j$.

Acknowledgements Zheng received research support from the Fundamental Research Funds for the Central Universities, China, SWJTU12ZT15. Datta's research was supported by grants from the United States National Science Foundation (DMS-0706965) and United States National Security Agency (H98230-11-1-0168). We thank an anonymous referee for suggesting a number of corrections to an earlier draft.

Thank you Hira for being a great teacher, friend and colleague! Over the years, I have enjoyed and learned from various interactions with you. Wish you many more healthy and productive years to come! Hope you enjoy reading this paper. (S.D.)

References

Aalen OO (1989) A linear regression model for the analysis of lifetimes. Stat Med 8:907–925

Akritas MG, Van Keilegom I (2003) Estimation of bivariate and marginal distributions with censored data. *J Royal Stat Soc*. Ser B 65:457–441

Beran R (1981) Nonparametric regression with randomly censored survival data. Technical report. University of California, Berkeley

Dabrowska DM (1988) Kaplan Meier estimate on the plane. Ann Stat 18:308–325

Dai H, Fu B (2012) A polar coordinate transformation for estimating bivariate survival functions with randomly censored and truncated data. J Stat Plan Inference 142:248–262

Datta S (2005) Estimating the mean life time using right censored data. Stat Methodol 2:65–69

Hanley JA, Parnes MN (1983) Nonparametric estimation of a multivariate distribution in the presence of censoring. Biometrics 39:129–139

Horvitz DG, Thompson DJ (1952) A generalization of sampling without replacement from a finite universe. J Amer Stat Assoc 47:663–685

Koul H, Susarla Y, Van Ryzin J (1981) Regression analysis with randomly right censored data. Ann Stat 9:1276–1288

Li G, Datta S (2001) A bootstrap approach to nonparametric regression for right censored data. Ann Inst Stat Math 53:708–729

Lin DY, Ying Z (1993) A simple nonparametric estimator of the bivariate survival function under univariate censoring. Biometrika 80:573–581

Meira-Machado L, Una-Alvarez J, Datta, S (2013) Nonparametric estimation of conditional transition probabilities in a non-Markov illness-death model. Preprint

McGilchrist CA, Aisbett CW.: Regression with frailty in survival analysis. Biometrics 47:461–466 (1991)

Prentice RL, Cai J (1992) Covariance and survivor function estimation using censored multivariate failure time data. Biometrika 79:495-512

Pruitt RC (1991) Strong consistency of self-consistent estimators: General theory and an application to bivariate survival analysis. Technical Report 543, University of Minnesota, MN, USA

Rotnitzky A, Robins JM (2005) Inverse probability weighted estimation in survival analysis. In: Amitage P, Colton T (ed) Encyclopedia of Biostatistics, 2nd edn Wiley, New York

Satten GA, Datta S (2001) The Kaplan-Meier estimator as an inverse-probability-of-censoring weighted average. Am Stat 55:207–210

Satten GA, Datta S, Robins JM (2001) Estimating the marginal survival function in the presence of time dependent covariates. Stat Probab Lett 54:397–403

Shen P (2010) Nonparametric estimation of the bivariate survival function for one modified form of doubly censored data. Comput Stat 25:203–213

Tsai WY, Crowley J (1998) A note on nonparametric estimators of the bivariate survival function under univariate censoring. Biometrika 85:573–580

van der Laan MJ (1996) Efficient estimation of the bivariate censoring model and repairing NPMLE. Ann Stat 24:596–627

van der Laan MJ (1997) Nonparametric estimators of the bivariate survival function under random censoring. Stat Neerl 51:178–200

Wang J, Zafra P (2009) Estimating bivariate survival function by Volterra estimator using dynamic programming techniques. J Data Sci 7:365–380

Wang WJ, Wells MT (1997) Nonparametric estimators of the bivariate survival function under simplified censoring conditions. Biometrika 84:863–880

Yang G (2005) A new bivariate survival function estimator under random right censoring, MS Thesis. Department of Statistics, University of Georgia, Athens

Chapter 6
On Equality in Distribution of Ratios $X/(X+Y)$ and $Y/(X+Y)$

Manish C. Bhattacharjee and Sunil K. Dhar

6.1 Introduction and Summary

One often comes across problems where the probability of male and that of female each being equal to 50 % is questioned. This question can be thought of in terms of the human sex ratio of $X : Y$ (which is currently 101 male to 100 female, CIA Fact Book, 2013) and the corresponding proportions being same to that of their corresponding distributions being identical. In this context, X and Y are thought to be nonnegative random variables. However, if the X and Y are independent identically distributed i.i.d.; it is well-known that the ratios $X/(X + Y)$ and $Y/(X + Y)$ are equal in distribution. This prompts the question: if we remove the assumption of mutual independence of X and Y, can the equidistribution of these ratios still hold, and under what reasonable conditions? In what follows, we explore some general answers to this question. We show that, if X and Y have the same distribution then $\frac{X}{X+Y}$ need not have the same distribution as $\frac{Y}{X+Y}$ and identify sufficient conditions for an affirmative answer. Extension of our main result to the case of n-dimensional random vectors $(X_1, \cdots, X_n)$ for $n \geq 2$ is indicated.

Generically, the cumulative distribution function (c.d.f.) of a random vector (X, Y) is denoted by $F_{X,Y}$ and its probability density function (p.d.f.), when it exists, by $f_{X,Y}$. For higher dimensional random vectors $(X_1, \cdots, X_n)$, $n \geq 2$; $F_{X_1,\cdots,X_n}$ and $f_{X_1,\cdots,X_n}$ correspondingly denote its c.d.f. and p.d.f., respectively. We use $\stackrel{d}{=}$ to denote equality in distribution of (r.v.s).

S. K. Dhar (✉) · M. C. Bhattacharjee
Center for Applied Mathematics & Statistics, Department of Mathematical Sciences,
New Jersey Institute of Technology, NJ 07102, Newark, United States
e-mail: dhar@njit.edu

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_6,

6.2 Counterexample

We show a *counterexample* to demonstrate that $X \stackrel{d}{=} Y$ does not guarantee equality of distribution of the ratios $\frac{X}{X+Y}$ and $\frac{Y}{X+Y}$. For this purpose, we use a suitable joint density of (X, Y), that we construct via the standard normal density

$$\phi(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right), \qquad -\infty < x < \infty.$$

Consider the joint density function on $R^2 = (-\infty, \infty) \times (-\infty, \infty)$, given by

$$f_{X,Y}(x, y) = [1 + xy\phi(x)\phi^2(y)]\phi(x)\phi(y).$$

To see that $f_{X,Y}$ is a valid joint density we need to observe that $\phi(x) < 1$ and that $|x\phi(x)| < 1$ (because $\frac{x^2}{2\pi} < exp(x^2)$). This in turn gives $1 + xy\phi(x)\phi^2(y) > 0$ and the fact that the mean of a scaled standard normal random variable is zero, which make $f_{X,Y}$ a valid density and both the marginals to be standard normal. Hence, X and Y have the same distribution. We will now derive the density of $V = \frac{Y}{X+Y}$ and then show that densities of V and $1 - V = \frac{X}{X+Y}$ are not the same. Let $W = X$ and $Y = \frac{VW}{1-V}$. The absolute value of the Jacobian is given by $\frac{|w|}{(1-v)^2}$. Hence, the joint density $f_{W,V}$ of (W, V) on the R^2 plane is given by,

$$f_{W,V}(w, v) = f_{X,Y}\left(w, \frac{wv}{(1-v)}\right) \frac{|w|}{(1-v)^2},$$

which simplifies to,

$$\frac{|w|}{2\pi(1-v)^2}\left[1 + \left(\frac{w^2 v}{1-v}\right)\frac{\exp\left(\frac{-w^2}{2} - \frac{w^2v^2}{(1-v)^2}\right)}{(2\pi)^{3/2}}\right]$$

$$\times \exp\left(\frac{-w^2}{2}\right)\exp\left(-\frac{w^2v^2}{2(1-v)^2}\right).$$

In the above joint density, we integrate out the w variable, to get the marginal density of V. Note that a closed form of the density of V can be obtained by using the facts that if N is a normal random variable with mean zero and variance σ_N^2 then $E|N| = \sqrt{\frac{2}{\pi}}\sigma_N$ and $E|N|^3 = 2\sqrt{\frac{2}{\pi}}\sigma_N^3$. Hence, the density of V is given by

$$f_V(v) = \int_{-\infty}^{\infty} f_{U,V}(u, v)du = \frac{1}{\pi(v^2 + (1-v)^2)} + \frac{v(1-v)}{\sqrt{2}\pi^{5/2}(2(1-v)^2 + 3v^2)^2}.$$

Clearly, $f_V(v) \neq f_V(1-v)$, and the latter is the density of $U := \frac{X}{X+Y}$. The two ratios U and V are not equal in distribution.

Dependence between X and Y in the counterexample does not establish the necessity of their statistical independence for the equality in distribution of the ratios U, V to hold. In fact, our results are typically based on the assumption of a joint distribution, and cover independence as a special case.

6.3 Main Results

For a random vector (X, Y), denote the ratios of the two component r.v.s to their sum, by

$$U := \frac{X}{X+Y}, \qquad V := \frac{Y}{X+Y}. \tag{6.1}$$

It may be noted that while $U + V = 1$, the r.v.s U and V cannot be thought of as the proportional contribution of the components of (X, Y) to their sum, as is obvious from the preceding counterexample.

If X, Y are absolutely continuous with a (joint) density, then so are U and V, with their respective densities related via

$$f_V(v) = f_U(1-v). \tag{6.2}$$

Standard calculations yield an expression for the density of U. In particular, choosing the transformation

$$U = \frac{X}{X+Y}, \qquad T = X + Y\ ;$$

the joint density of (U, T) is easily seen to be $f_{U,T}(u,t) = f_{X,Y}(ut,\ (1-u)t)\ |t|$, so that the marginal density of U is

$$f_U(u) = \int_{-\infty}^{\infty} f_{X,Y}\Big(ut\ ,\ (1-u)t\Big)\ |t|\ dt, \tag{6.3}$$

which together with (6.2) implies

$$\begin{aligned} f_V(v) &= \int_{-\infty}^{\infty} f_{X,Y}\Big((1-v)t\ ,\ vt\Big)\ |t|\ dt \\ &\neq \int_{-\infty}^{\infty} f_{X,Y}\Big(vt\ ,\ (1-v)t\Big)\ |t|\ dt = f_U(v), \quad -\infty < v < \infty, \end{aligned}$$

in general.

Define H to be *symmetric* in its arguments (x, y), if

$$H(x,y) = H(y,x), \text{all}\,(x,y).$$

If, however, $f_{X,Y}$ has this symmetry, then the earlier equality obviously holds. We thus have the following proposition.

Proposition 6.1 *If (X, Y) admits a joint density that is symmetric in its arguments, then the ratios in* (6.1) *are equal in distribution* ($U \stackrel{d}{=} V$).

Remark 1. There is no explicit assumption that $X \stackrel{d}{=} Y$ in the premise of the earlier proposition, as it is an easy consequence of the symmetry; viz,

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x, y)\, dy = \int_{-\infty}^{\infty} f_{X,Y}(y, x)\, dy = f_Y(x).$$

Remark 2. In view of the Remark 1 earlier, in the absolutely continuous case, the classic result that X, Y i.i.d. implies $U \stackrel{d}{=} V$ follows as a special case of proposition 6.1, since if X, Y are i.i.d. with a common p.d.f. $f_X(\cdot) \equiv f_Y(\cdot)$, then the joint p.d.f. satisfies

$$f_{X,Y}(x, y) = f_X(x) f_Y(y) = f_Y(x) f_X(y) = f_{X,Y}(y, x).$$

While proposition 6.1 provides an answer to our question when X, Y are absolutely continuous, an affirmative answer in the general case, where the joint c.d.f. of X, Y may also have discrete or/and singular components, is given by our next proposition. Note that $F(x, y)$ being symmetric in (x, y) implies that $P\{(X, Y) \in (-\infty, x] \times (-\infty, y]\} = P\{(Y, X) \in (-\infty, x] \times (-\infty, y]\}$ for all $(x, y) \in R^2$. This, in turn implies that $(X, Y) \stackrel{d}{=} (Y, X)$.

Proposition 6.2 *If the joint c.d.f. $F_{X,Y}(x, y)$ is symmetric in (x, y), then $U \stackrel{d}{=} V$.*

Proof. With $F_{X,Y}(x, y)$ also denoting the Lebesgue–Stieltjes measure on the plane induced by the joint c.d.f., we have,

$$\begin{aligned} E(e^{itU}) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(it\left(\frac{x}{x+y}\right)\right) dF_{X,Y}(x, y) \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} \exp\left(it\left(1 - \frac{y}{x+y}\right)\right) dF_{X,Y}(y, x) \\ &= E(e^{it(1-U)}) = E(e^{itV}), \qquad -\infty < t < \infty, \end{aligned} \tag{6.4}$$

where the second equality uses the symmetry condition of the joint c.d.f. and the two corresponding measures are the same because they are seen to be same of the relatively determining class of sets $(-\infty, x] \times (-\infty, y]$. Thus, the ratios U and V having the same *characteristic function* and therefore must be equal in distribution. Alternately, $F_{X,Y}(x, y) = F_{X,Y}(y, x)$ implies that $(X, Y) \stackrel{d}{=} (Y, X)$ and $h(x, y) = \frac{x}{x+y}$ being a continuous function gives $h(X, Y) \stackrel{d}{=} h(Y, X)$. Interestingly, converse of Proposition 6.2 is not true namely, $X/(X+Y) \stackrel{d}{=} Y/(X+Y)$ does not imply that X and Y have symmetric distribution functions. To see this, let (X, Y) take on the bivariate pairs (1,2) and (4,2) with probability 1/2 each. Then $X/(X + Y)$ and $Y/(X + Y)$ both have identical distributions, taking on the values 1/3 and 2/3 with probability 1/2 each. Yet, $1/2 = P[X = 1, Y = 2] \neq P[X = 2, Y = 1] = 0$.

The joint c.d.f.'s symmetry condition was motivated by the corresponding assumption in Proposition 6.1 and the following observation.

Lemma 6.3

(i) Suppose X, Y are absolutely continuous. Then $F_{X,Y}$ is symmetric in its arguments (x, y) if and only if so is $f_{X,Y}$.
(ii) The symmetry condition in Proposition 6.2 implies X and Y are identically distributed.

Proof.

(i) Suppose $f_{X,Y}$ is symmetric in (x, y). Then the nonnegativity of the integrand and Fubini's theorem implies,

$$\begin{aligned} F_{X,Y}(x,y) = P(X \le x,\ Y \le y) &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(u,v)\,dv\,du \\ &= \int_{-\infty}^{x}\int_{-\infty}^{y} f_{X,Y}(v,u)\,dv\,du \\ &= \int_{-\infty}^{y}\int_{-\infty}^{x} f_{X,Y}(v,u)\,du\,dv \\ &= P(X \le y,\ Y \le x) \equiv F_{X,Y}(y,x). \end{aligned}$$

Conversely, supposing $F_{X,Y}$ is symmetric in its argument (x, y), and has a joint density; we have,

$$f_{X,Y}(x,y) = \frac{\partial^2}{\partial x, \partial y} F_{X,Y}(x,y) = \frac{\partial^2}{\partial x, \partial y} F_{X,Y}(y,x) = f_{X,Y}(y,x).$$

(ii) Using the pointwise symmetry of $F_{X,Y}(\cdot,\cdot)$ on R^2,

$$P(X \le x) = \lim_{y\to\infty} F_{X,Y}(x,y) = \lim_{y\to\infty} F_{X,Y}(y,x) = P(Y \le x).$$

Remark 3. The symmetry condition in Proposition 6.2 is of course equivalent to X, Y being "*exchangeable*", i.e., $(X, Y) \stackrel{d}{=} (Y, X)$. For a pair of r.v.s however, it is much more simply stated as the property that the joint c.d.f. $F_{X,Y}(\cdot,\cdot) : R^2 \longrightarrow [0, 1]$ is symmetric in its arguments. For random vectors of higher dimensions, the corresponding condition that the c.d.f. $F_{X_1,\cdots,X_n}$ is permutation invariant in its arguments is more succinctly and elegantly descried as $X_1, \cdots, X_n$ being exchangeable; thus generalizing our earlier proposition as follows.

Proposition 6.4 *If $X_1, \cdots, X_n$ $(n \ge 2)$ is a finite, exchangeable sequence, then*

$$\frac{X_j}{S_n} \stackrel{d}{=} \frac{X_k}{S_n}, \qquad j,k \in \{1, 2, \cdots, n\},\ j \ne k$$

where $S_n := \sum_{i=1}^{n} X_i$.

Proof. Suppose $X_1, \cdots, X_n$ $(n \geq 2)$ are exchangeable, i.e., $(X_{i_1}, \cdots, X_{i_n}) \stackrel{d}{=} (X_1, \cdots, X_n)$ for all permutations $(i_1, \cdots, i_n)$ of $(1, \cdots, n)$. For brevity, denote by

$$\begin{aligned} \boldsymbol{X} &:= (X_1, \cdots, X_n), \text{and} \\ 0_j\boldsymbol{X} &:= (X_1, \cdots, X_{j-1}, X_{j+1}, \cdots, X_n), \end{aligned}$$

be the corresponding vector that skips the j-th coordinate X_j, and the corresponding values assumed as, $\boldsymbol{x}$ and $0_j\boldsymbol{X}$, respectively. When

$$\begin{aligned} E\left\{\exp\left(it\frac{X_j}{S_n}\right)\right\} &= \int_{-\infty}^{\infty} \exp\left(it\frac{u}{s_n}\right) dF_{\boldsymbol{X}}(x_1, \cdots, x_{j-1}, u, x_{j+1}, \cdots, x_n) \\ &= \int_{-\infty}^{\infty} \exp\left(it\frac{u}{s_n}\right) dF_{(X_j, 0_j\boldsymbol{X})}(u, 0_j\boldsymbol{x}) \\ &= \int_{-\infty}^{\infty} \exp\left(it\frac{u}{s_n}\right) dF_{(X_k, 0_k\boldsymbol{X})}(u, 0_k\boldsymbol{x}) \\ &= E\left\{\exp\left(it\frac{X_k}{S_n}\right)\right\}, \end{aligned}$$

where the value s_n of S_n is given by $s_n = u + \sum_{i=1_{i\neq j}}^{n} x_i$ or $s_n = u + \sum_{i=1_{i\neq k}}^{n} x_i$ in the second or third integrands earlier, respectively. Note, the two equalities preceding the last step hold, since $(X_j, 0_j\boldsymbol{X}) \stackrel{d}{=} \boldsymbol{X} \stackrel{d}{=} (X_k, 0_k\boldsymbol{X})$ for all pairs j, k, by exchangeability. Alternately, since $(X_j, 0_j\boldsymbol{X}) \stackrel{d}{=} (X_k, 0_k\boldsymbol{X})$ and $h(\boldsymbol{x}) = \frac{x_1}{s_n}$ is a continuous function, $h(X_j, 0_j\boldsymbol{X}) \stackrel{d}{=} h(X_k, 0_k\boldsymbol{X})$. Hence the result.

In conclusion, any *Archimedian copula* can be used as a generator of such exchangeable r.v.s, Nelson (1999) and Genest et al. (1986). These results are also applicable to Bayesian contexts, where the observations are conditionally i.i.d. given an environmental variable with a prior distribution.

Acknowledgements We thank an anonymous reviewer for improving the presentation of this paper. Prof. Hira Lal Koul is my "statistics-guru". I wish to thank him for all the knowledge he has given me.

References

The Central Intelligence Agency of the United States (2013) CIA Fact Book. https://www.cia.gov/library/publications/the-world-factbook/index.html

Nelsen RB (1999) An Introduction to Copulas. Lecture Notes in Statistics, # 139; Springer, New York

Genest C, MacKay J (1986) The joy of copulas: bivariate distributions with uniform marginals. Amer Statist 40:280–285

Chapter 7
Nonparametric Distribution-Free Model Checks for Multivariate Dynamic Regressions

J. Carlos Escanciano and Miguel A. Delgado

7.1 Introduction

Parametric time series regression models continue being attractive among practitioners because they describe, in a concise way, the relation between the response or dependent variable and the explanatory variables. Much of the existing statistical literature is concerned with the parametric modelling in terms of the conditional mean function of a response variable $Y_t \in \mathbb{R}$, given some conditioning variable at time $t-1$, $I_{t-1} \in \mathbb{R}^d$, $d \in \mathbb{N}$, say. More precisely, let $Z_t \in \mathbb{R}^m$, $m \in \mathbb{N}$, be a m-dimensional observable random variable (r.v) and $W_{t-1} = (Y_{t-1}, \ldots, Y_{t-s}) \in \mathbb{R}^s$. The conditioning set we consider at time $t-1$ is given by $I_{t-1} = (W'_{t-1}, Z'_t)'$, so $d = s+m$. We assume throughout the article that the time series process $\{(Y_t, Z'_t)' : t = 0, \pm 1, \pm 2, \ldots\}$ is strictly stationary and ergodic. Henceforth, A' denotes the matrix transpose of A.

It is well-known that under integrability of Y_t, we can write the tautological expression

$$Y_t = f(I_{t-1}) + \varepsilon_t,$$

where $f(z) = E[Y_t \mid I_{t-1} = z]$, $z \in \mathbb{R}^d$, is the conditional mean function almost surely (a.s.) of Y_t, given $I_{t-1} = z$, and $\varepsilon_t = Y_t - E[Y_t \mid I_{t-1}]$ satisfies, by construction, that $E[\varepsilon_t \mid I_{t-1}] = 0$ a.s.

Then, in parametric modelling one assumes the existence of a parametric family of functions $\mathcal{M} = \{f(\cdot, \theta) : \theta \in \Theta \subset \mathbb{R}^p\}$ and considers the following regression model

$$Y_t = f(I_{t-1}, \theta) + e_t(\theta), \tag{7.1}$$

AMS 2000 subject classification. 62M07, 62G09, 62G10.

M. A. Delgado (✉)
Universidad Carlos III de Madrid, Madrid, Spain
e-mail: delgado@est-econ.uc3m.es

J. C. Escanciano
Indiana University, Bloomington, USA
e-mail: jescanci@indiana.edu

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_7,

with $f(I_{t-1},\theta)$ a parametric specification for the conditional mean $f(I_{t-1})$, and $\{e_t(\theta) : t = 0, \pm 1, \pm 2, \ldots\}$ a sequence of r.v.'s, deviations of the model. Model (7.1) includes classes of linear and nonlinear regression models and linear and nonlinear autoregression models, such as Markov-switching, exponential or threshold autoregressive models, among many others (see Fan and Yao 2003).

The condition $f(\cdot) \in \mathcal{M}$ is tantamount to

$$H_0 : E[e_t(\theta_0) \mid I_{t-1}] = 0 \quad \text{a.s. for some } \theta_0 \in \Theta \subset \mathbb{R}^p.$$

We aim to test H_0 against the alternative hypothesis

$$H_A : P(E[e_t(\theta) \mid I_{t-1}] \neq 0) > 0, \text{for all } \theta \in \Theta \subset \mathbb{R}^p,$$

where $(\Omega, \mathcal{F}, P)$ is the probability space in which all the r.v.'s of this article are defined.

There is a vast literature on testing the correct specification of regression models. In an independent and identically distributed (i.i.d) framework, some examples of those tests have been proposed by Bierens (1982, 1990), Eubank and Spiegelman (1990), Eubank and Hart (1992), Härdle and Mammen (1993), Horowitz and Härdle (1994), Hong and White (1995), Fan and Li (1996), Zheng (1996), Stute (1997), Stute et al. (1998), Li and Wang (1998), Fan and Huang (2001), Horowitz and Spokoiny (2001), Li (2003), Khamaladze and Koul (2004), Guerre and Lavergne (2005) and Escanciano (2006a), to mention a few. Whereas, in a time series context some examples are Bierens (1984), Li (1999), de Jong (1996), Bierens and Ploberger (1997), Koul and Stute (1999), Chen et al. (2003), Stute et al. (2006) and Escanciano (2006b, 2007). This extensive literature can be divided into two approaches. In the first approach test statistics are based on nonparametric estimators of the local measure of dependence $E[e_t(\theta_0) \mid I_{t-1}]$. This local approach requires smoothing of the data in addition to the estimation of the finite-dimensional parameter vector θ_0, and leads to less precise fits, see Hart (1997) for some review of the local approach when $d = 1$. Tests within the local approach are in general asymptotic distribution-free (ADF).

The second class of tests avoids smoothing estimation by means of an infinite number of unconditional moment restrictions over a parametric family of functions, i.e., it is based on the equivalence

$$E[e_t(\theta_0) \mid I_{t-1}] = 0 \; a.s. \iff E[e_t(\theta_0) w(I_{t-1}, x)] = 0, \quad \text{almost everywhere (a.e.) in } \Pi \subset \mathbb{R}^q, \tag{7.2}$$

where $\Pi \subset \mathbb{R}^q$, $q \in \mathbb{N}$, is a properly chosen space, and the parametric family of functions $\{w(\cdot, x) : x \in \Pi\}$ is such that the equivalence (7.2) holds, see Stinchcombe and White (1998) and Escanciano (2006b) for primitive conditions on the family $\{w(\cdot, x) : x \in \Pi\}$ to satisfy this equivalence. We call the approach based on (7.2) the "integrated approach". In the integrated approach, test statistics are based on a distance from the sample analogue of $E[e_t(\theta_0) w(I_{t-1}, x)]$ to zero. This integrated approach is well known in the literature and was first proposed by Bierens (1982),

who used the exponential function $w(I_{t-1}, x) = \exp(ix'I_{t-1})$, where $i = \sqrt{-1}$ denotes the imaginary unit, see also Bierens (1990) and Bierens and Ploberger (1999). Stute (1997) using empirical process theory, proposed to use the indicator function $w(I_{t-1}, x) = 1(I_{t-1} \leq x)$ in an i.i.d context. Stinchcombe and White (1998) emphasized that there are many other possibilities in the choice of w. Recently, Escanciano (2006a) has considered in an i.i.d setup the family $w(I_{t-1}, x) = 1(\beta' I_{t-1} \leq u)$, $x = (\beta', u)' \in \Pi_{pro}$, where $\Pi_{pro} = \mathbb{S}^d \times [-\infty, \infty]$ is the auxiliary space with $\mathbb{S}^d$ the unit ball in $\mathbb{R}^d$, i.e., $\mathbb{S}^d = \{\beta \in \mathbb{R}^d : |\beta| = 1\}$. This new family combines the good properties of exponential and indicator families and delivers a Cramér-von Mises (CvM) test simple to compute and with excellent power properties in finite samples, see Escanciano (2006a) for further details. Escanciano (2007) provides a unified theory for specification tests based on the integrated approach for a general weighting function w, including but not restricting to indicators and exponential families.

A tenet in the integrated approach is that the asymptotic null distribution of resulting tests depends on the data generating process (DGP), the specified model and generally on the true parameter θ_0. Consequently, critical values for integrated tests have to be approximated with the assistance of resampling methods. In particular, Escanciano (2007) justified theoretically a wild bootstrap method to approximate the asymptotic critical values for general integrated-based tests. In contrast, Koul and Stute (1999) avoided resampling procedures by means of a martingale transformation in the spirit of that initially proposed by Khamaladze (1981). However, Koul and Stute's setup was restricted to homocedastic autoregressive models of order 1. Recently, Khamaladze and Koul (2004) have applied the martingale transform to residual marked processes in multivariate regressions with i.i.d data, but the resulting test is not ADF since it depends on the joint distribution of regressors. The main contribution of this article is to complement these approaches and extend them to heteroskedastic multivariate time series processes. We apply the martingale transform coupled with the Rossenblatt's transform on the multivariate regressors to get ADF test free of the joint design distribution. We formally justify the effect of these transformations on our test statistics using new asymptotic theory of function-parametric empirical processes under martingale conditions. Finally, we compare via a Monte Carlo experiment, our new model checks with existing bootstrap approximations.

The layout of the article is as follows. In Sect. 2 we present the ADF tests based on continuous functionals of a martingale transform of the function-parametric residual marked empirical process. We begin by establishing some heuristics for the martingale transform. In Sect. 3 we establish the asymptotic distribution of our test under the null. In Sect. 4 we compare the bootstrap approach with the martingale approach via a Monte Carlo experiment. Proofs are deferred to an appendix.

A word on notation. In the sequel C is a generic constant that may change from one expression to another. Throughout, $|A|$ denotes the Euclidean norm of A. $\overline{\mathbb{R}}^d$ denotes the extended d-dimensional Euclidean space, i.e., $\overline{\mathbb{R}}^d = [-\infty, \infty]^d$. Let $\|X\|_p$ be the L_p-norm of a r.v X, i.e., $\|X\|_p = (E|X|^p)^{1/p}$, $p \geq 1$. Let $N_{[]}(\varepsilon, \mathcal{H}, \|\cdot\|_p)$ be the ε-bracketing number of a class of functions $\mathcal{H}$ with respect to the norm $\|\cdot\|_p$, i.e., the minimal number N for which there exist ε-brackets $\{[l_j, u_j] : \|l_j - u_j\|_p \leq \varepsilon,$

$\|l_j\|_p < \infty$, $\|u_j\|_p < \infty$, $j =, 1 \ldots, N\}$ covering $\mathcal{H}$, see Definition 2.1.6 in van der Vaart and Wellner (1996). Let $\ell^\infty(\mathcal{H})$ be the metric space of all real-valued functions that are uniformly bounded on $\mathcal{H}$. As usual, $\ell^\infty(\mathcal{H})$ is endowed with the sup norm, i.e., $\|z\|_{\mathcal{H}} = \sup_{h\in\mathcal{H}} |z(h)|$. Let $\Longrightarrow$ denote weak convergence on $\ell^\infty(\mathcal{H})$, see Definition 1.3.3 in van der Vaart and Wellner (1996). Throughout the article, weak convergence on compacta in $\ell^\infty(\mathcal{H})$ means weak convergence on $\ell^\infty(\mathcal{C})$ for all compact subsets $\mathcal{C} \subset \mathcal{H}$. Also $\xrightarrow{P^*}$ and $\xrightarrow{as^*}$ denote convergence in outer probability and outer almost surely, respectively, see Definition 1.9.1 in Vaart and Wellner (1996). The symbol $\rightarrow_d$ denotes convergence in distribution of Euclidean random variables. All limits are taken as the sample size $n \rightarrow \infty$.

7.2 The Function-Parametric Residual Process and the Martingale Transform

In view of a sample $\{(Y_t, I'_{t-1})' : 1 \leq t \leq n\}$, and motivated from (7.2), we define the function-parametric empirical process,

$$R_n(b, \theta) = n^{-1/2} \sum_{t=1}^{n} e_t(\theta) b(I_{t-1}),$$

indexed by $(b, \theta) \in \mathcal{B} \times \Theta$, for a class of "check" functions $\mathcal{B}$ and a parameter space Θ. Examples of $\mathcal{B}$ will be specified later. Two important processes associated to $R_n(b, \theta)$ are the error-marked process $R_n(b) = R_n(b, \theta_0)$ and the residual-marked process

$$R_n^1(b) \equiv R_n(b, \theta_n) = n^{-1/2} \sum_{t=1}^{n} e_t(\theta_n) b(I_{t-1}),$$

where θ_n is a $\sqrt{n}$-consistent estimator for θ_0 (see Assumption A4 below). For convenience, we shall assume that $\mathcal{B} \subset L_2(\overline{\mathbb{R}}^d, G)$, the Hilbert space of all G -square integrable measurable functions, where $G(dx) = \sigma^2(x)F(dx)$, $F(\cdot)$ is the joint cumulative distribution function (cdf) of I_{t-1}, and $\sigma^2(\cdot)$ is the conditional error variance, i.e., $\sigma^2(y) = E[\varepsilon_t^2 \mid I_{t-1} = y]$. As usual, $L_2(\overline{\mathbb{R}}^d, G)$ is furnished with the inner-product

$$\langle f, g \rangle = \int_{\mathbb{R}^d} f(x) g(x) G(dx).$$

and the induced norm $\|h\| = \langle h, h \rangle^{1/2}$.

The aim of this section is to construct a suitable check space $\mathcal{B}$ such that the process $R_n^1(b)$, with $b \in \mathcal{B}$, delivers tests based on test statistics, $\Gamma(R_n^1)$ say, which are consistent and ADF. In this article we shall focus in a particular check space that

makes use of the martingale transformation proposed by Khmaladze (1981, 1993) for the problem of goodness-of-fit tests of distributions.

Let $g(I_0,\theta_0) = (\partial/\partial\theta')f(I_0,\theta_0)$ and $s(I_0,\theta_0) = \sigma^{-2}(I_0)g(I_0,\theta_0)$ be the non-standardized and standardized scores, respectively. From Theorem 1 in Sect. 3, under the null hypothesis and some mild regularity conditions, we have the following relation between $R_n(b)$ and $R_n^1(b)$, uniformly in $b \in \mathcal{B}$,

$$R_n^1(b) = R_n(b) - \langle b, s'\rangle\sqrt{n}(\theta_n - \theta_0) + o_P(1). \tag{7.3}$$

This relation gives us a clue about how to choose b for the test based on $R_n^1(b)$ being ADF. Namely, if b is orthogonal to the score, i.e., $\langle b, s'\rangle = 0$, we have the uniform representation

$$R_n^1(b) = R_n(b) + o_P(1),$$

and the estimation of θ_0 does not have any effect in the asymptotic null distribution of $R_n^1(b)$. Furthermore, it can be shown that the limit process of $R_n(b)$ is a standard function-parametric Brownian motion in $L_2(\overline{\mathbb{R}}^d, G)$, that is, a Gaussian process with zero mean and covariance function $\langle b_1, b_2\rangle$. Following ideas from Khmaladze (1993), a simple way to make b orthogonal to the score is to use a transformation $\mathcal{T}$ from $L_2(\overline{\mathbb{R}}^d, G)$ to $L_2(\overline{\mathbb{R}}^d, G)$ with values in the orthogonal complement of the space generated by the score s, and consider the transformed process $R_n^1(\mathcal{T}b)$. The covariance function of the limit process of $R_n^1(\mathcal{T}b)$ is then $\langle \mathcal{T}b_1, \mathcal{T}b_2\rangle$, so unless $\mathcal{T}$ is an isometry (i.e., $\langle \mathcal{T}b_1, \mathcal{T}b_2\rangle = \langle b_1, b_2\rangle$), the Brownian motion structure is lost. Therefore, we observe that a way to make the asymptotic null distribution "immune" to the estimation effect and, at the same time, preserve the original covariance structure is to consider $R_n^1(\mathcal{T}b)$, where $\mathcal{T}$ is an isometry with image orthogonal to the score. In other words, a suitable check space to obtain consistent and ADF tests is $\mathcal{B} = \{\mathcal{T}h : h \in \mathcal{H}\}$, for an isometry $\mathcal{T}$ with image orthogonal to the score (to obtain the ADF property) and with suitable large class of functions $\mathcal{H}$ (to obtain consistency in the test procedure).

A large class of isometries with the previous properties is the class of shift isometries. Let $bas = \{s, f_1, f_2, \ldots\}$ be an orthogonal basis of $L_2(\overline{\mathbb{R}}^d, G)$. Let us define the isometry $\mathcal{T}_{bas}$ in the following way

$$\mathcal{T}_{bas}s = f_1 \qquad \mathcal{T}_{bas}f_j = f_{j+1}, j > 1.$$

Then, it is easy to show that $\mathcal{T}$ is an isometry from $L_2(\overline{\mathbb{R}}^d, G)$ to $L_2(\overline{\mathbb{R}}^d, G)$ with values in the orthogonal complement of the score s. A remarkable example of a shift isometry is the Khmaladze's martingale transform (cf. Khmaladze 1981, 1993), that posseses the added property of having an explicit formula. We use the same notation as in Khmaladze and Koul (2004). Introduce the so called scanning family of measurable subsets $\mathcal{A} = \{A_\lambda : \lambda \in \mathbb{R}\}$ of $\overline{\mathbb{R}}^d$, such that

1: $A_z \subseteq A_u, \forall z \leq u$.

2: $G(A_{-\infty}) = 0$, $G(A_{\infty}) = 1$
3: $G(A_z)$ is a strictly increasing and absolutely continuous function of $z \in \mathbb{R}$.

An example of scanning family is the following. Assuming that $G(\beta' y)$ is absolutely continuous for some $\beta \in \overline{\mathbb{R}}^d$, then the family $\mathcal{A} = \{A_z : z \in \mathbb{R}\}$ with $A_z = \{y \in \overline{\mathbb{R}}^d : \beta' y \leq z\}$ is a scanning family. Now define $z(y) = \inf\{z : y \in A_z\}$ and

$$C_z = \int_{A_z^c} s(x, \theta_0) s'(x, \theta_0) G(dx),$$

where A_z^c is the complement of A_z. The linear operator T is given by

$$Tf(u) = f(u) - Kf(u), \tag{7.4}$$

where

$$Kf(u) = \int_{A_{z(u)}} f(x) s'(x, \theta_0) C_{z(x)}^{-1} G(dx) s(u, \theta_0) \tag{7.5}$$

and $f(\cdot) \in L_2(\overline{\mathbb{R}}^d, G)$. Such transformation was first proposed in the goodness-of-fit literature by Khmaladze (1981, 1993). In the statistical literature this transformation has been considered and extended to other problems in e.g. Stute et al. (1998), Koul and Stute (1999), Stute and Zhu (2002) or Koul and Khmaladze (2004). This transformation is becoming well-known in other areas and has been already applied to a variety of problems in Bai and Ng (2001), Koenker and Xiao (2002), Bai (2003), Delgado et al. (2008), Delgado and Stute (2008), Bai and Chen (2008), Song (2009, 2010) and Angrist and Kuersteiner (2011). It is not difficult to show that T defined by (7.4) is an isometry from $L_2(\overline{\mathbb{R}}^d, G)$ to $L_2(\overline{\mathbb{R}}^d, G)$ with values in the orthogonal complement of the score s, see Khmaladze and Koul (2004) for the proof.

The martingale transform[1] T depends on unknown quantities which can be estimated from a sample. The natural estimator of the transformation is

$$T_n f(u) = f(u) - \int_{A_{z(u)}} f(x) s_n'(x, \theta_n) C_{n,z(x)}^{-1} G_n(dx) s_n(u, \theta_n),$$

where

$$C_{n,z} = \int_{A_z^c} s_n(x, \theta_n) s_n'(x, \theta_n) G_n(dx),$$

with $G_n(dy) = \sigma_n^2(y) F_n(dy)$, F_n is the empirical cdf of $\{I_{t-1}\}_{t=1}^n$, $s_n(I_0, \theta) = \sigma_n^{-2}(I_0) g(I_0, \theta)$, θ_n is a $\sqrt{n}$-consistent estimator of θ_0, and $\sigma_n^2(y)$ is a consistent nonparametric estimator of $\sigma^2(y)$ (for instance, a Nadaraya-Watson estimator).

[1] The martingale trasform has also been variously referred to as: an innovation approach (Khmaladze, 1988), and an innovation process approach (Stute, Thies, and Zhu, 1998).

From the integrated approach we know that in the construction of consistent tests, it is not necessary to consider the whole space of functions $L_2(\overline{\mathbb{R}}^d, G)$. A parametric family that delivers well-known limit processes is the indicator class $\mathcal{B}_{ind} = \{1(\cdot \leq x) \equiv 1_x(\cdot) : x \in \overline{\mathbb{R}}^d\} \subset L_2(\overline{\mathbb{R}}^d, G)$. For the univariate case, i.e., $d = 1$, continuous functionals of standardizations of $R_n^1(T_n 1_x)$ deliver ADF tests for H_0, see Koul and Stute (1999). However, in the multivariate case, $d \geq 2$, the asymptotic null distribution of $R_n^1(T_n 1_x)$ still depends on the conditional variance and the design distribution. To overcome this problem we consider the so-called Rossenblatt's (1952) transformation. This transformation produces a multivariate distribution that is i.i.d on the d-dimensional unit cube, thereby, leading to tests that can be based on standardized tables. Let $I_t = (I_{t1}, I_{t2}, \ldots, I_{td})'$ and define the transformation $u = (u_1, \ldots, u_d)' = T_R(x)$ component-wise by $u_1 = F_1(x_1) = P(I_{t1} \leq x_1)$, $u_2 = F_2(x_2 \mid x_1) = P(I_{t2} \leq x_2 \mid I_{t1} = x_1), \ldots, u_d = F_d(x_d \mid x_1, \ldots, x_{d-1}) = P(I_{td} \leq x_d \mid I_{t1} = x_1, \ldots, I_{td-1} = x_{d-1})$. The inverse $x = T_R^{-1}(u)$ can be obtained recursively. Rossenblatt (1952) showed that $U_{t-1} = T_R(I_{t-1})$ has a joint distribution which marginals are uniform and independently distributed on $[0, 1]^d$.

In the next section, we shall show that under the null hypothesis and some mild regularity conditions the transformed process $J_n(u) = R_n^1(T_n(\sigma_n^{-1}(\cdot)1_u \circ T_R(\cdot)))$ converges weakly to a zero mean Gaussian process in $\ell^\infty(B_{x_0})$, for a suitable chosen set $B_{x_0} \subset [0, 1]^d$, with covariance function $u_1 \wedge u_2$, where for $a = (a_1, \ldots, a_d)'$ and $b = (b_1, \ldots, b_d)'$, $a \wedge b = \min\{a_1, b_1\} \times \cdots \times \min\{a_d, b_d\}$, that is, a standard Brownian sheet.

In practice the conditional distributions $F_1, \ldots, F_d$, are unknown and have to be estimated. Following Angrist and Kuersteiner (2004), we consider kernel estimators

$$\begin{aligned}
\widehat{F}_1(x_1) &= n^{-1}\sum_{t=1}^{n} 1(I_{t-11} \leq x_1) \\
&\vdots \\
\widehat{F}_d(x_d \mid x_1, \ldots, x_{d-1}) &= \frac{n^{-1}\sum_{t=1}^{n} 1(I_{t-1d} \leq x_d) K_{d-1}((x_d^- - I_{t-1d}^-)/h_n)}{n^{-1}\sum_{t=1}^{n} K_{d-1}((x_d^- - I_{t-1d}^-)/h_n)},
\end{aligned}$$

where $x_d^- = (x_1, \ldots, x_{d-1})'$, $I_{t-1d}^- = (I_{t-11}, \ldots, I_{t-1d-1})'$, $K_j(x) = (2\pi)^{-j/2} \sum_{h=1}^{w} \gamma_h |\sigma_h|^{-j} \exp(-0.5x'x/\sigma_h^2)$, $\sum_{h=1}^{w} \gamma_h = 1, \sum_{h=1}^{w} \gamma_h |\sigma_h|^{2\,l} = 0$, for $l = 1, 2, \ldots, w-1$, and $h_n = O(n^{-1/(2+d)})$ is a bandwidth sequence. Other higher order kernels or other nonparametric estimators are possible, as long as A6(ii) in the next section is satisfied.

Our final process is $\widehat{J}_n(u) = R_n^1(T_n(\sigma_n^{-1}(\cdot)1_u \circ \widehat{T}_R(\cdot)))$, where $\widehat{T}_R$ uses the previously described kernel estimation. $\widehat{J}_n(u)$ is called here the Khmaladze-Rossenblatt's transformed residual marked process. As a test statistic we consider in this article a

CvM functional

$$CvM_n = \int_{B_{x_0}} \left|\widehat{J}_n(u)\right|^2 F_{n,U}(du),$$

where $F_{n,U}(\cdot)$ is the empirical distribution function of the transformed sample $\{U_{t-1}\}_{t=1}^n$, $B_{x_0} = \{u \in [0,1]^d : \beta_1' T_R^{-1}(u) \leq x_0\}$, $\beta_1 \in \mathbb{R}^d$, and $x_0 < \infty$ is a user-chosen parameter necessary to avoid non-invertibility problems of the matrix $C_{n,z(x)}$, see Koul and Stute (1999) for a related situation. In the simulations we choose x_0 as the $(100-d)\%$ empirical quantile of the sample $\{\beta_1' I_{t-1}\}_{t=1}^n$. Other spaces B_{x_0}, threshold values x_0 and functionals different from the CvM are, of course, possible. Our test will reject the null hypothesis H_0 for "large" values of CvM_n. Next section establishes the asymptotic theory for CvM_n and Sect. 4 shows, via a Monte Carlo experiment, that it leads to a valuable diagnostic test.

7.3 Asymptotic Null Distribution

In this section we establish the limit distribution of $\widehat{J}_n$ under the null hypothesis H_0. First, we state a uniform representation for the function-parametric process $R_n^1(b)$, $b \in \mathcal{B}$, for a generic $\mathcal{B}$. This result is of independent interest. To derive these asymptotic results we consider the following notation and definitions. Let $\mathcal{F}_t = \sigma(I_t', I_{t-1}', \ldots, I_0')$ be the σ-field generated by the information set obtained up to time t. Let us endow $\mathcal{B}$ with the pseudo-metric $\|\cdot\|_{\mathcal{B}}$. Let us define $\mathcal{A} = \mathcal{B} \times \Theta$. For a given class of function $\mathcal{D}$ we define for $(r_1, r_2) \in \mathcal{D} \times \mathcal{D}$

$$d_{n,\mathcal{D}}^2(r_1, r_2) = n^{-1} \sum_{t=1}^n E\left[\varepsilon_t^2 \mid \mathcal{F}_{t-1}\right] |r_1(I_{t-1}) - r_2(I_{t-1})|^2$$

and

$$d_{\mathcal{D}}(r_1, r_2) = \|\varepsilon_t r_1(I_{t-1}) - \varepsilon_t r_2(I_{t-1})\|_2 .$$

Define the set $\Lambda_q = \{(r_1, r_2) \in \mathcal{D} \times \mathcal{D} : r_1 \leq r_2, d_{\mathcal{D}}^2(r_1, r_2) = 2^{-2q}\}$. If the family $\mathcal{D}$ satisfies that

$$\sup_{(r_1,r_2)\in\Lambda_q, q\in\mathbb{N}} \frac{d_{n,\mathcal{D}}^2(r_1, r_2)}{d_{\mathcal{D}}^2(r_1, r_2)} = O_P(1),$$

we say that $\mathcal{D}$ has bounded conditional quadratic variation with respect to $d_{\mathcal{D}}$. Also, we say that the class $\mathcal{D}$ satisfies a bracketing condition of order $p \geq 2$ and $s > 0$, in short $\mathcal{D}$ is $BEC(p, s)$, if

$$\int_0^\infty \left(\log\left(N_{[]}(\varepsilon^{1/s}, \mathcal{D}, \|\cdot\|_p)\right)\right)^{1/2} d\varepsilon < \infty.$$

The following assumptions are sufficient conditions for the weak convergence of $R_n^1(b)$ in $\ell^\infty(\mathcal{B})$ for a general $\mathcal{B}$.

Assumption A1: (*on the DGP*)

A1(a): $\{(Y_t, Z_t')' : t = 0, \pm 1, \pm 2, \ldots\}$ *is a strictly stationary and ergodic process.*

A1(b): $E[\varepsilon_t \mid \mathcal{F}_{t-1}] = 0$ *a.s. for all* $t \geq 1$, *and* $E|\varepsilon_1|^2 < C$.

Assumption A2: (*on the set of functions* $\mathcal{B}$)

A2(a): (*Locally Uniform* L_p*-Smoothness*) *Suppose that for some* $s > 0, C_1 > 0$, *and for* $p \geq 2$, *the following holds: for each* $b_1 \in \mathcal{B}$,

$$\left\| \sup_{b_2 \in \mathcal{B}: \|b_1 - b_2\|_{\mathcal{B}} < \delta} |\varepsilon_t b_1(I_{t-1}) - \varepsilon_t b_2(I_{t-1})| \right\|_p \leq C_1 \delta^s.$$

A2(b): (*control the size of* $\mathcal{B}$) *The class of functions* $\mathcal{B}$ is $BEC(p, s)$ *for* p *and* s *as in A2(a).*

A2(c): *The class* $\mathcal{B}$ *has bounded conditional quadratic variation with respect to* $d_{\mathcal{B}}$ *and the parametric space* Θ *is compact in* $\mathbb{R}^p$.

Assumption A3: (*on the model*) $f(\cdot, \theta)$ *is twice continuously differentiable in a neighborhood of* $\theta_0 \in \Theta$. *There exists a function* $M(I_{t-1})$ *with* $\sup_{\theta \in \Theta} \left| g(I_{t-1}, \theta) \right| \leq M(I_{t-1})$, *such that* $M(I_{t-1})$ *is* $F(\cdot)$*-square integrable.*

Assumption A4: (*on the parameter*)

A4(a): *The true parameter* θ_0 *belongs to the interior of* Θ. *There exists a unique* θ_1 *such that* $|\theta_n - \theta_1| = o_P(1)$.

A4(b): *The estimator* θ_n *satisfies* $\sqrt{n}(\theta_n - \theta_0) = O_P(1)$.

Assumption A1(a) is standard in the model checks literature under time series, see, e.g., Koul and Stute (1999). A1(b) is weaker than other related moment conditions in the literature and allows for most empirically relevant conditional heteroskedastic models. A2 is needed for the asymptotic tightness of the process $R_n^1(b)$. The bracketing entropy condition has been frequently used in the literature. Combined with locally uniform L_p-continuity, the bracketing entropy condition can be used to establish the stochastic equicontinuity of a process that involves non-smooth functions containing infinite dimensional parameters. Assumption A3 is classical in the model checks literature, see, e.g., Stute and Zhu (2002). Assumption A4 is satisfied for most estimators in the literature, such as the conditional nonlinear least squares estimator (NLSE), or its robust modifications (under further regularity assumptions), see Koul's (1992, 2002) monographs. Under H_0, a more efficient estimator than the NLSE (see Wefelmeyer 1996) is given by the M-estimator satisfying the equation

$$\sum_{t=1}^{n} \sigma^{-2}(I_{t-1}) g(I_{t-1}, \theta_n)(Y_t - f(I_{t-1}, \theta_n) = 0. \tag{7.6}$$

A4(a) and A4(b) imply that under the null $\theta_0 = \theta_1$, but they might be different under the alternative. A2(c) is a standard assumption to obtain weak convergence theorems under martingale assumptions, see Bae and Levental (1995) and Nishiyama (2000). Because this assumption is crucial in most of our asymptotic results, we now give primitive and simple-to-check conditions for a class of functions $\mathcal{D}$ being of

bounded conditional quadratic variation with respect to $d_{\mathcal{D}}$. See Escanciano and Mayoral (2010) for a related result. Let us define the quantity

$$G_t^{\mathcal{D}}(r) = E\left[E\left[\varepsilon_t^2 \mid I_{t-1}\right] r(I_{t-1}) \mid \mathcal{F}_{t-2}\right] \qquad r \in \mathcal{D},$$

Lemma 1: *Assume A1, A2(a-b) and that* $\left|G_t^{\mathcal{D}}(r_1) - G_t^{\mathcal{D}}(r_2)\right| \leq M_t d_{\mathcal{D}}^2(r_1, r_2)$, *where* M_t *is a stationary process with* $E[\,|M_1|\,] < \infty$. *Then,* $\mathcal{D}$ *has bounded conditional quadratic variation with respect to* $d_{\mathcal{D}}$.

Let V be a normal random vector with zero mean and variance–covariance matrix given by $L(\theta_0)$ (cf. A4(c)). Now, we are in position to state the asymptotic uniform representation of the process $R_n^1(b)$ and its weak convergence.

Theorem 1: *(i) Under Assumptions A1, A2 and A4(a) uniformly in* $b \in \mathcal{B}$,

$$\begin{aligned} R_n^1(\mathrm{b}) = & \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \{e_t(\theta_1) - E[e_t(\theta_1) \mid \mathcal{F}_{t-1}]\}\, b(I_{t-1}) \\ & + \frac{1}{\sqrt{n}} \sum_{t=1}^{n} \left\{ E[e_t(\theta) \mid \mathcal{F}_{t-1}]|_{\theta=\theta_n} - E[e_t(\theta_1) \mid \mathcal{F}_{t-1}] \right\} b(I_{t-1}) \\ & + \frac{1}{\sqrt{n}} \sum_{t=1}^{n} E[e_t(\theta_1) \mid \mathcal{F}_{t-1}] b(I_{t-1}) - E\left[E[e_t(\theta_1) \mid \mathcal{F}_{t-1}] b(I_{t-1})\right] \\ & + \sqrt{n} E\left[E[e_t(\theta_1) \mid \mathcal{F}_{t-1}] b(I_{t-1})\right] + o_P(1) \end{aligned}$$

(ii) If in addition, H_0, *A3 and A4(a) hold, then uniformly inb* $\in \mathcal{B}$,

$$R_n^1(b) = R_n(b) - \langle b, s' \rangle \sqrt{n}(\theta_n - \theta_0) + o_P(1).$$

The decomposition in Theorem 1(ii) paves the way for the discovery of appropriate martingale transforms of the residual marked process, see previous section. The analysis of function-parametric processes such as those considered in Theorem 1 provides simple methods of proof for the study of the asymptotic null distribution of $\widehat{J}_n$. To proceed further we need some regularity conditions.

Assumption A5: (*on the conditional variance and related quantities*)

A5(i): *The estimator* $\sigma_n^2(\cdot)$ *is a uniform consistent nonparametric estimator of* $\sigma^2(\cdot)$ *and* $0 < a \leq \sigma^2(y)$ for all $y \in \overline{\mathbb{R}}^d$ *and some positive* a.

A5(ii): $\sigma^{-j}(\cdot) \in \mathcal{W}$, $P(\sigma_n^{-j}(\cdot) \in \mathcal{W}) \to 1$ as $n \to \infty$ for $j = 1, 2$. *The class* $\mathcal{W}$ *satisfies A2(c), A2(a) for* $p > 2$ *and* $s = s_w > 0$ *and is* $BEC(p, r)$ *with* $r \leq \min(1, \mathrm{s_w})$. *Moreover,* $\mathcal{W}$ *has an envelope* $\overline{b}$, *such that* $\overline{b}(\cdot) < C < \infty$, *and the norm in* $\mathcal{W}$, $\|\cdot\|_{\mathcal{W}}$ *say, dominates the* L_2*-norm, i.e., there exists a* $C > 0$ *such that* $\|b\|_2 \leq C \|b\|_{\mathcal{W}}$, *for all* $b \in L_2(\overline{\mathbb{R}}^d, F)$.

A5(iii): $\mathcal{B}_{ind} = \{1_x(\cdot) : x \in \overline{\mathbb{R}}^d\}$ *satisfies A2(c) and* F *is absolutely continuous with respect to Lebesgue measure with density* $f(x) < \infty$ *for all* $x \in \overline{\mathbb{R}}^d$.

Assumption A6: A6(i): *The trimming constant x_0 is such that*

$$\inf_{x \in A_{x_0}} \left| C_{z(x)} \right| > \varepsilon > 0,$$

for some $\varepsilon > 0$ and where $A_{x_0} = \{x \in \overline{\mathbb{R}}^d : \beta_1' x \leq x_0\}$.

A6(ii): *The nonparametric estimators for the conditional distributions satisfy*

$$\sup_{x \in \mathbb{R}^d} \left| \widehat{F}_l(x_l \mid x_1, \ldots, x_{l-1}) - F_l(x_l \mid x_1, \ldots, x_{l-1}) \right| = o_P(1), l = 2, .., d,$$

A5(i) is standard in model checks under conditional heteroskedasticity, see Stute, Thies and Zhu (1998). Condition A5(ii) is necessary to obtain a uniform representation and tightness of the process $R_n^1(b)$ in $b \in \mathcal{B} = \{h 1_x : h \in \mathcal{W} \text{ and } x \in \overline{\mathbb{R}}^d\}$. A5(ii) can be relaxed using results for degenerate U-processes, but it simplifies the theory and it gives us a clue about what are the properties necessary in $\mathcal{W}$ to obtain the asymptotic tightness of $R_n^1(b)$ in $b \in \mathcal{B}$. If we assume that $\sigma^{-2}(\cdot)$ is smooth, usual examples of $\mathcal{W}$ are spaces of smooth functions such as Sobolev, Hölder, or Besov classes. Therefore, the covering number condition of Assumptions A2 or A5(ii) can be found in many books and articles on approximation theory. To give an example, define for any vector $(a_1, \ldots, a_d)$ of d integers the differential operator $D^a = \partial^{|a|} / \partial x_1^{a_1} \ldots \partial x_q^{a_d}$, where $|a| = \sum_{i=1}^d a_i$. Let R be a bounded, convex subset of $\mathbb{R}^d$, with nonempty interior. For any smooth function $h : R \subset \mathbb{R}^d \to \mathbb{R}$ and some $\eta > 0$, let η be the largest integer smaller than η, and

$$\|h\|_{\infty,\eta} = \max_{|a| \leq \underline{\eta}} \sup_x \left| D^a h(x) \right| + \max_{|a| = \underline{\eta}} \sup_{x_1 \neq x_2} \frac{|D^a h(x_1) - D^a h(x_2)|}{\|x_1 - x_2\|^{\eta - \underline{\eta}}}.$$

Further, let $C_c^\eta(R)$ be the set of all continuous functions $h : R \subset \mathbb{R}^d \to \mathbb{R}$ with $\|h\|_{\infty,\eta} \leq c$. If $\mathcal{W} = C_c^\eta(R)$, then $\mathcal{W}$ satisfies Assumption A5(ii) provided that $\eta > d$, see van der Vaart and Wellner (1996, Theorem 2.7.1). A5(i) implies the invertibility of the matrix $C_{z(x)}$, and it is assumed only for simplicity in the exposition, see Nikabadze (1997). Conditions for A6(ii) to hold are in abundance in the literature, see, for instance, Andrews (1995). A6(ii) implies that

$$\sup_{x \in \mathbb{R}^d} \left| \widehat{T}_R(x) - T_R(x) \right| = o_P(1)$$

holds.

Theorem 2: *Under the null hypothesis H_0, and Assumptions A1 to A6*

$$\widehat{J}_n \Longrightarrow J_\infty, in \ell^\infty(B_{x_0}),$$

where J_∞ is a standard Brownian Sheet, i.e, a continuous Gaussian process with zero mean and covariance function given by $(u_{11} \wedge u_{21}) \times \cdots \times (u_{1d} \wedge u_{2d})$, for $u_1 = (u_{11}, \ldots u_{1d})'$ and $u_2 = (u_{21}, \ldots u_{2d})'$ in $[0,1]^d$.

Next, using the last theorem and the Continuous Mapping Theorem (CMT), see, e.g., Theorem 1.3.6 in van der Vaart and Wellner (1996), we obtain the asymptotic null distribution of continuous functionals such as CvM_n.

Corollary 1: *Under the assumptions of Theorem 2, for any continuous (with respect to the sup norm) functional* $\Gamma(\cdot)$

$$\Gamma(\widehat{J}_n) \xrightarrow{d} \Gamma(J_\infty).$$

The integrating measure in CvM_n is a random measure, therefore, Corollary 1 is not readily applicable to the present case. However, an application of Lemma 3.1 in Chang (1990) shows that the estimation $F_{n,U}$ of the cdf of U_0, F_U say, does not affect the asymptotic theory for CvM_n as long as

$$\sup_{u \in B_{x_0}} \left| F_{n,U}(u) - F_U(u) \right| \longrightarrow 0 \text{ a.s.}$$

By the Glivenko-Cantelli's Theorem for ergodic and stationary time series, see e.g. Dehling and Philipp (2002, p. 4), jointly with A6(ii), the previous uniform convergence holds.

The power properties of CvM_n can be studied similarly to those established in Escanciano (2009). We do not discuss this issue here for the sake of space. A more important and difficult problem is the asymptotic power comparison between transformed and non-transformed tests from a theoretical point of view. This problem will be investigated elsewhere. Here, we focus on the finite-sample comparison between our ADF test and the bootstrap based tests via a Monte Carlo experiment in the next section.

7.4 Simulation Results

In this section we compare some bootstrap integrated CvM tests with our new ADF test via a Monte Carlo experiment. For the bootstrap CvM tests we consider the weighting functions $w(I_{t-1}, x) = \exp(ix' I_{t-1})$, $w(I_{t-1}, x) = 1(I_{t-1} \leq x)$ and $w(I_{t-1}, x) = 1(\beta' I_{t-1} \leq u)$, $x = (\beta', u)' \in \Pi_{pro} = \mathbb{S}^d \times [-\infty, \infty]$. Our Monte Carlo experiment complements that of Koul and Sakhanenko (2005) in the context of goodness of fit for error distributions.

We briefly describe our simulation setup. Let $I_{t-1} = (Y_{t-1}, Y_{t-2})$ be the information set at time $t-1$. For our ADF test we consider $A_z = \{y \in \overline{\mathbb{R}}^2 : \beta_1' y \leq z\}$, with $\beta_1 = (1,1)'$. Let $F_{n,\beta}(u)$ be the empirical distribution function of the projected information set $\{\beta' I_{t-1} : 1 \leq t \leq n\}$. Escanciano (2006a) proposed the CvM test

$$CVM_{n,pro} = \int_{\Pi_{pro}} (R^1_{n,pro}(\beta, u))^2 F_{n,\beta}(du) d\beta,$$

where

$$R^1_{n,pro}(\beta, u) = \frac{1}{\widehat{\sigma}_e \sqrt{n}} \sum_{t=1}^{n} e_t(\theta_n) 1(\beta' I_{t-1} \leq u)$$

and

$$\widehat{\sigma}_e^2 = \frac{1}{n} \sum_{t=1}^{n} e_t^2(\theta_n).$$

For a simple algorithm to compute $CVM_{n,pro}$ see Appendix B in Escanciano (2006a).

Bierens (1982) proposed to use $w(I_{t-1}, x) = \exp(i I'_{t-1} x)$ as the weighting function in (7.2) and considered the CvM test statistic

$$CvM_{n,\exp} = \int_{\Pi} \left| R^1_{n,\exp}(x) \right|^2 \Psi(dx),$$

where

$$R^1_{n,\exp}(x) = \frac{1}{\widehat{\sigma}_e \sqrt{n}} \sum_{t=1}^{n} e_t(\theta_n) \exp(i x' I_{t-1}),$$

and with $\Psi(dx)$ a suitable chosen integrating function. In order that $CvM_{n,\exp}$ has a closed expression, we consider the weighting function $\Psi(dx) = \phi(x)$, where $\phi(x)$ is the probability density function of the standard normal bivariate r.v. In that case, $CvM_{n,\exp}$ simplifies to

$$CvM_{n,\exp} = \frac{1}{\widehat{\sigma}_e^2 n} \sum_{t=1}^{n} \sum_{s=1}^{n} e_t(\theta_n) e_s(\theta_n) \exp\left(-\frac{1}{2} |I_{t-1} - I_{s-1}|^2 \right).$$

Escanciano (2007) considered the CvM test based on the indicator function, which is given by

$$CvM_{n,ind} = \frac{1}{\widehat{\sigma}_e^2 n^2} \sum_{j=1}^{n} \left[\sum_{t=1}^{n} e_t(\theta_n) 1(I_{t-1} \leq I_{j-1}) \right]^2.$$

We consider the wild bootstrap approximation for all these test statistics as described in Sect. 3 of Escanciano (2007).

Our null model is an AR(2) model:

$$Y_t = a + b Y_{t-1} + c Y_{t-2} + \varepsilon_t.$$

We examine the adequacy of this model under the following DGP:

1. AR(2) model: $Y_t = 0.6 Y_{t-1} - 0.5 Y_{t-2} + \varepsilon_t$.
2. AR(2) model with heteroskedasticity (ARHET): $Y_t = 0.6 Y_{t-1} - 0.5 Y_{t-2} + h_t \varepsilon_t$, where $h_t^2 = 0.1 + 0.1 Y_{t-1}^2 + 0.3 Y_{t-1}^2$.
3. Bilinear model (BIL): $Y_t = 0.6 Y_{t-1} + 0.7 \varepsilon_{t-1} Y_{t-2} + \varepsilon_t$.
4. Nonlinear Moving Average model (NLMA): $Y_t = 0.6 Y_{t-1} + 0.7 \varepsilon_{t-1} \varepsilon_{t-2} + \varepsilon_t$.
5. TAR(2) model: $Y_t = \begin{cases} 0.6 Y_{t-1} + \varepsilon_t, & if\, Y_{t-2} < 1, \\ -0.5 Y_{t-1} + \varepsilon_t, & if\, Y_{t-2} \geq 1. \end{cases}$

We consider for the experiments the sample sizes $n = 50, 100$, and 300. The number of Monte Carlo experiments is 1000 and the number of bootstrap replications is $B = 500$. In all the replications 200 pre-sample data values of the processes were

Table 7.1 *Empirical critical values for* CvM_n

$n\backslash\alpha$	10 %	5 %	1 %
50	0.55557	0.74353	1.18788
100	0.56371	0.75706	1.21756
300	0.61113	0.81060	1.35720

generated and discarded. For a fair comparison, the critical values for the new tests are approximated using 10000 replications of model 1. These critical values are given in Table 7.1.

In Table 7.2 we show the empirical rejection probabilities (RP) associated with the nominal levels 10, 5 and 1 %. The empirical levels of the test statistics are closed to the nominal level. Only in the heteroskedastic case the tests presents some small size distortion (underrejection).

In Table 7.3 we report the empirical power against the BIL, NLMA and TAR(2) alternatives. The RP increase with the sample size n for all test statistics, as expected.

Table 7.2 Empirical size of tests

		AR(2)			ARHET		
		10 %	5 %	1 %	10 %	5 %	1 %
	CvM_n	9.4	4.8	0.8	14.1	7.4	1.7
$n = 50$	$CvM_{n,\exp}$	10.5	5.5	1.1	13.6	7.8	0.8
	$CvM_{n,ind}$	10.3	4.3	1.3	12.4	6.5	1.0
	$CvM_{n,pro}$	11.6	5.7	0.8	13.1	5.9	1.0
	CvM_n	9.0	4.3	1.2	12.4	7.1	2.1
$n = 100$	$CvM_{n,\exp}$	13.4	7.0	1.0	11.7	6.9	2.7
	$CvM_{n,ind}$	11.3	6.5	1.4	12.7	5.8	1.4
	$CvM_{n,pro}$	11.2	6.4	1.6	13.4	7.1	2.0
	CvM_n	10.5	4.8	0.6	11.9	6.4	1.2
$n = 300$	$CvM_{n,\exp}$	10.3	6.0	1.9	12.3	6.1	1.5
	$CvM_{n,ind}$	9.6	4.7	0.5	11.8	6.2	2.0
	$CvM_{n,pro}$	12.5	5.7	1.8	13.2	7.1	1.6

Table 7.3 Empirical power of tests.

	BIL				NLMA			TAR(2)		
		10 %	*5 %*	*1 %*	*10 %*	*5 %*	*1 %*	*10 %*	*5 %*	*1 %*
	CvM_n	29.8	21.7	7.2	19.8	13.4	*4.7*	*53.3*	*40.8*	*19.6*
$n = 50$	$CvM_{n,\exp}$	29.4	18.0	4.4	16.0	8.6	1.5	23.0	13.4	2.0
	$CvM_{n,ind}$	32.2	22.8	8.1	*24.6*	*15.3*	4.6	39.8	30.0	10.5
	$CvM_{n,pro}$	*39.6*	*25.2*	*9.0*	22.9	11.6	2.3	38.5	27.2	9.7
	CvM_n	*56.1*	*43.0*	*24.6*	36.7	27.0	12.9	*76.3*	*69.1*	*49.7*
$n = 100$	$CvM_{n,\exp}$	43.8	30.0	10.7	28.6	16.2	3.8	43.2	27.5	8.2
	$CvM_{n,ind}$	50.0	39.4	19.1	*45.1*	*33.5*	*13.3*	65.4	54.8	34.9
	$CvM_{n,pro}$	55.7	42.3	20.1	41.0	26.8	9.0	62.0	51.3	28.2
	CvM_n	*96.6*	*93.1*	*81.5*	76.3	64.3	41.6	*99.5*	*99.0*	*95.9*
$n = 300$	$CvM_{n,\exp}$	77.2	66.0	36.9	75.6	61.0	28.4	92.5	86.4	61.1
	$CvM_{n,ind}$	76.2	68.4	50.8	88.8	*82.7*	*59.2*	98.5	96.9	88.1
	$CvM_{n,pro}$	75.2	65.8	44.8	*89.4*	80.8	51.9	98.7	96.6	86.5

The highest RP are presented in italics. It is shown that no test is better than the others uniformly for all alternatives, levels and sample sizes. The new ADF Cramér-von Mises test CvM_n performs quite well, being the best in many cases. In particular, it has the highest empirical power for BIL and TAR(2) alternatives uniformly in the level for $n = 300$. The empirical power for $CvM_{n,\exp}$ is low for these alternatives and, in general, less than $CvM_{n,ind}$. The test statistic $CvM_{n,ind}$ has good power against the BIL alternative for $n = 50$ and for the NLMA alternative for $n = 100$, and moderate power against the TAR(2). $CvM_{n,pro}$ performs similarly to $CvM_{n,ind}$, but with a little less empirical power in general.

Summarizing, we conclude from this limited Monte Carlo experiment that our new CvM test compares very well to bootstrap-based integrated tests, with power against all alternatives considered, and in many cases presenting the highest power performance. To conclude, we summarize the properties of our CvM test as follows: (i) it is asymptotically distribution-free; (ii) it is valid under fairly general regularity conditions on the underlying DGP, in particular, under conditional heteroskedasticity of unknown form and multivariate regressors; and (iii) it is simple to compute and has an excellent finite sample performance as has been shown in the Monte Carlo experiment. All these properties make of our test a valuable tool for time series modelling.

Acknowledgements Hira Koul aroused our interest in spectification testing based on empirical processes many years ago. We are most thankful for his support to our research and the many discussions we had during these years. We are pleased to contribute to this volume with warm wishes for many more birthdays. Research funded by Spanish "Plan Nacional de I+D+i" reference number ECO2012–33053.

Appendix: Proofs

First, we shall state a weak convergence theorem which is a trivial extension of Theorem A1 in Delgado and Escanciano (2007). Let for each $n \geq 1$, $I'_{n,0}, \ldots, I'_{n,n-1}$, be an array of random vectors in $\mathbb{R}^p$, $p \in \mathbb{N}$, and $\varepsilon_{n,1}, \ldots, \varepsilon_{n,n}$, be an array of real random variables (r.v.'s). Denote by $(\Omega_n, \mathcal{A}_n, P_n)$, $n \geq 1$, the probability space in which all the r.v.'s $\{\varepsilon_{n,t}, I'_{n,t}\}_{t=1}^n$ are defined. Let $\mathcal{F}_{n,t}$, $0 \leq t \leq n$, be a double array of sub σ-fields of $\mathcal{A}_n$ such that $\mathcal{F}_{n,t} \subset \mathcal{F}_{n,t+1}$, $t = 0, \ldots, n-1$ and such that for each $n \geq 1$ and each $\gamma \in \mathcal{H}$,

$$E[w(\varepsilon_{n,t}, I_{n,t-1}, \gamma) \mid \mathcal{F}_{n,t-1}] = 0 \qquad a.s., 1 \leq t \leq n, \forall n \geq 1. \tag{7.7}$$

Moreover, we shall assume that $\{w(\varepsilon_{n,t}, I_{n,t-1}, \gamma), \mathcal{F}_{n,t}, 0 \leq t \leq n\}$ is a square-integrable martingale difference sequence for each $\gamma \in \mathcal{H}$, that is, (7.7) holds, $Ew^2(\varepsilon_{n,t}, I_{n,t-1}, \gamma) < \infty$ and $w(\varepsilon_{n,t}, I_{n,t-1}, \gamma)$ is $\mathcal{F}_{n,t}$-measurable for each $\gamma \in \mathcal{H}$ and $\forall t, 1 \leq t \leq n, \forall n \in \mathbb{N}$. The following result gives sufficient conditions for the

weak convergence of the empirical process

$$\alpha_{n,w}(\gamma) = n^{-1/2}\sum_{t=1}^{n} w(\varepsilon_{n,t}, I_{n,t-1}, \gamma) \qquad \gamma \in \mathcal{H}.$$

Under mild conditions the empirical process $\alpha_{n,w}$ can be viewed as a mapping from Ω_n to $\ell^\infty(\mathcal{H})$, the space of all real-valued functions that are uniformly bounded on $\mathcal{H}$. The weak convergence theorem that we present here is funded on results by Levental (1989), Bae and Levental (1995) and Nishiyama (2000). In Theorem A1 in Delgado and Escanciano (2007) $\mathcal{H}$ was finite-dimensional, but here we allow for an infinite-dimensional $\mathcal{H}$. The proof of theorem does not change by this possibility, however.

An important role in the weak convergence theorem is played by the conditional quadratic variation of the empirical process $\alpha_{n,w}$ on a finite partition $\mathcal{B} = \{H_k; 1 \leq k \leq N\}$ of $\mathcal{H}$, which is defined as

$$\alpha_{n,w}(\mathcal{B}) = \max_{1\leq k\leq N} n^{-1}\sum_{t=1}^{n} E\left[\sup_{\gamma_1,\gamma_2\in H_k} \left|w(\varepsilon_{n,t}, I_{n,t-1}, \gamma_1) - w(\varepsilon_{n,t}, I_{n,t-1}, \gamma_2)\right|^2 \mid \mathcal{F}_{n,t-1}\right].$$

Then, for the weak convergence theorem we need the following assumptions.

W1: For each $n \geq 1$, $\{(\varepsilon_{n,t}, I_{n,t-1})' : 1 \leq t \leq n\}$ is a strictly stationary and ergodic process. The sequence $\{w(\varepsilon_{n,t}, I_{n,t-1}, \gamma), \mathcal{F}_{n,t}, 0 \leq t \leq n\}$ is a square-integrable martingale difference sequence for each $\gamma \in \mathcal{H}$. Also, there exists a function $C_w(\gamma_1, \gamma_2)$ on $\mathcal{H} \times \mathcal{H}$ to $\mathbb{R}$ such that uniformly in $(\gamma_1, \gamma_2) \in \mathcal{H} \times \mathcal{H}$

$$n^{-1}\sum_{t=1}^{n} w(\varepsilon_{n,t}, I_{n,t-1}, \gamma_1)w(\varepsilon_{n,t}, I_{n,t-1}, \gamma_2) = C_w(\gamma_1, \gamma_2) + o_{P_n}(1).$$

W2: The family $w(\varepsilon_{n,t}, I_{n,t-1}, \gamma)$ is such that $\alpha_{n,w}$ is a mapping from Ω_n to $\ell^\infty(\mathcal{H})$ and for every $\varepsilon > 0$ there exists a finite partition $\mathcal{B}_\varepsilon = \{H_k; 1 \leq k \leq N_\varepsilon\}$ of $\mathcal{H}$, with N_ε being the elements of such partition, such that

$$\int_0^\infty \sqrt{\log(N_\varepsilon)}d\varepsilon < \infty \tag{7.8}$$

and

$$\sup_{\varepsilon\in(0,1)\cap\mathbb{Q}} \frac{\alpha_{n,w}(\mathcal{B}_\varepsilon)}{\varepsilon^2} = O_{P_n}(1). \tag{7.9}$$

Let $\alpha_{\infty,w}(\cdot)$ be a Gaussian process with zero mean and covariance function given by $C_w(\gamma_1, \gamma_2)$. We are now in position to state the following

Theorem A1: *If Assumptions W1 and W2 hold, then it follows that*

$$\alpha_{n,w} \Longrightarrow \alpha_{\infty,w} \text{ in } \ell^\infty(\mathcal{H}).$$

Proof of Theorem A1: Theorem A1 in Delgado and Escanciano (2007).

Proof of Lemma 1: By A2(a-b) we can form for any $\varepsilon > 0$ a finite partition $\mathcal{B}_\varepsilon = \{B_k; 1 \leq k \leq N_{[]}(\varepsilon, \mathcal{B}, \|\cdot\|_p)\}$ of $\mathcal{B}$ in ε-brackets $B_k = [\underline{b}_k, \overline{b}_k]$. Denote $v = 1/s$, with s as in A2(a), and define for every $q \in \mathbb{N}$, $q \geq 1$, $\varepsilon = 2^{-qv}$. We denote the previous partition associated to $\varepsilon = 2^{-qv}$ by $\mathcal{B}_q = \{B_{qk}; 1 \leq k \leq N_q \equiv N_{[]}(2^{-qv}, \mathcal{B}, \|\cdot\|_p)\}$. Without loss of generality we can assume that the finite partitions in the sequence $\{\mathcal{B}_q\}$ are nested. By A2(b), we have

$$\sum_{q=1}^{\infty} 2^{-q} \sqrt{\log N_q} < \infty.$$

Furthermore, by definition of the brackets

$$\begin{aligned} R_n(\mathcal{B}_q) &= \max_{1 \leq k \leq N_q} \left| n^{-1} \sum_{t=1}^{n} E\left[\varepsilon_t^2 \mid \mathcal{F}_{t-1}\right] \sup_{r_1, r_2 \in B_{qk}} |r_1(I_{t-1}) - r_2(I_{t-1})|^2 \right| \\ &= \max_{1 \leq k \leq N_q} \left| n^{-1} \sum_{t=1}^{n} E\left[\varepsilon_t^2 \mid \mathcal{F}_{t-1}\right] \left|\underline{b}_k(I_{t-1}) - \overline{b}_k(I_{t-1})\right|^2 \right| \\ &= \max_{1 \leq k \leq N_q} d_n^2(\underline{b}_k, \overline{b}_k). \end{aligned} \tag{7.10}$$

Define the event

$$V_n = \left\{ \sup_{q \in \mathbb{N}} \max_{1 \leq k \leq N_q} \frac{d_n^2(\underline{b}_k, \overline{b}_k)}{2^{-2q}} \geq \gamma \right\}.$$

We shall show that for all $\eta > 0$, there exists some $\gamma > 0$ such that $\limsup_{n \to \infty} P_n(V_n) \leq \eta$. Note that

$$P_n(V_n) \leq \sum_{q=1}^{\infty} P_n \left(\max_{1 \leq k \leq N_q} \frac{d_n^2(\underline{b}_k, \overline{b}_k)}{2^{-2q}} \geq \gamma \right) \equiv \sum_{q=1}^{\infty} V_{nq} \tag{7.11}$$

Now, define the process

$$\widetilde{\alpha}_{n,w}(r) = n^{-1} \sum_{t=1}^{n} E\left[\varepsilon_t^2 \mid \mathcal{F}_{t-1}\right] r(I_{t-1}),$$

and the quantities for $1 \leq t \leq n$, $\widetilde{\beta}_t(r) = E\left[\varepsilon_t^2 \mid \mathcal{F}_{t-1}\right] r(I_{t-1}) - G_t^{\mathcal{B}}(r)$. Hence,

$$\widetilde{\alpha}_{n,w}(r) = n^{-1} \sum_{t=1}^{n} \widetilde{\beta}_t(r) + n^{-1} \sum_{t=1}^{n} G_t^{\mathcal{B}}(r).$$

By triangle's inequality

$$\begin{aligned} V_{nq} &\leq P_n\left(\max_{1\leq k\leq N_q}\left|n^{-1}\sum_{t=1}^{n}\left|\widetilde{\beta}_t(\underline{b}_k)-\widetilde{\beta}_t(\overline{b}_k)\right|\right|\geq 2^{-2q}\gamma\right) \\ &\quad + P_n\left(\max_{1\leq k\leq N_q}\left|n^{-1}\sum_{t=1}^{n}\left|G_t^{\mathcal{B}}(\underline{b}_k)-G_t^{\mathcal{B}}(\overline{b}_k)\right|\right|\geq 2^{-2q}\gamma\right) \\ &\equiv A_{1nq}+A_{2nq}. \end{aligned}$$

Notice that $\{\widetilde{\beta}_{n,w}(r),\mathcal{F}_{n,t-2}\}$ is a martingale difference sequence for each $r\in\mathcal{B}$, by construction. By a truncation argument, it can be assumed without loss of generality that $\max_{1\leq k\leq N_q}|\varepsilon_t|\left|\underline{b}_k(I_{t-1})-\overline{b}_k(I_{t-1})\right|^2\leq\sqrt{n}a_{q-1}$, where henceforth $a_q=2^{-q\rho}/\sqrt{\log(N_{q+1})}$ with $1<\rho<2$. See Theorem A1 in Delgado and Escanciano (2006). Define the set

$$B_n=\left\{\left(n^{-1}\sum_{t=1}^{n}M_t\right)\leq K\right\}.$$

Now, by Freedman's (1975) inequality in Lemma A2 and Lemma 2.2.10 in van der Vaart and Wellner (1996),

$$\begin{aligned} &E\max_{1\leq k\leq N_q}\left|n^{-1}\sum_{t=1}^{n}\left|\widetilde{\beta}_t(\underline{b}_k)-\widetilde{\beta}_t(\overline{b}_k)\right|\right|1(B_n) \\ &\quad\leq C\left(a_{q-1}^2\log(1+N_q)+a_{q-1}2^{-qv/2}\sqrt{\log(1+N_q)}\right). \end{aligned}$$

Hence, by Markov's inequality and the definition of a_q, on the set B_n,

$$\begin{aligned} A_{1nq} &\leq C\frac{a_{q-1}^2\log(1+N_q)+a_{q-1}2^{-qv/2}\sqrt{\log(1+N_q)}}{2^{-2q}\gamma} \\ &= C\gamma^{-1}2^{-2q(\rho-1)}+C\gamma^{-1}2^{-q(\rho+\frac{v}{2}-1)}. \end{aligned}$$

On the other hand, by (D) and by Markov's inequality

$$\begin{aligned} A_{2nq} &\leq \gamma^{-1}s_n^{-2}\sum_{t=1}^{n}E\max_{1\leq k\leq N_q}2^{2q}\left|n^{-1}\sum_{t=1}^{n}\left|G_t^{\mathcal{B}}(\underline{b}_k)-G_t^{\mathcal{B}}(\overline{b}_k)\right|\right| \\ &\leq \gamma^{-1}2^{-q(v-2)}\left(n^{-1}\sum_{t=1}^{n}M_t\right)\leq K\gamma^{-1}2^{-q(v-2)}, \end{aligned}$$

on the set B_n. Therefore, by our previous arguments and the last three bounds,

$$P_n(V_n)\leq C\gamma^{-1}\sum_{q=1}^{\infty}\left(2^{-2q(\rho-1)}+2^{-q(\rho+\frac{v}{2}-1)}+2^{-q(v-2)}\right)+P_n(B_n^c),$$

which can be made arbitrarily small by choosing a sufficiently large γ and K. Hence, $\mathcal{B}$ has bounded quadratic variation. □

Lemma A0: (*Uniform Law of Large Numbers*) *If the class* $\mathcal{B}$ *is such that* $\log\left(N_{[]}(\varepsilon, \mathcal{B}, \|\cdot\|_1\right) < \infty$ *for each* $\varepsilon > 0$, *with envelope* $\overline{b}$, $g(I_{t-1}, \theta)$ *satisfies A3 and* $E\left|M(I_{t-1})\overline{b}(I_{t-1})\right| < \infty$, *then uniformly in* $(\theta, b) \in \Theta \times \mathcal{B}$,

$$\left|\frac{1}{n}\sum_{t=1}^{n} g(I_{t-1}, \theta)b(I_{t-1}) - E\left[g(I_{t-1}, \theta)b(I_{t-1})\right]\right| = o_P(1).$$

Proof of Lemma A0: Under the assumptions of the lemma, the class $\{g(I_{t-1}, \theta)b(I_{t-1}) : \theta \in \Theta, b \in \mathcal{B}\}$ is Glivenko-Cantelli. □

Proof of Theorem 1: First we shall show that the process

$$S_n(b, \theta) = \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left\{e_t(\theta) - E\left[e_t(\theta) \mid \mathcal{F}_{t-1}\right]\right\} b(I_{t-1}) \tag{7.12}$$

is asymptotically tight with respect to $(b, \theta) \in \mathcal{A}$.

Let us define the class $\mathcal{K} = \{\left\{e_t(\theta) - E\left[e_t(\theta) \mid \mathcal{F}_{t-1}\right]\right\} b(I_{t-1}) : (b, \theta) \in \mathcal{A}\}$. Denote $X_{t-1} = (I_{t-1}, I_{t-2}, \ldots)'$. Let $\mathcal{B}_\varepsilon = \{B_k; 1 \le k \le N_\varepsilon \equiv N_{[]}(\varepsilon, \mathcal{K}, \|\cdot\|_p\}$, with $B_k = [\underline{w}_k(Y_t, X_{t-1}), \overline{w}_k(Y_t, X_{t-1})]$, be a partition of $\mathcal{K}$ in ε-brackets with respect to $\|\cdot\|_p$. Notice that A2 implies

$$\left\| \sup_{\substack{((b_2,\theta_2)\in\mathcal{A}:|\theta_1-\theta_2|<\delta \\ \|b_1-b_2\|_{\mathcal{B}}<\delta}} \left|\left\{e_t(\theta_1) - E\left[e_t(\theta_1) \mid \mathcal{F}_{t-1}\right]\right\} b_1(I_{t-1}) \right.\right.$$

$$\left.\left. - \left\{e_t(\theta_2) - E\left[e_t(\theta_2) \mid \mathcal{F}_{t-1}\right]\right\} b_2(I_{t-1})\right| \right\|_p$$

$$\le C_1\delta^s.$$

Theorem 3 in Chen et al. (2003) and A2 imply that (7.8) holds for such partition. On the other hand

$$\max_{1\le k\le N_\varepsilon} n^{-1}\sum_{t=1}^{n} E\left[\left|\sup_{w_1,w_2\in B_k} |w_1(Y_t, X_{t-1}) - w_2(Y_t, X_{t-1})|\right|^2 \mid \mathcal{F}_{t-1}\right]$$

$$\le \max_{1\le k\le N_\varepsilon} n^{-1}\sum_{t=1}^{n} E\left[\left|\underline{w}_k(Y_t, X_{t-1}) - \overline{w}_k(Y_t, X_{t-1})\right|^2 \mid \mathcal{F}_{t-1}\right]. \tag{7.13}$$

Therefore, A2(c) yields that (7.9) follows, and condition W2 of Theorem A1 holds. The asymptotically tightness of $S_n(b, \theta)$ is then proved.

Then, the last statement and A4(a)

$$R_n^1(\cdot) \quad = \quad \frac{1}{\sqrt{n}}\sum_{t=1}^{n}\{e_t(\theta_1) - E[e_t(\theta_1) \mid \mathcal{F}_{t-1}]\} b(I_{t-1})$$

$$+\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left\{E[e_t(\theta)\mid\mathcal{F}_{t-1}]|_{\theta=\theta_n}-E[e_t(\theta_1)\mid\mathcal{F}_{t-1}]\right\}b(I_{t-1})$$

$$+\frac{1}{\sqrt{n}}\sum_{t=1}^{n}E[e_t(\theta_1)\mid\mathcal{F}_{t-1}]b(I_{t-1})-E\left[E[e_t(\theta_1)\mid\mathcal{F}_{t-1}]b(I_{t-1})\right]$$

$$+\sqrt{n}E\left[E[e_t(\theta_1)\mid\mathcal{F}_{t-1}]b(I_{t-1})\right]+o_P(1),$$

uniformly in $b\in\mathcal{B}$. Part (i) is proved.

As for (ii), A3 and A4(a) imply by the Mean Value Theorem

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n}\left\{E[e_t(\theta)\mid\mathcal{F}_{t-1}]|_{\theta=\theta_n}-E[e_t(\theta_0)\mid\mathcal{F}_{t-1}]\right\}b(I_{t-1})$$

$$=-n^{1/2}(\theta_n-\theta_0)'\frac{1}{n}\sum_{t=1}^{n}g(I_{t-1},\theta_{ni})b(I_{t-1}),$$

and where θ_{ni} satisfies $|\theta_{ni}-\theta_0|\leq|\theta_n-\theta_0|$. Now, A3, A2(b) and Lemma A0 imply that, uniformly in $b\in\mathcal{B}$,

$$\left|\frac{1}{n}\sum_{t=1}^{n}g(I_{t-1},\theta_{ni})b(I_{t-1})-E\left[g(I_{t-1},\theta_0)b(I_{t-1})\right]\right|=o_P(1).$$

From (i) and the last display, (ii) is proved. □

Before proving Theorem 2 we need several useful Lemmas. Let us define $A_{x_0}=\{x\in\overline{\mathbb{R}}^d:\beta_1'x\leq x_0\}$.

Lemma A1: *Under the assumptions of Theorem 2, uniformly in* $x\in A_{x_0}$,

$$R_n^1(T\sigma_n^{-1}(\cdot)1_x)=R_n(T\sigma^{-1}(\cdot)1_x)+o_P(1).$$

Lemma A2: *Under the assumptions of Theorem 2, uniformly in* $x\in A_{x_0}$,

$$R_n^1(T_n\sigma_n^{-1}(\cdot)1_x)=R_n^1(T\sigma_n^{-1}(\cdot)1_x)+o_P(1).$$

Lemma A3: *Under the assumptions of Theorem 2, uniformly in* $u\in B_{x_0}$

$$R_n^1(T_n(\sigma_n^{-1}(\cdot)1_u\circ\widehat{T}_R(\cdot)))=R_n^1(T_n(\sigma_n^{-1}(\cdot)1_u\circ T_R(\cdot)))+o_P(1).$$

Before proving Lemmas A1 to A3 we shall prove two more Lemmas. We need to define first the classes of functions $\mathcal{S}=\{Th1_x(\cdot):h\in\mathcal{W}\text{ and }x\in A_{x_0}\}$ and $\mathcal{B}=\{h1_x:h\in\mathcal{W}\text{ and }x\in A_{x_0}\}$. Define the semimetric

$$d_{ind}(x_1,x_2)=\left\|\varepsilon_t1_{x_1}(I_{t-1})-\varepsilon_t1_{x_2}(I_{t-1})\right\|_2,$$

and recall that $\mathcal{B}_{ind}=\{1(\cdot\leq x)\equiv1_x(\cdot):x\in\mathbb{R}^d\}$.

Lemma B1: *Assume that $\mathcal{B}_{ind}$ satisfies A2(c). Then, if $\mathcal{W}$ satisfies A5(ii) then $\mathcal{B}$ satisfies A2 with $p = 2$.*

Lemma B2: *Assume A3, A5 and A6(i). Then, if $\mathcal{B}$ satisfies A2 with $p = 2$ then $\mathcal{S}$ satisfies A2 with $p = 2$.*

Proof of Lemma B1: We shall start with A2(a). Assume $0 < \delta < 1$. By the triangle inequality, for each $h_1 \in \mathcal{W}$ and each $x_1 \in \overline{\mathbb{R}}^d$

$$\begin{aligned}
&\left\| \sup_{x_2 \in \overline{\mathbb{R}}^d, h_2 \in \mathcal{W}: \|h_1 - h_2\|_{\mathcal{W}} < \delta, d_{ind}(x_1, x_2) < \delta} \left| \varepsilon_t h_1 1_{x_1}(I_{t-1}) - \varepsilon_t h_2 1_{x_2}(I_{t-1}) \right| \right\|_2 \\
\leq\ & C \left\| \sup_{x_2 \in \overline{\mathbb{R}}^d, h_2 \in \mathcal{W}: \|h_1 - h_2\|_{\mathcal{W}} < \delta, d_{ind}(x_1, x_2) < \delta} |\varepsilon_t h_1(I_{t-1})| \left| 1_{x_1}(I_{t-1}) - 1_{x_2}(I_{t-1}) \right| \right\|_2 \\
& + C \left\| \sup_{x_2 \in \overline{\mathbb{R}}^d, h_2 \in \mathcal{W}: \|h_1 - h_2\|_{\mathcal{W}} < \delta, d_{ind}(x_1, x_2) < \delta} 1_{x_2}(I_{t-1}) |\varepsilon_t h_1(I_{t-1}) - \varepsilon_t h_2(I_{t-1})| \right\|_2 \\
\leq\ & C\delta^1 + C\delta^{s_w} \\
\leq\ & C\delta^s,
\end{aligned}$$

with $s = \min(1, s_w)$, where the second inequality is by A5(ii). A2(b) follows from Theorem 6 in Andrews (1994) and A5(ii), because $\mathcal{B}_{ind}$ is $BEC(p, 1/2)$ for all $p \geq 2$. A2(c) follows from the previous arguments, using A5(ii) and that $\mathcal{B}_{ind}$ and $\mathcal{W}$ satisfy A2(c). □

Proof of Lemma B2: We shall start with A2(a). Assume $0 < \delta < 1$. By the triangle inequality, for each $h_1 \in \mathcal{W}$ and each $x_1 \in \overline{\mathbb{R}}^d$

$$\begin{aligned}
&\left\| \sup_{x_2 \in \overline{\mathbb{R}}^d, h_2 \in \mathcal{W}: \|h_1 - h_2\|_{\mathcal{W}} < \delta, d_{ind}(x_1, x_2) < \delta,} \left| \varepsilon_t T h_1 1_{x_1}(I_{t-1}) - \varepsilon_t T h_2 1_{x_2}(I_{t-1}) \right| \right\|_2 \\
\leq\ & C \left\| \sup_{x_2 \in \overline{\mathbb{R}}^d, h_2 \in \mathcal{W}: \|h_1 - h_2\|_{\mathcal{W}} < \delta, d_{ind}(x_1, x_2) < \delta,} \left| \varepsilon_t h_1 1_{x_1}(I_{t-1}) - \varepsilon_t h_2 1_{x_2}(I_{t-1}) \right| \right\|_2 \\
& C \left\| \sup_{x_2 \in \overline{\mathbb{R}}^d, h_2 \in \mathcal{W}: \|h_1 - h_2\|_{\mathcal{W}} < \delta, d_{ind}(x_1, x_2) < \delta,} \left| \varepsilon_t K h_1 1_{x_1}(I_{t-1}) - \varepsilon_t K h_2 1_{x_2}(I_{t-1}) \right| \right\|_2,
\end{aligned}$$

where K is defined in (7.5). Then, it is only necessary to consider the second term in the last inequality. Now, by the linearity of K and the triangle inequality this term is bounded by

$$\leq C \left\| \sup_{x_2 \in \overline{\mathbb{R}}^d, h_2 \in \mathcal{W}: \|h_1 - h_2\|_{\mathcal{W}} < \delta, d_{ind}(x_1, x_2) < \delta,} \varepsilon_t K\{h_1(\cdot)(1_{x_1}(\cdot) - 1_{x_2}(\cdot))\}(I_{t-1}) \right\|_2$$

$$+C\left\|\sup_{x_2\in\overline{\mathbb{R}}^d,h_2\in\mathcal{W}:\|h_1-h_2\|_{\mathcal{W}}<\delta,d_{ind}(x_1,x_2)<\delta,}\varepsilon_t K1_{x_2}(\cdot)(h_1(\cdot)-h_2(\cdot))(I_{t-1})\right\|_2$$

$$\equiv A_1+A_2.$$

A_1^2 is equal to

$$E\left[\sup\varepsilon_t{}^2\left(\int 1(y\in A_{z(I_{t-1})})h_1(\cdot)(1_{x_1}(\cdot)-1_{x_2}(\cdot))s'(x,\theta_0)C_{z(x)}^{-1}G(dx)s(I_{t-1},\theta_0)\right)^2\right],$$

where the sup is computed over $d_{ind}(x_1,x_2)<\delta$. By Cauchy-Schwartz's inequality (C-S), A3, A5 and A6(i) the integral is bounded by

$$C\left|\int h_1^2(\cdot)(1_{x_1}(\cdot)-1_{x_2}(\cdot))^2G(dx)\right|\leq Cd_{ind}^2(x_1,x_2),$$

and hence $|A_1|\leq C\delta$. The proof for A_2 follows from the same steps that for A_1, and hence, it is omitted.

The proof of A2(b) is straightforward. A2(c) can be proved following the arguments in the proof of A2(a). These proofs are omitted for the sake of space. □

Proof of Lemma A1: By Lemmas B1 and B2, $\mathcal{B}$ and $\mathcal{S}$ satisfies A2 with $p=2$. Hence, by Theorem 1,

$$R_n^1(Tb(\cdot)1_x)=R_n(Tb(\cdot)1_x)+o_P(1),$$

uniformly in $x\in A_{x_0}$ and $b\in\mathcal{W}$. Now, by the convergence of σ_n^{-1},

$$R_n^1(T\sigma_n^{-1}(\cdot)1_x)=R_n^1(T\sigma^{-1}(\cdot)1_x)+o_P(1),$$

uniformly in $x\in A_{x_0}$. □

Proof of Lemma A2: Write $R_n^1((T-T_n)\sigma_n^{-1}(\cdot)1_x)$ as

$$\begin{aligned}
&\int\sigma_n^{-1}(y)1_x(y)R_n^1\left(s'(\cdot,\theta_0)1(\cdot\in A_{z(y)}^c)\right)C_{z(y)}^{-1}g(y,\theta_0)F(dy)\\
&-\int\sigma_n^{-1}(y)1_x(y)R_n^1\left(s_n'(\cdot,\theta_n)1(\cdot\in A_{z(y)}^c)\right)C_{n,z(y)}^{-1}g(y,\theta_n)F_n(dy)\\
=\;&\int\sigma_n^{-1}(y)1_x(y)\beta_n(\cdot,\sigma^{-2}(\cdot),\theta_0)\left[F(dy)-F_n(dy)\right]\\
&-\int\sigma_n^{-1}(y)1_x(y)\left[\beta_n(\cdot,\sigma_n^{-2}(\cdot),\theta_n)-\beta_n(\cdot,\sigma^{-2}(\cdot),\theta_0)\right]F_n(dy)\\
\equiv\;&A_{1n}(x)-A_{2n}(x),
\end{aligned}$$

where

$$\beta_n(y,b,\theta)=R_n^1\left(g'(\cdot,\theta)b(\cdot)1(\cdot\in A_{z(y)}^c)\right)C_{z(y)}^{-1}g(y,\theta).\tag{7.14}$$

Putting

$$\alpha_n(y) = \sigma_n^{-1}(y)1_x(y)\beta_n(\cdot, \sigma^{-2}(\cdot), \theta_0),$$

and using our Theorem 1 it is not difficult to show that the sequence $\{\alpha_n(\cdot)\}$ is asymptotically tight. Hence, by Lemma 3.4 in Stute, Thies and Zhu (1998)

$$\sup_{x \in A_{x_0}} |A_{1n}(x)| = o_P(1).$$

Similarly, it can be proved that $\beta_n(y, b, \theta)$ is uniformly tight in $(y, b, \theta) \in B_{x_0} \times \mathcal{W} \times \Theta$ (see Lemmas B1 and B2) and continuous in θ, but θ_n converges in probability to θ_0, and hence, again by Lemma 3.4 in Stute, Thies and Zhu (1998)

$$\sup_{x \in A_{x_0}} |A_{2n}(x)| = o_P(1).$$

□

Proof of Lemma A3: Define

$$\widehat{\gamma}_u(I_{t-1}) = 1_u \circ \widehat{T}_R(I_{t-1}),$$

$$\widetilde{\gamma}_u(I_{t-1}) = 1_u \circ T_R(I_{t-1})$$

and

$$d_u(\cdot) = \widehat{\gamma}_u(\cdot) - \widetilde{\gamma}_u(\cdot).$$

Then, write $R_n^1(T_n\sigma_n^{-1}(d_u(\cdot)))$ as

$$R_n^1(\sigma_n^{-1}(d_u(\cdot))) - \int d_u(\cdot)\sigma_n^{-1}(y)R_n^1\left(s_n'(\cdot, \theta_n)1(\cdot \in A_{z(y)}^c)\right) C_{n,z(y)}^{-1} g_n(y, \theta_n)F_n(dy)$$

$$\equiv A_{n1} - A_{n2}.$$

$|A_{n1}|$ is bounded by

$$\left|n^{-1/2}\sum_{t=1}^{n} e_t(\theta_0)\sigma_n^{-1}(I_{t-1})d_u(I_{t-1})\right| + \left|n^{-1/2}\sum_{t=1}^{n}\{e_t(\theta_n) - e_t(\theta_0)\}\sigma_n^{-1}(I_{t-1})d_u(I_{t-1})\right|$$

$$= \left|R_n\left(\sigma_n^{-1}d_u(\cdot)\right)\right| + \left|\sqrt{n}(\theta_n - \theta_0)'n^{-1}\sum_{t=1}^{n} g(I_{t-1}, \theta_{ni})\sigma_n^{-1}(I_{t-1})d_u(I_{t-1})\right|$$

$$\equiv |B_{n1}(u)| + |B_{n2}(u)|.$$

Now, the stochastic equicontinuity of $R_n b1_x$ in $b \in \mathcal{W}$ and $1_x \in \mathcal{B}_{ind}$, and A6(ii) yield

$$\sup_{u \in [0,1]^d} |B_{1n}(u)| = o_P(1).$$

On the other hand, by Lemma A0, uniformly in $b \in \mathcal{B}$,

$$\left| \frac{1}{n} \sum_{t=1}^{n} g(I_{t-1}, \theta_{ni}) b(I_{t-1}) - E\left[g(I_{t-1}, \theta_0) b(I_{t-1}) \right] \right| = o_P(1).$$

Therefore, A4(b) and the last display yield

$$\sup_{u \in [0,1]^d} |B_{2n}(u)| = o_P(1).$$

As for A_{n2}, by C-S,

$$\left[\int \left[\widehat{\gamma}_u(y) - \widetilde{\gamma}_u(y) \right]^2 F_n(dy) \right]^{1/2} \left[\int \sigma_n^{-2}(y) \beta_n^2(y, \sigma_n^{-1}, \theta_n) F_n(dy) \right]^{1/2},$$

where β_n is defined in (7.14). Both integrants are asymptotically tight (see the arguments of Lemma A2). Hence, Lemma 3.1 in Chang (1990) yields

$$\int \left[\widehat{\gamma}_u(y) - \widetilde{\gamma}_u(y) \right]^2 F_n(dy) = \int \left[\widehat{\gamma}_u(y) - \widetilde{\gamma}_u(y) \right]^2 F(dy) + o_P(1)$$

and

$$\int \sigma_n^{-2}(y) \beta_n^2(y, \sigma_n^{-1}, \theta_n) F_n(dy) = O_P(1).$$

Now, we shall show that A6(ii) and A6(iii) imply

$$\sup_{u \in B_{x_0}} \left| \int \left[\widehat{\gamma}_u(y) - \widetilde{\gamma}_u(y) \right]^2 F(dy) \right| = o_P(1). \tag{7.15}$$

To that end, from A6(ii) we have that

$$\sup_{x \in \mathbb{R}^d} \left| \widehat{T}_R(x) - T_R(x) \right| = o_P(1),$$

Hence, for a given $\varepsilon > 0$, there exists and n_0 such that for all $n \geq n_0$

$$\sup_{x \in \mathbb{R}^d} \left| \widehat{T}_R(x) - T_R(x) \right| < \varepsilon$$

with probability tending to one. Therefore, on that set

$$\sup_{u \in B_{x_0}} \left| \int \left[\widehat{\gamma}_u(y) - \widetilde{\gamma}_u(y) \right]^2 F(dy) \right| \leq \sup_{u \in B_{x_0}} \left| E\left[1(u - \varepsilon \leq U_{t-1} \leq u + \varepsilon \right] \right| \leq 2\varepsilon.$$

Hence, as ε was arbitrary (7.15) holds, and Lemma A3 is proved. □

References

Andrews DWK (1994) Empirical process method in econometrics. In: Engle RF, McFadden DL (ed) The Handbook of Econometrics, vol. IV. North-Holland, Amsterdam

Andrews DWK (1995) Nonparametric kernel estimation for semiparametric models. Econometric Theor 11:560–596

Angrist JD, Kuersteiner GM (2011) Causal effects of monetary shocks: Semiparametric conditional independence tests with a multinomial propensity score. Rev Econ Stat 93:725–747

Bai J., 2003, Testing parametric conditional distributions of dynamic models. Review of Economics and Statistics 85:531–549

Bai J, Ng S (2001) A consistent test for conditional symmetry in time series models. J Economet 103:225–258

Bai J, Chen Z (2008) Testing multivariate distributions in GARCH models. J Economet 143:19–36

Bae J, Levental S (1995) Uniform CLT for markov chains and its invariance principle: A martingale approach. J Theor Probab 8:549–570

Bierens HJ (1982) Consistent model specification tests. J Economet 20:105–134

Bierens HJ (1984) Model specification testing of time series regressions. J Economet 26:323–353

Bierens HJ (1990) A consistent conditional moment test of functional form. Econometrica 58: 1443–1458

Bierens HJ, Ploberger W (1997) Asymptotic theory of integrated conditional moment tests. Econometrica 65:1129–1151

Chang NM (1990) Weak convergence of a self-consistent estimator of a survival function with doubly censored data. Ann Stat 18:391–404

Chen SX, Härdle W, Li M (2003) An empirical likelihood goodness-of-fit test for time series. J Roy Statist Soc Ser B 65:663–678

de Jong RM (1996) The Bierens' tests under data dependence. J Economet 72:1–32

Dehling H, Philipp W., 2002, Empirical process techniques for dependent data. In: Dehling H, Mikosch T, Sø rensen M (ed) Empirical process techniques for dependent data (Birkhaüser), 3–113

Delgado MA, Escanciano JC (2007) Nonparametric tests for conditional symmetry in dynamic models. J Economet 141:652–682

Delgado MA, Hidalgo J, Velasco C (2008) Distribution free goodness-of-fit tests for linear models. Ann Stat 33:2568–2609

Delgado MA, Stute W (2008) Distribution free specification tests of conditional models. J Economet 143:37–55

Escanciano JC (2006a) A consistent diagnostic test for regression models using projections. Econometr Theor 22:1030–1051

Escanciano JC (2006b) Goodness-of-fit tests for linear and non-linear time series models. J Amer Stat Assoc 101:531–541

Escanciano JC (2007) Model checks using residual marked empirical processes. Stat Sin 17: 115–138

Escanciano JC (2009) On The Lack of Power of Omnibus Specification Tests. Econometr Theor 25:162–194

Escanciano JC, Mayoral S (2010) Data-driven smooth tests for the martingale difference hypothesis. Comput Stat Data An 54:1983–1998

Eubank R, Hart J (1992) Testing goodness-of-fit in regression via order selection criteria. Ann Stat 20:1412–1425

Eubank R, Spiegelman S (1990) Testing the goodness of fit of a linear model via nonparametric regression techniques. J Amer Stat Assoc 85:387–392

Fan J, Huang L (2001) Goodness-of-fit tests for parametric regression models. J Amer Stat Assoc 96:640–652

Fan J, Yao Q, 2003, Nonlinear time series: Nonparametric and parametric methods. Springer-Verlag, New York

Fan Y, Li Q (1996) Consistent model specification tests: Omitted variables, parametric and semiparametric functional forms. Econometrica 64:865–890
Guerre E, Lavergne P (2005) Rate-optimal data-driven specification testing for regression models. Ann Stat 33:840–870
Härdle W, Mammen E (1993) Comparing nonparametric versus parametric regression fits. Ann Stat 21:1926–1974
Hart JD, 1997, Nonparametric smoothing and lack-of-fit tests. Springer Verlag, New-York
Hong Y, White H (1995) Consistent specification testing via nonparametric series regression. Econometrica 63:1133–1159
Horowitz JL, Härdle W (1994) Testing a parametric model against a semiparametric alternative. Economet Theor 10:821–848
Horowitz JL, Spokoiny VG (2001) An adaptive rate-optimal test of a parametric mean-regression model against a nonparametric alternative. Econometrica 69:599–632
Khmaladze EV., 1981, Martingale approach to the goodness of fit tests. Theory Probab Appl 26:246–265
Khmaladze EV (1993) Goodness of fit problem and scanning innovation martingales. Ann Stat 21:798–829
Khmaladze EV, Koul H (2004) Martingale transforms goodness-of-fit tests in regression models. Ann Stat 32:995–1034
Koenker R, Xiao Z (2002) Inference on the quantile regression process. Econometrica 70:1583–1612
Koul HL 1992. Weighted empiricals and linear models. IMS Lecture Notes-Monograph Series, vol. 21. Hayward, California
Koul HL., 2002, Weighted Empirical Processes in Dynamic Nonlinear Models, 2nd ed, Lect Notes Stat, Vol. 166, Springer
Koul HL, Sakhanenko L (2005) Goodness-of-fit testing in regression: A finite sample comparison of bootstrap methodology and Khamaladze transformation. Stat Probab Lett 74:290–302
Koul HL, Stute W (1999) Nonparametric model checks for time series. Ann Stat 27:204–236
Levental S (1989) A uniform CLT for uniformly bounded families of martingale differences. J Theor Probab 2:271–287
Li Q (1999) Consistent model specification test for time series econometric models. J Economet 92:101–147
Li Q, Hsiao C, Zinn J (2003) Consistent specification tests for semiparametric/nonparametric models based on series estimation methods. J Economet 112:295–325
Li Q, Wang S (1998) A simple consistent bootstrap test for a parametric regression functional form. J Economet 87:145–165
Nikabadze A., 1997, Scanning innovations and goodness of fit tests for vector random variables against the general alternative. A. Razmadze Mathematical Institute, Tbilisi, Preprint
Nishiyama Y (2000) Weak convergence of some classes of martingales with jumps. Ann Probab 28:685–712
Rossenblatt M (1952) Remark on multivariate transformation. Ann Math Stat 23:470–472
Song KK (2009) Testing conditional independence via rosenblatt transforms. Ann Stat 37: 4011–4045
Song KK (2010) Testing semiparametric conditional moment restrictions using conditional martingale transforms. J Economet 154:74–84
Stinchcombe M, White H (1998) Consistent specification testing with nuisance parameters present only under the alternative. Economet Theor 14:295–325
Stute W (1997) Nonparametric model checks for regression. Ann Stat 25:613–641
Stute W, Gonzalez-Manteiga W, Presedo-Quindimil M (1998) Bootstrap approximations in model checks for regression. J Amer Stat Assoc 93:141–149
Stute W, Thies S, Zhu LX (1998) Model checks for regression: An innovation process approach. Ann Stat 26:1916–1934
Stute W, Presedo-Quindimil M, González-Manteiga W, Koul HL (2006) Model checks of higher order time series. Stat Probab Lett 76:1385–1396
Stute W, Zhu LX (2002) Model checks for generalized linear models. Scand J Stat 29:535–545

van der Vaart AW, Wellner JA (1996) Weak convergence and empirical processes. Springer, New York
Wefelmeyer W (1996) Quasi-likelihood models and optimal inference. Ann Stat 24:405–422
Zheng X (1996) A consistent test of functional form via nonparametric estimation technique. J Economet 75:263–289

Chapter 8
Ridge Autoregression R-Estimation: Subspace Restriction

A. K. Md. Ehsanes Saleh

8.1 Introduction

Consider the usual AR(p)-model

$$X_t = \rho_1 X_{t-1} + \cdots + \rho_p X_{t-p} + e_t,\ t = 0, \pm 1, \pm 2, \cdots \tag{8.1}$$

where $e_{\pm 1}, \cdots, e_{\pm t}$ are i.i.d.r.v. with a cdf F defined on R^1, Let $Y_0 = (X_0, X_{-1}, \cdots, X_{1-p})'$ be an observable random vector independent of $e_1, e_2, \cdots$. We assume that all the roots of p-degree polynomial (See Brockwell and Davis, 1987)

$$x^p - \rho_1 x^{p-1} - \cdots - \rho_p = 0 \text{ are in}(-1, 1). \tag{8.2}$$

Here, $\boldsymbol{\rho} = (\rho_1, \cdots, \rho_p)' \in R^p$ is vector of unknown autoregressive parameters. Assume further that $\boldsymbol{\rho}$ is suspected to belong to the linear subspace $\mathbf{H}\boldsymbol{\rho} = \mathbf{h}$, where $\mathbf{H}$ is a $q \times p$ matrix of known constants and $\mathbf{h}$, is a q-vector of known constants. If $\mathbf{H} = \begin{pmatrix} \mathbf{I}_{p1} & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{p2} \end{pmatrix}$ and $\mathbf{h} = \begin{pmatrix} \boldsymbol{\rho}_{(1)} \\ \mathbf{0} \end{pmatrix}$, then $\mathbf{H}\boldsymbol{\rho} = \begin{pmatrix} \boldsymbol{\rho}_{(1)} \\ \boldsymbol{\rho}_{(2)} \end{pmatrix} = \begin{pmatrix} \boldsymbol{\rho}_{(1)} \\ \mathbf{0} \end{pmatrix}$ leading to the subhypothesis, that $\boldsymbol{\rho}_{(2)} = \mathbf{0}$. To this end, we first consider the theory of R-estimation of $\boldsymbol{\rho}$ based on a class of rank statistics and define a class of rank test for the null-hypothesis, $H_0 : \mathbf{H}\boldsymbol{\rho} = \mathbf{h}$ Vs $\mathbf{H}\boldsymbol{\rho} \neq \mathbf{h}$. To obtain the asymptotic properties of the R-estimators, we use Koul and Saleh (1995) AUL results for the class of rank statistics. These results are then used to investigate the asymptotic properties of R-estimators of $\boldsymbol{\rho}$ and their properties.

For the AR(p)-model (1.1), let $\tilde{\boldsymbol{\rho}}_n$ be R-estimator of $\boldsymbol{\rho}$ and $\hat{\boldsymbol{\rho}}_n$ be the R-estimator of $\boldsymbol{\rho}$ under $\mathbf{H}\boldsymbol{\rho} = \mathbf{h}$. We designate $\tilde{\boldsymbol{\rho}}_n$ as "unrestricted R-estimator" (URE) of $\boldsymbol{\rho}$ and $\hat{\boldsymbol{\rho}}_n$ as the "restricted R-estimator" (RRE) of $\boldsymbol{\rho}$ respectively. The RRE performs better than the URE when $\mathbf{H}\boldsymbol{\rho} = \mathbf{h}$ holds. But, if $\boldsymbol{\rho}$ departs from this subspace, RRE may be considerably biased, inefficient and even inconsistent, while URE retains all the performance characteristics for the variations of $\boldsymbol{\rho}$ around the subspace. Further, since $\mathbf{H}\boldsymbol{\rho} = \mathbf{h}$ is suspected to hold, we consider the rank statistics,

A. K. Md. E. Saleh (✉)
School of Mathematics and Statistics, Carleton University, Ottawa, Canada
e-mail: esaleh@math.carleton.ca

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_8,

$\mathcal{L}_n$ to test this restriction by a suitable form. Let c_α be the upper level critical value for the distribution of $\mathcal{L}_n$ under $H_0 : \mathbf{H}\boldsymbol{\rho} = \mathbf{h}$, then we may define the "preliminary test R-estimator" (PTRE), $\hat{\boldsymbol{\rho}}_n^{PT}$ and the Stein-type R-estimator (SRE), $\hat{\boldsymbol{\rho}}_n^{s}$ and the "positive-rule Stein-type R-estimator" (PRSRE), $\hat{\boldsymbol{\rho}}_n^{s+}$ respectively as in Koul and Saleh (1995) specific to the suspected restriction, $\mathbf{H}\boldsymbol{\rho} = \mathbf{h}$. The relative merits of these R-estimators are studied in terms of asymptotic distributional risks (ADR) as in Koul and Saleh (1995) and Saleh and Kibria (2011) and Sen and Saleh (1987). Finally, we modify these five estimators using the "ridge factors" to define "ridge autoregression estimators" (RARE) and study their asymptotic dominance properties. The main results on the asymptotic properties of different ridge autoregression R-estimators (RARRE) are presented in Sect. 8.5 and 8.6 with a concluding remarks in Sect. 8.7.

8.2 R-estimation of $\boldsymbol{\rho}$ for AR(p)-model

Let $\mathbf{Y}_i = (X_i, \cdots, X_{i-p+1})^{'}, 1 \le i \le n$ and define $R_i(b)$ as the rank of $(X_i - b\mathbf{Y}_{i-1})$ among $\{X_j - b\mathbf{Y}_{j-1}, 1 \le j \le n\}$ for $i = 1, \cdots, n$. Set $R_i(b) = 0$ if $i \le 0$. Let φ be a nondecreasing function from $[0, 1]$ to R^1 and define the vector of rank statistics, $\mathbf{L}_n(\mathbf{b}) = (L_{1n}(b), \cdots, L_{pn}(b))^{'}$ where

$$L_{jn}(b) = n^{-\frac{1}{2}} \sum_{i=j+1}^{n} X_{i-j}\, \varphi\left(\frac{R_i(b)}{n+1}\right),\ i \le j \le p,\ \mathbf{b} \in R^p \tag{8.3}$$

It is natural to define an R-estimator of $\boldsymbol{\rho}$ by the relation

$$\inf_{\mathbf{b} \in R^p} \left\| \mathbf{L}_n(\mathbf{b}) \right\| = \mathbf{L}_n(\tilde{\boldsymbol{\rho}}_n). \tag{8.4}$$

An alternative way to define R-estimator of $\boldsymbol{\rho}$ is to follow Jaeckel (1972) to AR($\boldsymbol{\rho}$)-model. Accordingly, set $a_n(i) = \varphi\left(\frac{i}{n+1}\right)$ and $Z_{(i)}(b) = i^{th}$ largest residuals $\{X_k - \mathbf{b}^{'} Y_{k-1},\ 1 \le k \le n\},\ 1 \le i \le n$, and

$$T_n(\mathbf{b}) = \sum_{i=1}^{n} a_n(i)\, Z_{(i)}(\mathbf{b}),\ \mathbf{b} \in R^p. \tag{8.5}$$

According to Jaeckel (1972), if $\sum_{i=1}^{n} a_n(i) = 0$, then $T_n(\mathbf{b})$ can be shown to be convex on R^p with a.e. differential equal to $-\mathbf{L}_n(\mathbf{b})$. Thus, the minimizer $\boldsymbol{\rho}_J$ of $T_n(\mathbf{b})$ exists and has the property that makes $\left\| \mathbf{L}_n(\mathbf{b}) \right\|$ small. It follows from the linearity results given below that $\boldsymbol{\rho}_J$ and $\tilde{\boldsymbol{\rho}}_n$ are asymptotically equivalent.

Theorem 2.1 *(Koul and Saleh (1995)) Assume that (8.1) and (8.2) hold. In addition, assume the following:*

(*a*) (*i*) $E(e_t) = 0$ and $E(e_t^4) < \infty\ \forall\ t$. (ii) F has uniformly continuous density $f, f > 0$ a.e.

(b) φ is non-decreasing function and differentiable with its derivative φ' being uniformly continuous on [0,1].

Then, for every $0 < c < \infty$,

$$\sup_{\|\mathbf{\Delta}\| < c} \left\| L_n(\rho + n^{-\frac{1}{2}}\mathbf{\Delta}) - \mathbf{L}_n^* + \mathbf{\Delta}'\mathbf{\Sigma}\,\gamma \right\| = o_p(1) \tag{8.6}$$

where $\mathbf{L}_n^* = (L_{1n}^*, \cdots, L_{pn}^*)'$ with

$$L_{jn}^* = n^{-\frac{1}{2}} \sum_{i=j+1}^{n} (X_{j+1} - \bar{X}_j)[\varphi\,(F(e_i)) - \bar{\varphi}],\ \bar{\varphi} = \int \varphi(t)dF(t) \tag{8.7}$$

$\bar{X}_j = n^{-1} \sum_{i=j+1}^{n} X_{i-j},\ \gamma = \int f d\varphi(F),$ and $\mathbf{\Sigma}$ is the Toeplitz matrix defined as

$$\mathbf{\Sigma} = ((\,\beta(i-j)\,)),\ i, j = 1, \cdots, p,\ cov(X_0, X_k) = \beta(k), 1 \le k \le p. \tag{8.8}$$

Note that the above Theorem covers Wilcoxon's type score but not normal score.

Further, under (a) and (b) of Theorem 2.1 for every $0 < c < \infty$

$$\sup_{\|\mathbf{\Delta}\| < c} \left\| L_n(\rho + n^{-\frac{1}{2}}\mathbf{\Delta}) - \mathbf{L}_n(\rho) + \mathbf{\Delta}\mathbf{\Sigma}\,\gamma \right\| = o_p(1) \tag{8.9}$$

Arguing as in Koul (1985, Lemma 3.1) or in Jeackel (1972) one may conclude

$$\left\| n^{\frac{1}{2}}(\tilde{\boldsymbol{\rho}}_n - \boldsymbol{\rho}) \right\| = O_p(1).$$

Consequently, by Theorem 2.1

$$n^{\frac{1}{2}}(\tilde{\boldsymbol{\rho}}_n - \boldsymbol{\rho}) = \gamma^{-1}\mathbf{\Sigma}^{-1}\mathbf{L}_n^* + o_p(1). \tag{8.10}$$

Observe that $\mathbf{L}_n^*$ is a vector of square intergrable mean zero martingales with $E[\mathbf{L}_n^*\mathbf{L}_n^{*'}] = \sigma_\varphi^2\mathbf{\Sigma}$, $\sigma_\varphi^2 = var[\varphi(u)]$. Thus, by routine Cramer-Wold device and Corollary 3.1 of Hall and Hyde (1980) one obtains

$$\mathbf{L}_n^* \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, \sigma_\varphi^2\mathbf{\Sigma}). \text{ Hence } n^{\frac{1}{2}}(\tilde{\boldsymbol{\rho}}_n - \boldsymbol{\rho}) \xrightarrow{\mathcal{D}} N_p(\mathbf{0}, \gamma^{-2}\sigma_\varphi^2\mathbf{\Sigma}^{-1}). \tag{8.11}$$

Now, consider the restricted R-estimator, $\hat{\boldsymbol{\rho}}_n$ of $\boldsymbol{\rho}$ under $\mathbf{H}\boldsymbol{\rho} = \mathbf{h}$ as

$$\hat{\boldsymbol{\rho}}_n = \tilde{\boldsymbol{\rho}}_n - \mathbf{\Sigma}_n^{-1}\mathbf{H}'(\mathbf{H}\mathbf{\Sigma}_n^{-1}\mathbf{H}')^{-1}(\mathbf{H}\,\tilde{\boldsymbol{\rho}}_n - \mathbf{h}), \tag{8.12}$$

where, $\mathbf{\Sigma}_n = \left(\sum_i \mathbf{Y}_i\mathbf{Y}_{i-1}'\right)$, where $n^{-1}\left\|\mathbf{\Sigma}_n\right\| \xrightarrow{\mathcal{P}} \mathbf{\Sigma}$. We draw two relations from (8.6) of Theorem 2.1 given by

$$(i) \quad L_n(\hat{\boldsymbol{\rho}}_n) - L_n(\boldsymbol{\rho}) + \gamma\, n^{\frac{1}{2}}(\hat{\boldsymbol{\rho}}_n - \boldsymbol{\rho})\mathbf{\Sigma} = o_p(1) \tag{8.13}$$

$$(ii) \quad L_n(\boldsymbol{\rho}) - \gamma\, n^{\frac{1}{2}}(\tilde{\boldsymbol{\rho}}_n - \boldsymbol{\rho})\mathbf{\Sigma} = o_p(1). \tag{8.14}$$

As a result of (8.13) and (8.14) we have

$$L_n(\hat{\boldsymbol{\rho}}_n) = \gamma\, n^{\frac{1}{2}}(\tilde{\boldsymbol{\rho}}_n - \hat{\boldsymbol{\rho}})\boldsymbol{\Sigma} + o_p(1) \tag{8.15}$$

Thus, for the rank statistics for the test of $\mathbf{H}\boldsymbol{\rho} = \mathbf{h}$, one may use the quadratic form

$$\begin{aligned}
\mathcal{L}_n &= [L_n(\hat{\boldsymbol{\rho}}_n)]^{'}\boldsymbol{\Sigma}_n[L_n(\hat{\boldsymbol{\rho}}_n)] && (8.16)\\
&= n\frac{\gamma^2}{\sigma_\varphi^2}\,(\tilde{\boldsymbol{\rho}}_n - \hat{\boldsymbol{\rho}}_n)^{'}\boldsymbol{\Sigma}(\tilde{\boldsymbol{\rho}}_n - \hat{\boldsymbol{\rho}}_n) + o_p(1) && (8.17)\\
&= n\frac{\gamma^2}{\sigma_\varphi^2}\,(\mathbf{H}\tilde{\boldsymbol{\rho}}_n - \mathbf{h})^{'}(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}^{'})^{-1}(\mathbf{H}\tilde{\boldsymbol{\rho}}_n - \mathbf{h}) + o_p(1) && (8.18)
\end{aligned}$$

Hence, one may show that

$$\lim_{n\to\infty} P(\mathcal{L}_n < x|\mathbf{H}\boldsymbol{\rho} = \mathbf{h}) = \mathcal{H}_q(x;0) \tag{8.19}$$

where $\mathcal{H}_q(x;0)$ is the cdf of the central chi-square distribution with q degrees of freedom. For the application of $\mathcal{L}_n$ one may have to estimate γ consistently using the methods suggested by Koul (2002, p. 128).

8.3 Various R-estimators of $\boldsymbol{\rho}$ and Their Asymptotic Distributional Properties

First, we consider the following quasi-empirical Bayes R-estimators of $\boldsymbol{\rho}$ when one suspects that $\boldsymbol{\rho}$ may belong to the linear subspace $\mathbf{H}\boldsymbol{\rho} = \mathbf{h}$, as follows using Saleh (2006). (i) the unrestricted R-estimator (URE), $\tilde{\boldsymbol{\rho}}_n$ (ii) the restricted R-estimator (RRE), $\hat{\boldsymbol{\rho}}_n$ (iii) the preliminary test R-estimator (PTRE), $\hat{\boldsymbol{\rho}}_n^{PT}$

$$\hat{\boldsymbol{\rho}}_n^{PT} = \tilde{\boldsymbol{\rho}}_n - (\tilde{\boldsymbol{\rho}}_n - \hat{\boldsymbol{\rho}}_n)I(\mathcal{L}_n < \chi_q^2(\alpha)) \tag{8.20}$$

where $\chi_q^2(\alpha)$ is the α-level critical value from the asymptotic null distribution of $\mathcal{L}_n$ and $I(A)$ is the indicator function of the set A. (iv) the Stein-type R-estimator (SRE), $\hat{\boldsymbol{\rho}}_n^s$

$$\hat{\boldsymbol{\rho}}_n^s = \tilde{\boldsymbol{\rho}}_n - (q-2)(\tilde{\boldsymbol{\rho}}_n - \hat{\boldsymbol{\rho}}_n)\mathcal{L}_n^{-1} \tag{8.21}$$

and (v) the positive-rule Stein-type R-estimator (PRSRE), $\hat{\boldsymbol{\rho}}_n^{s+}$

$$\hat{\boldsymbol{\rho}}_n^{s+} = \hat{\boldsymbol{\rho}}_n - I(\mathcal{L}_n < q-2) + \hat{\boldsymbol{\rho}}_n^s I(\mathcal{L}_n \geq q-2). \tag{8.22}$$

Note that PTRE and PRSRE are convex combinations of $\hat{\boldsymbol{\rho}}_n$ and $\tilde{\boldsymbol{\rho}}_n$ and $\hat{\boldsymbol{\rho}}_n$ and $\hat{\boldsymbol{\rho}}_n^s$ respectively, while $\hat{\boldsymbol{\rho}}_n^s$ is not.

Now, we consider the asymptotic distributional bias (ADB), MSE (ADMSE) and risks (ADQR) of the five R-estimators of $\boldsymbol{\rho}$. It may be verified that $\mathcal{L}_n$ is a consistent

test for the test of the null hypothesis, $H_0 : \mathbf{H}\boldsymbol{\rho} = \mathbf{h}$. As a result $\hat{\boldsymbol{\rho}}_n^{PT}$, $\hat{\boldsymbol{\rho}}_n^{s}$ and $\hat{\boldsymbol{\rho}}_n^{s+}$ are asymptotically equivalent to $\tilde{\boldsymbol{\rho}}_n$ and $\mathcal{L}_n \xrightarrow{\mathcal{P}} \infty$ as $n \to \infty$ and the asymptotic distribution of $\hat{\boldsymbol{\rho}}_n$ degenerates. To avoid this asymptotic degeneracy, we consider a sequence of local alternatives

$$K_{(n)} : \mathbf{H}\boldsymbol{\rho} = \mathbf{h} + n^{-\frac{1}{2}}\boldsymbol{\xi},\ \boldsymbol{\xi} \in R^q. \tag{8.23}$$

Note that when $\boldsymbol{\xi} = \mathbf{0}$, $K_{(n)}$ reduces to H_0. Now, we use the technique by Saleh (2006) to obtain the following Theorem. But, first we let $G_p(.; \boldsymbol{\mu}, \boldsymbol{\Sigma})$ and $\mathcal{H}_q(.; \Delta^2)$ respectively stand for the p-dimensional normal distribution with mean, $\boldsymbol{\mu}$ and covariance matrix, $\boldsymbol{\Sigma}$ and a non central chi-square cdf with q degrees of freedom and noncentrality parameter, Δ^2, then

Theorem 3.1 *Under $\{K_{(n)}\}$ and assumed conditions of Theorem 2.1 and a (ii). The error distribution cdf F has an absolutely continuous density f with its a.e. derivative, $f^{'}$ with finite Fisher information*

$$I(f) = \int_{-\infty}^{\infty} \{-\frac{f^{'}(u)}{f(u)}\}^2 f(u)du < \infty \tag{8.24}$$

hold. Then, as $n \to \infty$

$$(a)\quad \begin{pmatrix} \sqrt{n}\,(\tilde{\boldsymbol{\rho}}_n - \boldsymbol{\rho})^{'} \\ \sqrt{n}\,(\hat{\boldsymbol{\rho}}_n - \boldsymbol{\rho})^{'} \\ \sqrt{n}\,(\tilde{\boldsymbol{\rho}}_n - \hat{\boldsymbol{\rho}}_n)^{'} \end{pmatrix} \xrightarrow{\mathcal{D}} N_{3p}\left\{ \begin{pmatrix} \mathbf{0} \\ -\boldsymbol{\delta} \\ \boldsymbol{\delta} \end{pmatrix}; \gamma^{-2}\sigma_{\varphi}^2 \begin{pmatrix} \boldsymbol{\Sigma}^{-1} & \boldsymbol{\Sigma}^{-1} - \mathbf{A} & \mathbf{A} \\ \boldsymbol{\Sigma}^{-1} - \mathbf{A} & \boldsymbol{\Sigma}^{-1} - \mathbf{A} & \mathbf{0} \\ \mathbf{A} & \mathbf{0} & \mathbf{A} \end{pmatrix} \right\} \tag{8.25}$$

where

$$\mathbf{A} = \boldsymbol{\Sigma}^{-1}\mathbf{H}^{'}(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}^{'})^{-1}\mathbf{H}\boldsymbol{\Sigma}^{-1} \text{ and } \boldsymbol{\delta} = \boldsymbol{\Sigma}^{-1}\mathbf{H}^{'}(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}^{'})^{-1}\boldsymbol{\xi}$$

$$(b)\qquad \lim_{n\to\infty} P\{\mathcal{L}_n < x | K_{(n)}\} = \mathcal{H}_q(x; \Delta^2),\ \Delta^2 = \sigma_{\varphi}^{-2}\gamma^2(\boldsymbol{\delta}^{'}\boldsymbol{\Sigma}\boldsymbol{\delta})$$

$$\begin{aligned}(c)\quad &\lim_{n\to\infty} P\{\sqrt{n}\,(\hat{\boldsymbol{\rho}}_n^{PT} - \boldsymbol{\rho})^{'} \le x | K_{(n)}\} \\ &= \mathcal{H}_q(\chi_q^2(\alpha); \Delta^2) G_p[\mathbf{x} + \boldsymbol{\delta}, \mathbf{0}, \sigma_{\varphi}^2\gamma^{-2}(\boldsymbol{\Sigma}^{-1} - \mathbf{A})] \\ &\quad + \int_{E(\Delta)} G_p[\mathbf{x} - \boldsymbol{\Sigma}^{-1}\mathbf{H}^{'}(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}^{'})^{-1}\mathbf{Z}; 0, \sigma_{\varphi}^2\gamma^{-2}(\boldsymbol{\Sigma}^{-1} - \mathbf{A})] dG_p \\ &\quad [\mathbf{Z}, \mathbf{0}, \sigma_{\varphi}^2\gamma^{-2}(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}^{'})]\end{aligned}$$

where $E(\Delta) = \{\mathbf{Z} : \frac{\gamma^2}{\sigma_{\varphi}^2}(\mathbf{Z} + \Delta)^{'}(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}^{'})^{-1}(\mathbf{Z} + \Delta) \ge \chi_q^2(\alpha)\}, \mathbf{Z} \sim N_p(\mathbf{0}, \frac{\gamma^2}{\sigma_{\varphi}^2}\boldsymbol{\Sigma})$

$$(d)\qquad \sqrt{n}\,(\hat{\boldsymbol{\rho}}_n^{s} - \boldsymbol{\rho}) \xrightarrow{\mathcal{D}} \mathbf{Z} - \frac{p(q-2)\boldsymbol{\Sigma}^{-1}\mathbf{H}^{'}(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}^{'})^{-1}(\mathbf{H}\mathbf{Z} + \Delta)}{(\mathbf{Z} + \Delta)^{'}(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}^{'})^{-1}(\mathbf{Z} + \Delta)}$$

$$(e)\ \sqrt{n}\,(\hat{\boldsymbol{\rho}}_n^{s+}-\boldsymbol{\rho}) \xrightarrow{\mathcal{D}} [\mathbf{Z}-\frac{p(q-2)\boldsymbol{\Sigma}^{-1}\mathbf{H}'(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}')^{-1}(\mathbf{H}\mathbf{Z}+\Delta)}{(\mathbf{Z}+\Delta)'(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}')^{-1}(\mathbf{Z}+\Delta)}]$$
$$\times I\left((\mathbf{Z}+\Delta)'(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}')^{-1}(\mathbf{Z}+\Delta)\geq q-2\right)+\boldsymbol{\Sigma}^{-1}\mathbf{H}'(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}')^{-1}(\mathbf{H}\mathbf{Z}+\Delta)$$
$$\times I\left((\mathbf{H}\mathbf{Z}+\Delta)'(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}')^{-1}(\mathbf{H}\mathbf{Z}+\Delta)\leq q-2\right)$$

Assume that for a given estimator $\boldsymbol{\rho}^*$ of $\boldsymbol{\rho}$

$$\lim_{n\to\infty} P\{\sqrt{n}\,(\boldsymbol{\rho}_n^*-\boldsymbol{\rho})'\leq x|K_{(n)}\}=G_p(x;\mathbf{B}^*,\boldsymbol{\Sigma}^*)$$

where $\mathbf{B}^*$ is the bias and $\boldsymbol{\Sigma}^*$ is the cov-matrix of $\sqrt{n}\,(\boldsymbol{\rho}_n^*-\boldsymbol{\rho})$. Then, the asymptotic distributional bias and MSE matrix is given by

$$\mathbf{B}^*=\lim E[\sqrt{n}\,(\boldsymbol{\rho}_n^*-\boldsymbol{\rho})]$$
$$M^*(\boldsymbol{\rho}_n^*)=\boldsymbol{\Sigma}^*+\mathbf{B}^*\mathbf{B}^{*'}$$

The asymptotic distributional quadratic risk (ADQR) is then given by

$$R(\boldsymbol{\rho}_n^*;\mathbf{Q})=tr[\boldsymbol{\Sigma}^*\mathbf{Q}]+\mathbf{B}^{*'}\mathbf{Q}\mathbf{B}^*$$

where $\mathbf{Q}$ is the matrix associated with the loss function

$$L(\boldsymbol{\rho}_n^*;\boldsymbol{\rho})=n(\boldsymbol{\rho}_n^*-\boldsymbol{\rho})'\mathbf{Q}(\boldsymbol{\rho}_n^*-\boldsymbol{\rho}).$$

Then, the following Theorem gives the asymptotic bias, MSE matrices and risk expressions.

Theorem 3.2 *Under $K_{(n)}$ and the assumed conditions of Theorem 3.1, as $n\to\infty$, the following holds.*

$$(a)\quad b_1(\tilde{\boldsymbol{\rho}}_n)=0,\ M_1(\tilde{\boldsymbol{\rho}}_n)=\sigma_\varphi^2\gamma^{-2}\boldsymbol{\Sigma}^{-1}\text{ and }R_1(\tilde{\boldsymbol{\rho}}_n;\mathbf{Q})=tr[\mathbf{Q}\boldsymbol{\Sigma}^{-1}].$$
$$(b)\quad b_2(\hat{\boldsymbol{\rho}}_n)=-\boldsymbol{\delta},\ M_2(\hat{\boldsymbol{\rho}}_n)=\sigma_\varphi^2\gamma^{-2}(\boldsymbol{\Sigma}^{-1}-\mathbf{A})+\boldsymbol{\delta}\boldsymbol{\delta}'$$
$$R_2(\hat{\boldsymbol{\rho}}_n;\mathbf{Q})=\sigma_\varphi^2\gamma^{-1}\,tr[\mathbf{Q}(\boldsymbol{\Sigma}^{-1}-\mathbf{A})]+\boldsymbol{\delta}'\mathbf{Q}\boldsymbol{\delta}.$$
$$(c)\quad b_3(\hat{\boldsymbol{\rho}}_n^{PT})=-\boldsymbol{\delta}\mathcal{H}_q(\chi_q^2(\alpha);\Delta^2),\ \Delta^2=\sigma_\varphi^{-2}\gamma^2(\boldsymbol{\delta}'\boldsymbol{\Sigma}^{-1}\boldsymbol{\delta})$$
$$M_3(\hat{\boldsymbol{\rho}}_n^{PT})=\sigma_\varphi^2\gamma^{-2}\boldsymbol{\Sigma}^{-1}-\sigma_\varphi^2\gamma^{-2}\mathbf{A}\mathcal{H}_q(\chi_q^2(\alpha);\Delta^2)$$
$$+\ \boldsymbol{\delta}\boldsymbol{\delta}'\{2\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2)-\mathcal{H}_{q+4}(\chi_q^2(\alpha);\Delta^2)\}$$
$$R_3(\hat{\boldsymbol{\rho}}_n^{PT};\mathbf{Q})=\sigma_\varphi^2\gamma^{-2}\,tr(\mathbf{Q}\boldsymbol{\Sigma}^{-1})-\sigma_\varphi^2\gamma^{-2}\,tr(\mathbf{Q}\mathbf{A})\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2)$$
$$+\ (\boldsymbol{\delta}'\mathbf{Q}\boldsymbol{\delta})\{2\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2)-\mathcal{H}_{q+4}(\chi_q^2(\alpha);\Delta^2)\}.$$
$$(d)\quad b_4(\hat{\boldsymbol{\rho}}_n^{s})=-(q-2)\boldsymbol{\delta}E[\chi_{q+2}^{-2}(\Delta^2)]$$
$$M_4(\hat{\boldsymbol{\rho}}_n^{s})=\sigma_\varphi^2\gamma^{-2}\boldsymbol{\Sigma}^{-1}-(q-2)\sigma_\varphi^2\gamma^{-2}\mathbf{A}\{2E[\chi_{q+2}^{-2}(\Delta^2)]$$

$$
\begin{aligned}
&- (q-2)E(\chi_{q+2}^{-4}(\Delta^2)\} + (q-2)\boldsymbol{\delta}\boldsymbol{\delta}^{'}\{2E[\chi_{q+2}^{-2}(\Delta^2)] \\
&- 2E[\chi_{q+2}^{-4}(\Delta^2)] + (q-2)E[\chi_{q+4}^{-4}(\Delta^2)]\} \\
R_4(\hat{\boldsymbol{\rho}}_n^s;\mathbf{Q}) &= \sigma_\varphi^2\gamma^{-2}\, tr(\mathbf{Q}\boldsymbol{\Sigma}^{-1}) - (q-2)\sigma_\varphi^2\gamma^{-2}\, tr(\mathbf{QA})\{2E[\chi_{q+2}^{-2}(\Delta^2)] \\
&- (q-2)E[\chi_{q+2}^{-4}(\Delta^2)]\} + (q^2-4)(\boldsymbol{\delta}^{'}\mathbf{Q}\boldsymbol{\delta})E[\chi_{q+4}^{-4}\Delta^2)].
\end{aligned}
$$

$$
\begin{aligned}
(e)\quad b_5(\hat{\boldsymbol{\rho}}_n^{s+}) &= -(q-2)\boldsymbol{\delta}\{E[\chi_{q+2}^{-2}(\Delta^2)] + \mathcal{H}_{q+2}(\chi_{q-2}^2(\alpha);\Delta^2) \\
&- E[\left(1-(q-2)\chi_{q+2}^{-2}(\Delta^2)\right) I\left(\chi_{q+2}^2(\Delta^2) < q-2\right)]\} \\
M_5(\hat{\boldsymbol{\rho}}_n^{s+}) &= M_4(\hat{\boldsymbol{\rho}}_n^s) - (q-2)\sigma_\varphi^2\gamma^{-2}\mathbf{A}E[\left(1-(q-2)\chi_{q+2}^{-2}(\Delta^2)\right)^2 \\
&I\left(\chi_{q+2}^2(\Delta^2) < q-2\right)] \\
&+ \boldsymbol{\delta}\boldsymbol{\delta}^{'}\{2E[\left(1-(q-2)\chi_{q+2}^{-2}(\Delta^2)\right) I\left(\chi_{q+2}^2(\Delta^2) < q-2\right)] \\
&- E[\left(1-(q-2)\chi_{q+4}^{-2}(\Delta^2)\right)^2 I\left(\chi_{q+4}^2(\Delta^2) < q-2\right)]\} \\
R_5(\hat{\boldsymbol{\rho}_n}^{s+};\mathbf{Q}) &= tr[\mathbf{Q}M_4(\hat{\boldsymbol{\rho}}_n^s)] - (q-2)\sigma_\varphi^2\gamma^{-2}\, tr(\mathbf{QA}) \\
&\times E[\left(1-(q-2)\chi_{q+4}^{-2}(\Delta^2)\right)^2 I\left(\chi_{q+4}^2(\Delta^2) < q-2\right)] \\
&+ (\boldsymbol{\delta}^{'}\mathbf{Q}\boldsymbol{\delta})\{2E[\left(1-(q-2)\chi_{q+2}^{-2}(\Delta^2)\right) I\left(\chi_{q+2}^2(\Delta^2) < q-2\right)] \\
&- E[\left(1-(q-2)\chi_{q+4}^{-2}(\Delta^2)\right)^2 I\left(\chi_{q+4}^2(\Delta^2) < q-2\right)]\}
\end{aligned}
$$

It is well-known that (see Saleh (2006, ch7)) the risk ordering is given by

$$R_5(\hat{\boldsymbol{\rho}}_n^{s+};\mathbf{Q}) \le R_4(\hat{\boldsymbol{\rho}}_n^s;\mathbf{Q}) \le R_1(\tilde{\boldsymbol{\rho}}_n;\mathbf{Q})\ \forall\ \Delta^2 \in R^+.$$

and under H_0, it is given by

$$R_2(\hat{\boldsymbol{\rho}}_n;\mathbf{Q}) \le R_3(\hat{\boldsymbol{\rho}}_n^{PT};\mathbf{Q}) \le R_5(\hat{\boldsymbol{\rho}}_n^{s+};\mathbf{Q}) \le R_4(\hat{\boldsymbol{\rho}}_n^s;\mathbf{Q}) \le R_1(\tilde{\boldsymbol{\rho}}_n;\mathbf{Q}).$$

The position of $\hat{\boldsymbol{\rho}}_n^{PT}$ changes between $R_2(\hat{\boldsymbol{\rho}}_n;\mathbf{Q})$ and $R_5(\hat{\boldsymbol{\rho}}_n^{s+};\mathbf{Q})$ to in between $R_4(\hat{\boldsymbol{\rho}}_n^s;\mathbf{Q})$ and $R_1(\tilde{\boldsymbol{\rho}}_n;\mathbf{Q})$. The picture changes when Δ^2 moves from the origin. For details see page 362 of Saleh (2006).

8.4 Ridge Autoregression R-estimators of $\boldsymbol{\rho}$

In this section, we define the following ridge autoregression R-estimators of $\boldsymbol{\rho}$ using Hoerl and Kennard (1970) ridge regression estimators

$$\boldsymbol{\rho}_n^*(k) = \mathbf{R}_n(k)\boldsymbol{\rho}_n^*,\ \ \mathbf{R}_n(k) = \left(\mathbf{I}_p + k(\frac{1}{n}\boldsymbol{\Sigma}_n)^{-1}\right)^{-1}. \tag{8.26}$$

where $\mathbf{R}_n(k)$ is the "ridge factor" and $\boldsymbol{\rho}_n^*$ stands for $\tilde{\boldsymbol{\rho}}_n, \hat{\boldsymbol{\rho}}_n, \hat{\boldsymbol{\rho}}_n^{PT}, \hat{\boldsymbol{\rho}}_n^{s}$ and $\hat{\boldsymbol{\rho}}_n^{s+}$ respectively. Note that

$$P_{\lim}\, \mathbf{R}_n(k) = \mathbf{R}(k) = \left(\mathbf{I}_p + k(\boldsymbol{\Sigma})^{-1}\right)^{-1} \tag{8.27}$$

We may now find the asymptotic distributional biases, MSE matrices and risk expressions of these estimators based on the following Theorem.

First we consider H-K ridge autoregression R-estimators.

Theorem 4.1 *Under $\{K_{(n)}\}$ and the assumed regularity conditions of Theorem 2.1 as $n \to \infty$, the following holds.*

$$(a) \quad \begin{pmatrix} \sqrt{n}\,(\tilde{\boldsymbol{\rho}}_n(k) - \boldsymbol{\rho}) \\ \sqrt{n}\,(\hat{\boldsymbol{\rho}}_n(k) - \boldsymbol{\rho}) \\ \sqrt{n}\,(\tilde{\boldsymbol{\rho}}_n(k) - \hat{\boldsymbol{\rho}}(k)) \end{pmatrix} \xrightarrow{\mathcal{D}} N_{3\,p}\left\{ \begin{pmatrix} -k\mathbf{R}^{-1}(k)\boldsymbol{\rho} \\ -[k\mathbf{R}^{-1}(k)\boldsymbol{\rho} + \mathbf{R}(k)\boldsymbol{\rho}] \\ \mathbf{R}(k)\boldsymbol{\rho} \end{pmatrix}; \sigma_\varphi^2 \gamma^{-2} \boldsymbol{\Sigma}^* \right\} \tag{8.28}$$

$$where\ \mathbf{R}^{-1}(k) = (\boldsymbol{\Sigma} + k\mathbf{I}_p)^{-1},\ \boldsymbol{\delta} = \boldsymbol{\Sigma}^{-1}\mathbf{H}'(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}')^{-1}\boldsymbol{\xi}$$

and $\quad \mathbf{A} = \boldsymbol{\Sigma}^{-1}\mathbf{H}'(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}')^{-1}\mathbf{H}\boldsymbol{\Sigma}^{-1}$ with

$$\boldsymbol{\Sigma}^* = \begin{pmatrix} \mathbf{R}(k)\boldsymbol{\Sigma}^{-1}\mathbf{R}'(k) & \mathbf{R}(k)(\boldsymbol{\Sigma}^{-1} - \mathbf{A})\mathbf{R}'(k) & \mathbf{R}(k)\mathbf{A}\mathbf{R}'(k) \\ \mathbf{R}(k)(\boldsymbol{\Sigma}^{-1} - \mathbf{A})\mathbf{R}'(k) & \mathbf{R}(k)(\boldsymbol{\Sigma}^{-1} - \mathbf{A})\mathbf{R}'(k) & 0 \\ \mathbf{R}(k)\mathbf{A}\mathbf{R}'(k) & 0 & \mathbf{R}(k)\mathbf{A}\mathbf{R}'(k) \end{pmatrix} \tag{8.29}$$

(b) The asymptotic distributional bias (ADB), MSE (ADMSE) matrices and quadratic risks (ADQR) are given by

$$\begin{aligned} (i) \quad b_1(\tilde{\boldsymbol{\rho}}_n(k)) &= [\mathbf{R}(k) - \mathbf{I}_p]\boldsymbol{\rho} = \mathbf{R}(k)[\mathbf{I}_p - (\mathbf{I}_p + k\boldsymbol{\Sigma}^{-1})]\boldsymbol{\rho} = -k\mathbf{R}(k)\boldsymbol{\Sigma}^{-1}\boldsymbol{\rho} \\ &= -k[\boldsymbol{\Sigma} + k\mathbf{I}_p]^{-1}\boldsymbol{\rho} = -k\mathbf{R}^{-1}(k)\boldsymbol{\rho}. \\ M_1(\tilde{\boldsymbol{\rho}}_n(k)) &= \sigma_\varphi^2\gamma^{-2}[\mathbf{R}(k)\boldsymbol{\Sigma}_{xx}^{-1}\mathbf{R}'(k)] + [\mathbf{R}(k) - \mathbf{I}_p]\boldsymbol{\rho}\boldsymbol{\rho}'[\mathbf{R}(k) - \mathbf{I}_p]'. \\ R_1(\tilde{\boldsymbol{\rho}}_n(k); \mathbf{W}) &= \sigma_\varphi^2\gamma^{-2} tr\left(\mathbf{W}[\mathbf{R}(k)\boldsymbol{\Sigma}_{xx}^{-1}\mathbf{R}'(k)]\right) + \boldsymbol{\rho}'[\mathbf{R}(k) - \mathbf{I}_p]'\mathbf{W}[\mathbf{R}(k) - \mathbf{I}_p]\boldsymbol{\rho}. \end{aligned}$$

$$\begin{aligned} (ii) \quad b_2(\hat{\boldsymbol{\rho}}_n(k)) &= [\mathbf{R}(k) - \mathbf{I}_p]\boldsymbol{\rho} + \mathbf{R}(k)\boldsymbol{\delta} = \mathbf{B} \text{ say.} \\ M_2(\hat{\boldsymbol{\rho}}_n(k)) &= \sigma_\varphi^2\gamma^{-2}\left\{[\mathbf{R}(k)\boldsymbol{\Sigma}^{-1}\mathbf{R}'(k)] - [\mathbf{R}(k)\mathbf{A}\mathbf{R}'(k)]\right\} + \mathbf{B}\mathbf{B}'. \\ R_2(\hat{\boldsymbol{\rho}}_n(k); \mathbf{W}) &= \sigma_\varphi^2\gamma^{-2} tr\left(\mathbf{W}[\mathbf{R}(k)(\boldsymbol{\Sigma}^{-1} - \mathbf{A})\mathbf{R}'(k)]\right) + \mathbf{B}'\mathbf{W}\mathbf{B}. \end{aligned}$$

$$\begin{aligned} (iii) \quad b_3(\hat{\boldsymbol{\rho}}_n^{PT}(k)) &= \left([\mathbf{R}(k) - \mathbf{I}_p]\boldsymbol{\rho} + \mathbf{R}(k)\,\boldsymbol{\delta}\mathcal{H}_{q+2}(\chi_q^2(\alpha); \Delta^2)\right). \\ M_3(\hat{\boldsymbol{\rho}}_n^{PT}(k)) &= \sigma_\varphi^2\gamma^{-2}[\mathbf{R}(k)\boldsymbol{\Sigma}_{xx}^{-1}\mathbf{R}'(k)] - \sigma_\varphi^2\gamma^{-2}[\mathbf{R}(k)\mathbf{A}\mathbf{R}'(k)]\mathcal{H}_{q+2}(\chi_q^2(\alpha); \Delta^2) \end{aligned}$$

$$
\begin{aligned}
&+\ [\mathbf{R}(k)\boldsymbol{\delta}\boldsymbol{\delta}^{'}\mathbf{R}^{'}(k)]\{2\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2)-\mathcal{H}_{q+4}(\chi_q^2(\alpha);\Delta^2)\}\\
&+\ k^2[\mathbf{R}(k)-\mathbf{I}_p]\boldsymbol{\rho}\boldsymbol{\rho}^{'}[\mathbf{R}(k)-\mathbf{I}_p]^{'}+k[\mathbf{R}(k)\boldsymbol{\delta}\boldsymbol{\rho}^{'}(\mathbf{R}(k)-\mathbf{I}_p)\\
&+\ \left\{(\mathbf{R}(k)-\mathbf{I}_p)\boldsymbol{\rho}-\mathbf{R}(k)\boldsymbol{\delta}\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2)\right\}\\
&\times\ \left\{[\mathbf{R}(k)-\mathbf{I}_p]\boldsymbol{\rho}-\mathbf{R}(k)\boldsymbol{\delta}\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2)\right\}^{'}.\\
R_3(\hat{\boldsymbol{\rho}}_n^{PT}(k);\mathbf{W})&=\sigma_\varphi^2\gamma^{-2}\ tr[\mathbf{W}\left(\mathbf{R}(k)\{\boldsymbol{\Sigma}^{-1}-\mathbf{A}\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2)\}\mathbf{R}^{'}(k)\right)]\\
&+\ tr[\mathbf{W}\left(\mathbf{R}(k)\boldsymbol{\delta}\boldsymbol{\delta}^{'}\mathbf{R}^{'}(k)\right)]\{2\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2)\\
&-\mathcal{H}_{q+4}(\chi_q^2(\alpha);\Delta^2)\}+\ k^2\boldsymbol{\rho}^{'}[\mathbf{R}(k)-\mathbf{I}_p]^{'}\mathbf{W}[\mathbf{R}(k)-\mathbf{I}_p]\boldsymbol{\rho}\\
&+k\ tr[\mathbf{W}\{\mathbf{R}(k)\boldsymbol{\delta}\boldsymbol{\rho}^{'}(\mathbf{R}(k)-\mathbf{I}_p]\\
&+\ \left\{(\mathbf{R}(k)-\mathbf{I}_p)\boldsymbol{\rho}-\mathbf{R}(k)\boldsymbol{\delta}\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2)\right\}^{'}\\
&\times\mathbf{W}\left\{[\mathbf{R}(k)-\mathbf{I}_p]\boldsymbol{\rho}-\mathbf{R}(k)\boldsymbol{\delta}\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2)\right\}.
\end{aligned}
$$

$$
\begin{aligned}
(iv)\ \ b_4(\hat{\boldsymbol{\rho}}_n^s(k))&=-\left([\mathbf{R}(k)-\mathbf{I}_p]+(q-2)\mathbf{R}(k)\boldsymbol{\delta}E[\chi_{q+2}^{-2}(\Delta^2)]\right)\\
M_4(\hat{\boldsymbol{\rho}}_n^s(k))&=[\mathbf{R}(k)\boldsymbol{\Sigma}^{-1}\mathbf{R}^{'}(k)]-(q-2)[\mathbf{R}(k)\mathbf{A}\mathbf{R}^{'}(k)]\{2E[\chi_{q+2}^{-2}(\Delta^2)]\\
&-\ (q-2)E[\chi_{q+2}^{-4}(\Delta^2)]\}+(q^2-4)[\mathbf{R}(k)\boldsymbol{\delta}\boldsymbol{\delta}^{'}\mathbf{R}^{'}(k)]E[\chi_{q+4}^{-4}(\Delta^2)]\\
&+\ \left\{[\mathbf{R}(k)-\mathbf{I}_p]\boldsymbol{\rho}-\mathbf{R}(k)\boldsymbol{\delta}E[\chi_{q+2}^{-2}(\Delta^2)]\right\}\\
&\left\{[\mathbf{R}(k)-\mathbf{I}_p]\boldsymbol{\rho}-\mathbf{R}(k)\boldsymbol{\delta}E[\chi_{q+2}^{-2}(\Delta^2)]\right\}^{'}\\
R_4(\hat{\boldsymbol{\rho}}_n^s(k);\mathbf{W})&=\sigma_\varphi^2\gamma^{-2}\ tr(\mathbf{W}[\mathbf{R}(k)\boldsymbol{\Sigma}^{-1}\mathbf{R}^{'}(k)])-(q-2)\sigma_\varphi^2\gamma^{-2}\ tr\left(\mathbf{W}\mathbf{R}(k)\mathbf{A}\mathbf{R}^{'}(k)\right)\\
&\times\{2E[\chi_{q+2}^{-2}(\Delta^2)]-(q-2)E[\chi_{q+2}^{-4}(\Delta^2)]\}+(q^2-4)\ tr\\
&\left(\mathbf{W}[\mathbf{R}(k)\boldsymbol{\delta}\boldsymbol{\delta}^{'}\mathbf{R}^{'}(k)]\right)\times E[\chi_{q+4}^{-4}\Delta^2)]\\
&+\left\{[\mathbf{R}(k)-\mathbf{I}_p]\boldsymbol{\rho}-\mathbf{R}(k)\boldsymbol{\delta}E[\chi_{q+2}^{-2}(\Delta^2)]\right\}^{'}\mathbf{W}\\
&\times\ \left\{[\mathbf{R}(k)-\mathbf{I}_p]\boldsymbol{\rho}-\mathbf{R}(k)\boldsymbol{\delta}E[\chi_{q+2}^{-2}(\Delta^2)]\right\}.
\end{aligned}
$$

$$
\begin{aligned}
(v)\quad b_5(\hat{\boldsymbol{\rho}}_n^{s+}(k))&=[\mathbf{R}(k)-\mathbf{I}_p]\boldsymbol{\rho}-\mathbf{R}(k)\boldsymbol{\delta}\{\mathcal{H}_{q+2}(\chi_{q-2}^2(\alpha);\Delta^2)\\
&-(q-2)\{E[\chi_{q+2}^2(\alpha);\Delta^2)]\}-\ E[\chi_{q+2}^{-2}(\Delta^2)]I\left(\chi_{q+2}^2(\Delta^2)\right)]\}\\
M_5(\hat{\boldsymbol{\rho}}_n^{s+}(k))&=M_4(\hat{\boldsymbol{\rho}}_n^s(k))-\sigma_\varphi^2\gamma^{-2}[\mathbf{R}(k)\mathbf{A}\mathbf{R}^{'}(k)]E[\left(1-(q-2)\chi_{q+2}^{-2}(\Delta^2)\right)^2\\
&\times\ I\left(\chi_{q+2}^2(\Delta^2)<q-2\right)]
\end{aligned}
$$

$$
\begin{aligned}
&+ [\mathbf{R}(k)\boldsymbol{\delta}\boldsymbol{\delta}^{'}\mathbf{R}^{'}(k)]\{2E[\left(1-(q-2)\chi_{q+2}^{-2}(\Delta^2)\right)\\
&\times I\left(\chi_{q+2}^2(\Delta^2)<q-2\right)]\\
&- E[\left(1-(q-2)\chi_{q+4}^{-2}(\Delta^2)\right)^2 I\left(\chi_{q+4}^2(\Delta^2)<q-2\right)]\}\\
&+ [b_5(\hat{\boldsymbol{\rho}}_n^{s+}(k))][b_5(\hat{\boldsymbol{\rho}}_n^{s+}(k))]^{'}.
\end{aligned}
$$

$$
\begin{aligned}
R_5(\hat{\boldsymbol{\rho}}_n^{s+}(k);\mathbf{W}) &= R_4(\hat{\boldsymbol{\rho}}_n^{s}(k);\mathbf{W})] - \{\sigma_{\varphi}^2\gamma^{-2}\, tr(\mathbf{R}(k)\mathbf{A}\mathbf{R}^{'}(k))\\
&E[\left(1-(q-2)\chi_{q+4}^{-2}(\Delta^2)\right)^2 \times I\left(\chi_{q+4}^2(\Delta^2)<q-2\right)]\}\\
&+ (\boldsymbol{\delta}^{'}\mathbf{R}^{'}(k)\mathbf{W}\mathbf{R}(k)\boldsymbol{\delta})\{2E[\left(1-(q-2)\chi_{q+2}^{-2}(\Delta^2)\right)^2\\
&I\left(\chi_{q+2}^2(\Delta^2)<q-2\right)] - E[\left(1-(q-2)\chi_{q+2}^{-2}(\Delta^2)\right)\\
&I\left(\chi_{q+2}^2(\Delta^2)<q-2\right)]\} + [b_5(\hat{\boldsymbol{\rho}}_n^{s+}(k))]^{'}\mathbf{W}[b_5(\hat{\boldsymbol{\rho}}_n^{s+}(k))].
\end{aligned}
$$

8.5 Comparison of the Five Ridge Autoregression R-estimators

In this section, we consider the comparison of the Ridge Autoregression R-estimators under a quadratic loss functions. Notice that the risk expression of the five ridge autoregression rank estimators are functions of the departure parameter Δ^2 as well as the "ridge constant", k. First, we consider the comparisons when the risk expressions are function of k in Sect. 8.5.1 and in Sect. 8.5.2 we consider the comparison as a function of Δ^2. In this respect, we present the comparison in a sequence theorems that follow in each section.

8.5.1 Comparison of RARE's as a function of ridge constant

It is clear that, $\boldsymbol{\Sigma}$ is a positive definite matrix so that there exist an orthogonal matrix $\boldsymbol{\Gamma}$ such that $\boldsymbol{\Sigma} = \boldsymbol{\Gamma}\boldsymbol{\Lambda}\boldsymbol{\Gamma}^{'}$ and $\boldsymbol{\Lambda} = \boldsymbol{\Gamma}^{'}\boldsymbol{\Sigma}\boldsymbol{\Gamma} = \text{Diag}(\lambda_1, \cdots, \lambda_p)$ where $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p > 0$ are the eigenvalues of $\boldsymbol{\Sigma}$. It is easy to see that the eigenvalues of $\mathbf{R}(k)$ and $\mathbf{R}^{-1}(k) = \boldsymbol{\Sigma} + k\mathbf{I}_p$ are $\frac{\lambda_1}{\lambda_1+k}, \cdots, \frac{\lambda_p}{\lambda_p+k}$ and $\lambda_1 + k, \cdots, \lambda_p + k$ respectively. with this background, we get the following identities:

$$
\begin{aligned}
&(i) \quad \boldsymbol{\rho}^{'}\mathbf{R}^{-1}(k)\boldsymbol{\rho} = \boldsymbol{\alpha}^{'}(\boldsymbol{\Lambda} + k\mathbf{I}_p)\boldsymbol{\alpha} = \textstyle\sum_{j=1}^{p} \frac{\alpha_j^2}{(\lambda_j+k)^2},\ \boldsymbol{\alpha} = \boldsymbol{\Gamma}^{'}\boldsymbol{\rho} \qquad (8.30)\\
&(ii) \quad tr(\mathbf{R}(k)\boldsymbol{\Sigma}^{-1}\mathbf{R}(k)) = \textstyle\sum_{j=1}^{p} \frac{\lambda_j}{(\lambda_j+k)^2}\\
&(iii) \quad tr[\mathbf{R}(k)\boldsymbol{\Sigma}^{-1}\mathbf{H}^{'}(\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{H}^{'})^{-1}\mathbf{H}\boldsymbol{\Sigma}^{-1}\mathbf{R}(k)] = \textstyle\sum_{j=1}^{p} \frac{h_{jj}^*}{(\lambda_j+k)^2}
\end{aligned}
$$

where $h_{jj}^{*} \geq 0$ is the j^{th} diagonal element of

$$(iv) \quad \mathbf{\Gamma}^{'}\mathbf{H}^{'}(\mathbf{H}\mathbf{\Sigma}^{-1}\mathbf{H}^{'})^{-1}\mathbf{H}\mathbf{\Gamma} = \mathbf{H}^{*},\ \boldsymbol{\delta}^{'}\mathbf{\Sigma}\boldsymbol{\delta} = \sum_{j=1}^{p}\delta_{j}^{*2}$$

where $\boldsymbol{\delta} = \mathbf{\Sigma}^{-1}\mathbf{H}^{'}(\mathbf{H}\mathbf{\Sigma}^{-1}\mathbf{H}^{'})^{-1}(\mathbf{H}\boldsymbol{\rho} - h)$. Assume that $\mathbf{W} = \mathbf{I}_{p}$ in Theorem 4.1 (b).

8.5.1.1 Comparison of $\tilde{\boldsymbol{\rho}}_{n}(k)$ and $\tilde{\boldsymbol{\rho}}_{n}$

The comparison results are presented in the following Theorem.

Theorem 5.1.1 *Under $K_{(n)}$ and basics assumptions, there exists a $k \in (0, k_0)$ where $k_0 = \frac{\sigma_{\varphi}^{2}\gamma^{-2}}{\alpha_{\max}}$ such that the unrestricted ridge autoregression estimator $\tilde{\boldsymbol{\rho}}_{n}(k)$ has smaller mean square error (mse) than the unrestricted estimator, $\tilde{\boldsymbol{\rho}}_{n}$ as $n \to \infty$.*

Proof Consider the asymptotic distributional mse of $\tilde{\boldsymbol{\rho}}_{n}(k)$ given by

$$R_{1}(\tilde{\boldsymbol{\rho}}_{n}(k); \mathbf{I}_{p}) = \sigma_{\varphi}^{2}\gamma^{-2}\sum_{j=1}^{p}\frac{\lambda_{j}}{(\lambda_{j}+k)^{2}} + k^{2}\sum_{j=1}^{p}\frac{\alpha_{j}^{2}}{(\lambda_{j}+k)^{2}}. \tag{8.31}$$

It is obvious that for $k = 0$, the first term equals $\sigma_{\varphi}\gamma^{-2}\sum_{j=1}^{p}\frac{1}{\lambda_{j}} = R_{1}(\tilde{\boldsymbol{\rho}}_{n}; \mathbf{I}_{p})$ and second term equals zero respectively. The first term is a continuous, monotonically decreasing function of k and its derivative w.r.t k approaches $-\infty$ as $k \to 0^{+}$ and $\lambda_{p} \to 0$. The second term is also continuous, monotonically increasing function of k and its derivative tends to zero as $k \to 0^{+}$. We note that the second term tends to $\boldsymbol{\rho}^{'}\boldsymbol{\rho}$ as $k \to \infty$. Differentiating (8.31) w.r.t k we get

$$\frac{\partial R_{1}}{\partial k}(\tilde{\boldsymbol{\rho}}_{n}(k); \mathbf{I}_{p}) = 2\sum_{j=1}^{p}\frac{\lambda_{j}}{(\lambda_{j}+k)^{3}}(k\alpha_{j} - \sigma_{\varphi}^{2}\gamma^{-2}). \tag{8.32}$$

Thus, a sufficient condition for (8.32) to be negative is that there exists a $k \in (0, k_0)$ such that $\tilde{\boldsymbol{\rho}}_{n}(k)$ has smaller mse than that of $\tilde{\boldsymbol{\rho}}_{n}$, where $k_0 = \frac{\sigma_{\varphi}^{2}\gamma^{-2}}{\max_{1\leq j\leq p}\{\alpha_{j}\}}$.

8.5.1.2 Comparison between $\hat{\boldsymbol{\rho}}_{n}(k)^{PT}$ and $\hat{\boldsymbol{\rho}}_{n}^{PT}$

First, note that for $\alpha = 0$, $\mathcal{H}_{q+2}(\chi_{q}^{2}(\alpha); \Delta^{2}) = 1$. In this case, one compares between $\hat{\boldsymbol{\rho}}_{n}(k)$ and $\hat{\boldsymbol{\rho}}_{n}$. On the other hand, if $\alpha = 1$, then $\mathcal{H}_{q+2}(\chi_{q}^{2}(\alpha); \Delta^{2}) = 0$. Hence, compares $\tilde{\boldsymbol{\rho}}_{n}(k)$ and $\tilde{\boldsymbol{\rho}}_{n}$ which have done in Sect. 8.5.1.1.

The comparison of $\hat{\boldsymbol{\rho}}_{n}(k)^{PT}$ and $\hat{\boldsymbol{\rho}}_{n}^{PT}$ is given in the following Theorem.

Theorem 5.1.2 *Under $K_{(n)}$ and the regularity conditions as $n \to \infty$, a sufficient condition for the mse of $\hat{\boldsymbol{\rho}}_n(k)^{PT}$ is less than the mse $\hat{\boldsymbol{\rho}}_n^{PT}$ is that there exists a $k \in (0, k_{PT}(\Delta^2, \alpha))$ where*

$$k_{PT}(\Delta^2, \alpha) = \frac{f_1(\Delta^2, \alpha)}{g_1(\Delta^2, \alpha)} \tag{8.33}$$

with

$$\begin{aligned} f_1(\Delta^2, \alpha) = \min_{1 \le j \le p} [&\sigma_\varphi^2 \gamma^{-2} \left\{\lambda_j - h_{jj}^* \mathcal{H}_{q+2}(\chi_q^2(\alpha); \Delta^2)\right\} \\ &+ \lambda_j^2 \delta_j^{*2} \left\{2\mathcal{H}_{q+2}(\chi_q^2(\alpha); \Delta^2) - \mathcal{H}_{q+4}(\chi_q^2(\alpha); \Delta^2)\right\} \\ &- \alpha_j \lambda_j^2 \delta_j^* \mathcal{H}_{q+2}(\chi_q^2(\alpha); \Delta^2)] \\ \text{and } g_1(\Delta^2, \alpha) = \max_{1 \le j \le p} &\left[\alpha_j \lambda_j \left\{\alpha_j - \delta_j^* \mathcal{H}_{q+2}(\chi_q^2(\alpha); \Delta^2)\right\}\right]. \end{aligned}$$

Proof

$$\begin{aligned} R_3(\hat{\boldsymbol{\rho}}_n^{PT}(k); \mathbf{I}_p) = \sum_{j=1}^{p} \frac{1}{(\lambda_j + k)^2} &\{\sigma_\varphi^2 \gamma^{-2} [\lambda_j - h_{jj}^* \mathcal{H}_{q+2}(\chi_q^2(\alpha); \Delta^2)] \\ &+ \lambda_j^2 \delta_j^{*2} [2\mathcal{H}_{q+2}(\chi_q^2(\alpha); \Delta^2) - \mathcal{H}_{q+4}(\chi_q^2(\alpha); \Delta^2)] \\ &+ k^2 \alpha_j^2 + 2k\alpha_j \lambda_j \delta_j^{*2} \mathcal{H}_{q+2}(\chi_q^2(\alpha); \Delta^2)\} \end{aligned}$$

Differentiating with respect to k, we obtain

$$\begin{aligned} \frac{\partial R_3}{\partial k}(\hat{\boldsymbol{\rho}}_n^{PT}(k); \mathbf{I}_p) = 2 \sum_{j=1}^{p} \frac{1}{(\lambda_j + k)^3} &\{k\alpha_j \lambda_j [\alpha_j - h_{jj}^* \mathcal{H}_{q+2}(\chi_q^2(\alpha); \Delta^2)] \\ &- [\sigma_\varphi^2 \gamma^{-2} \left(\lambda_j - h_{jj}^* \mathcal{H}_{q+2}(\chi_q^2(\alpha); \Delta^2)\right) \\ &+ \lambda_j \delta_j^{*2} [2\mathcal{H}_{q+2}(\chi_q^2(\alpha); \Delta^2) - \mathcal{H}_{q+4}(\chi_q^2(\alpha); \Delta^2)] \\ &- \alpha_j \lambda_j^2 \delta_j^* \mathcal{H}_{q+2}(\chi_q^2(\alpha); \Delta^2)]\} \end{aligned} \tag{8.34}$$

Hence, a sufficient condition that $\hat{\boldsymbol{\rho}}_n^{PT}(k)$ has mse less than Hannan (1970) the mse of $\hat{\boldsymbol{\rho}}_n^{PT}$ is that there exists a $k \in (0, k_{PT}(\Delta^2, \alpha))$ where $k_{PT}(\Delta^2, \alpha)$ is defined by (8.33). Consequently, the mse of $\hat{\boldsymbol{\rho}}_n(k)$ is less than the mse of $\hat{\boldsymbol{\rho}}_n$.

8.5.1.3 Comparison between $\hat{\boldsymbol{\rho}}_n^s(k)$ and $\hat{\boldsymbol{\rho}}_n^s$

The following theorem gives the sufficient conditions for the dominance of $\hat{\boldsymbol{\rho}}_n^s(k)$ over $\hat{\boldsymbol{\rho}}_n^s$.

Theorem 5.1.3 *Under $K_{(n)}$ and the assumed regularity conditions, $R_4(\hat{\boldsymbol{\rho}}_n^s; \mathbf{I}_p) \geq R_4(\hat{\boldsymbol{\rho}}_n^s(k); \mathbf{I}_p)$ holds whenever $k \in (0, k_s)$, where $k_s = \frac{f_2(\Delta^2)}{g_2(\Delta^2)}$ with*

$$\begin{aligned} f_2(\Delta^2) &= \min_{1\leq j\leq p} [\sigma_\varphi^2\gamma^{-2}(\lambda_j - (q-2)h_{jj}^*\{(q-2)E[\chi_{q+2}^{-4}(\Delta^2)] \\ &\quad + 2(1 - \frac{(q+2)\lambda_j^2\delta_j^{*2}}{2\sigma_\varphi^2\gamma^{-2}\Delta^2 h_{jj}^*})\Delta^2 E[\chi_{q+2}^{-4}(\Delta^2)]\} \\ &\quad + (q-2)\lambda_j^2\delta_j^{*2}E[\chi_{q+2}^{-2}(\Delta^2)]] \qquad (8.35) \\ \text{and } g_2(\Delta^2) &= \max_{1\leq j\leq p} \left\{\alpha_j\lambda_j[\alpha_j - (q-2)\delta_j^* E[\chi_{q+2}^{-2}(\Delta^2)]]\right\} \qquad (8.36) \end{aligned}$$

Proof Consider the mse expression for $\hat{\boldsymbol{\rho}}_n^s(k)$ given by

$$\begin{aligned} R_4(\hat{\boldsymbol{\rho}}_n^s(k); \mathbf{I}_p) = \sum_{j=1}^p \frac{1}{(\lambda_j + k)^2}\{\sigma_\varphi^2\gamma^{-2}[\lambda_j - (q-2)h_{jj}^*\{(q-2)E[\chi_{q+4}^{-4}\Delta^2)] \\ + (1 - \frac{(q+2)\lambda_j^2\delta_j^{*2}}{2\sigma_\varphi^2\gamma^{-2}\Delta^2 h_{jj}^*})\Delta^2 E[\chi_{q+4}^{-4}(\Delta^2)]\} \\ + 2k(q-2)\alpha_j\lambda_j\delta_j^* E[\chi_{q+2}^{-2}(\Delta^2)]]\}. \qquad (8.37) \end{aligned}$$

The derivative of $R_4(\hat{\boldsymbol{\rho}}_n^s(k); \mathbf{I}_p)$ w.r.t. k is given by

$$\begin{aligned} \frac{\partial R_4}{\partial k}(\hat{\boldsymbol{\rho}}_n^s(k); \mathbf{I}_p) = 2\sum_{j=1}^p \frac{1}{(\lambda_j + k)^3}\{k\alpha_j\lambda_j[\alpha_j - (q-2)\delta_j^* E[\chi_{q+2}^{-2}(\Delta^2)] \\ - \sigma_\varphi^2\gamma^{-2}\{\lambda_j - (q-2)h_{jj}^*[(q-2)E[\chi_{q+2}^{-2}(\Delta^2)] \\ + (1 - \frac{(q+2)\lambda_j^2\delta_j^{*2}}{2\sigma_\varphi^2\gamma^{-2}\Delta^2 h_{jj}^*}) \\ \times 2\Delta^2 E[\chi_{q+4}^{-4}(\Delta^2)]\} - (q-2)\alpha_j\lambda_j^2\delta_j^* E[\chi_{q+2}^{-2}(\Delta^2)]\}. \qquad (8.38) \end{aligned}$$

Thus, a sufficient condition for (8.37) to be negative is that $k \in (0, k_s)$ where k_s is given by (8.35) and (8.36). QED.

8.5.1.4 Comparison between $\hat{\boldsymbol{\rho}}_n^{s+}(k)$ and $\hat{\boldsymbol{\rho}}_n^{s+}$

The comparison between $\hat{\boldsymbol{\rho}}_n^{s+}(k)$ and $\hat{\boldsymbol{\rho}}_n^{s+}$ is given by the following theorem.

Theorem 5.1.4 *Under $K_{(n)}$ and the assumed regularity conditions as $n \to \infty$, a sufficient condition for the mse of $\hat{\boldsymbol{\rho}}_n^{s+}(k)$ is less than Hannan (1970) $\hat{\boldsymbol{\rho}}_n^{s+}$ is that there exists a $k \in (0, k_{s+})$ where $k_{s+} = \frac{f_3(\Delta^2)}{g_3(\Delta^2)}$ and $f_3(\Delta^2)$ and $g_3(\Delta^2)$ are given by*

$$f_3(\Delta^2) = \min_{1\leq j\leq p} \{\sigma_\varphi^2\gamma^{-2}\{\lambda_j - (q-2)a_{ii}^*[(q-2)E[\chi_{q+2}^{-2}(\Delta^2)]$$

$$
\begin{aligned}
&+ (1 - \frac{(q+2)\lambda_j^2\delta_j^{*2}}{2\sigma_\varphi^2\gamma^{-2}\Delta^2 a_{ii}^*})(2\Delta^2)E[\chi_{q+4}^{-4}(\Delta^2)] \\
&- a_{ii}^* E[(1-(q-2)\chi_{q+2}^{-2}(\Delta^2))^2 I(\chi_{q+2}^2(\Delta^2) \le (q-2))] \\
&- \lambda_i^2\delta_i^{*2} E[(1-(q-2)\chi_{q+4}^{-2}(\Delta^2))^2 I(\chi_{q+4}^2(\Delta^2) \le (q-2))] \\
&+ (\theta_i - 2\delta_i^*)\lambda_i^2\delta_i^* E[((q-2)\chi_{q+4}^{-2}(\Delta^2) - 1) I(\chi_{q+2}^2(\Delta^2) \le (q-2))] \\
&+ dq\theta_i\delta_i^*)\lambda_i^2 E[\chi_{q+2}^{-2}(\Delta^2)]\}\} \qquad (8.39)
\end{aligned}
$$

$$
\begin{aligned}
\text{and } g_3(\Delta^2) = \max_{1\le i\le p} &[\lambda_i\theta_i\{\theta_i + \delta_i^* E((q-2)\chi_{q+2}^{-2}(\Delta^2) - 1) \\
&I(\chi_{q+2}^2(\Delta^2) \le (q-2)) \\
&- (q-2)\delta_i^* E[\chi_{q+2}^{-2}(\Delta^2)]\}].
\end{aligned}
$$

Proof The risk function of $\hat{\boldsymbol{\rho}}_n^{s+}(k)$ can be expressed as

$$
\begin{aligned}
R_5(\hat{\boldsymbol{\rho}}_n^{s+}(k);\mathbf{I}_p) = R_4(\hat{\boldsymbol{\rho}}_n^{s}(k);\mathbf{I}_p) - \sum_{i=1}^{p} \frac{1}{(\lambda_i+k)^2}\{\{\sigma_\varphi^2\gamma^{-2} \\
E\left[(1-(q-2)\chi_{q+2}^{-2}(\Delta^2))^2 I(\chi_{q+2}^2(\Delta^2) \le (q-2))\right] \\
+ \lambda_i^2\delta_i^{*2} E\left[(1-(q-2)\chi_{q+4}^{-2}(\Delta^2))^2 I(\chi_{q+4}^2(\Delta^2) \le (q-2))\right]\} \\
+ 2\lambda_i^2\delta_i^{*2} E\left[((q-2)\chi_{q+2}^{-2}(\Delta^2) - 1) I(\chi_{q+2}^2(\Delta^2) \le (q-2))\right]\} \\
+ 2k\theta_i\lambda_i\delta_i^* E[((q-2)\chi_{q+2}^{-2}(\Delta^2) - 1) I(\chi_{q+2}^2(\Delta^2) \le (q-2))]\}. \qquad (8.40)
\end{aligned}
$$

where $R_4(\hat{\boldsymbol{\rho}}_n^{s}(k);\mathbf{I}_p)$ is given by (8.37). Differentiating $R_5(\hat{\boldsymbol{\rho}}_n^{s+}(k);\mathbf{I}_p)$ with respect to k, we obtain

$$
\begin{aligned}
\frac{\partial R_5(\hat{\boldsymbol{\rho}}_n^{s+}(k);\mathbf{I}_p)}{\partial k} = \frac{\partial R_4(\hat{\boldsymbol{\rho}}_n^{s}(k);\mathbf{I}_p)}{\partial k} + 2\sum_{i=1}^{p} \frac{1}{(\lambda_i+k)^3}\{k\alpha_i\lambda_i\delta_i^* \\
E\left[(q-2)\chi_{q+2}^{-2}(\Delta^2) - 1) I(\chi_{q+2}^2(\Delta^2) \le (q-2))\right] \\
+ \sigma_e^2\{h_{ii}^* E\left[(1-(q-2)\chi_{q+2}^{-2}(\Delta^2))^2 I(\chi_{q+2}^2(\Delta^2) \le (q-2))\right] \\
+ \frac{\lambda_i^2\delta_i^{*2}}{\sigma_\varphi^2\gamma^{-2}} E\left[(1-(q-2)\chi_{q+4}^{-2}(\Delta^2))^2 I(\chi_{q+4}^2(\Delta^2) \le (q-2))\right]\} \\
- (\alpha_i - 2\delta_i^*)\lambda_i^2\delta_i^* E\left[((q-2)\chi_{q+2}^{-2}(\Delta^2) - 1) I(\chi_{q+2}^2(\Delta^2) \le (q-2))\right]\}, \qquad (8.41)
\end{aligned}
$$

where $\frac{\partial R_4(\hat{\boldsymbol{\rho}}_n^s(k);\mathbf{I}_p)}{\partial k}$ is given by (8.38). Hence, a sufficient condition for (8.41) to be negative is that $0 < k < k_{s+}$ where

$$k_{s+} = \frac{f_3(\Delta^2)}{g_3(\Delta^2)}$$

whenever $g_3(\Delta^2) > 0$. QED.

8.5.1.5 Comparison between $\hat{\boldsymbol{\rho}}_n^s(k)$ and $\tilde{\boldsymbol{\rho}}_n$

In this case, we may obtain the Theorem given below.

Theorem 5.1.5 *Under $K_{(n)}$ and the assumed regularity conditions, as $n \to \infty$, a sufficient condition for $\hat{\boldsymbol{\rho}}_n^s(k)$ to have mse value is less than or equal to the mse of $\tilde{\boldsymbol{\rho}}_n$ is that there exists a value of $k \in (0, k_s^*)$ where*

$$k_s^* = \frac{\sigma_\varphi^2 \gamma^{-2} \min_{1\le j\le p} \left\{ h_{jj}^*(q-2)E[\chi_{q+2}^{-4}(\Delta^2)] + (1 - \frac{(q+2)\lambda_j^2 \delta_j^{*2}}{2\sigma_\varphi^2 \gamma^{-2}\Delta^2 h_{jj}^*})(2\Delta^2)E[\chi_{q+4}^{-4}(\Delta^2)] \right\}}{\max_{1\le j\le p} \left\{ 2\alpha_j \lambda_j \delta_j^* E[\chi_{q+2}^{-2}(\Delta^2)] \right\}} \tag{8.42}$$

For proof consider the mse difference $R_1(\tilde{\boldsymbol{\rho}}_n(k);\mathbf{I}_p) - R_4(\hat{\boldsymbol{\rho}}_n^s(k);\mathbf{I}_p) \ge 0$, then k_s^* follows.

8.5.1.6 Comparison of $\hat{\boldsymbol{\rho}}_n^{s+}(k)$ and $\hat{\boldsymbol{\rho}}_n^s(k)$

The result is presented in the following Theorem.

Theorem 5.1.6 *Under $K_{(n)}$ and the assumed regularity conditions, as $n \to \infty$, $\hat{\boldsymbol{\rho}}_n^{s+}(k)$ has smaller mse than $\hat{\boldsymbol{\rho}}_n^s(k)$ for all $k \ge 0$.*

Proof Consider the mse difference of $\hat{\boldsymbol{\rho}}_n^{s+}(k)$ and $\hat{\boldsymbol{\rho}}_n^s(k)$ given by

$$\begin{aligned}
R_4 \quad & (\hat{\boldsymbol{\rho}}_n^s(k);\mathbf{I}_p) - R_5(\hat{\boldsymbol{\rho}}_n^{s+}(k);\mathbf{I}_p) \\
&= \sum_{j=1}^{p} \frac{1}{(\lambda_j + k)^2} \{\sigma_\varphi^2 \gamma^{-2} \\
&\quad \left(h_{jj}^* E\left[(1 - (q-2)\chi_{q+2}^{-2}(\Delta^2))^2 I(\chi_{q+2}^2(\Delta^2) \le (q-2)) \right] \right) \\
&\quad - \lambda_j^2 \delta_j^{*2} \{ 2E\left[(1 - (q-2)\chi_{q+2}^{-2}(\Delta^2)) I(\chi_{q+2}^2(\Delta^2) \le (q-2)) \right] \\
&\quad - E\left[(1 - (q-2)\chi_{q+4}^{-2}(\Delta^2))^2 I(\chi_{q+4}^2(\Delta^2) \le (q-2)) \right] \} \\
&\quad - 2k\alpha_j \lambda_j \delta_j^* E[(1 - (q-2)\chi_{q+2}^{-2}(\Delta^2)) I(\chi_{q+2}^2(\Delta^2) \le (q-2))].
\end{aligned} \tag{8.43}$$

Since $1-(q-2)\chi^{-2}_{q+2}(\Delta^2) \le 0 \Rightarrow \chi^2_{q+2}(\Delta^2) < q-2$ and the expectation of a negative r.v. is negative, hence, the R.H.S is non-negative for all $k \ge 0$ and the mse of $\hat{\boldsymbol{\rho}}^{s+}_n(k)$ is smaller than that of $\hat{\boldsymbol{\rho}}^{s}_n(k)$ uniformly in $k \ge 0$.

Further, we have the following corollary.

Corollary *A sufficient condition that the dominance relation is given by*

$$R_5(\hat{\boldsymbol{\rho}}^{s+}_n(k);\mathbf{I}_p) \le R_4(\hat{\boldsymbol{\rho}}^{s}_n(k);\mathbf{I}_p) \le R_1(\tilde{\boldsymbol{\rho}}_n(k);\mathbf{I}_p) \tag{8.44}$$

is that there exists a k such that $k \in (0, k^*_s)$ where k^*_s is given by (8.42).

8.6 Comparison of the Five RRE's as a Function of Δ^2

Consider a mse differences of $\tilde{\boldsymbol{\rho}}_n(k)$ and $\hat{\boldsymbol{\rho}}^{s}_n(k)$ and $\hat{\boldsymbol{\rho}}_n(k)$ and $\hat{\boldsymbol{\rho}}^{s+}_n(k)$ are given by

$$\begin{aligned}(a)\qquad & R_1(\tilde{\boldsymbol{\rho}}_n(k);\mathbf{I}_p) - R_4(\hat{\boldsymbol{\rho}}^{s}_n(k);\mathbf{I}_p)\\ &= \sigma^2_\varphi \gamma^{-2}(q-2)tr[R^2(k)A]\{(q-2)E[\chi^{-2}_{q+2}(\Delta^2)]\\ &\quad + [1-\frac{(q+2)\boldsymbol{\delta}' R^2(k)\boldsymbol{\delta}}{\sigma^2_\varphi\gamma^{-2}(2\Delta^2)tr[R^2(k)A]}](2\Delta^2)E[\chi^{-4}_{q+4}(\Delta^2)]\}\\ &\quad + 2k(q-2)[\boldsymbol{\delta}' R^{-1}(k)R(k)\boldsymbol{\rho}]E[\chi^{-2}_{q+2}(\Delta^2)] \ge 0\end{aligned}$$

uniformly in Δ^2 since $\frac{tr[R^2(k)C^{-1}]}{ch_{max}(R^2(k)C^{-1})} \ge \frac{q+2}{2}$.

Hence, $R_4(\hat{\boldsymbol{\rho}}^{s}_n(k);\mathbf{I}_p) \le R_1(\tilde{\boldsymbol{\rho}}_n(k);\mathbf{I}_p)$ uniformly in Δ^2 for fixed $k \in (0,\infty)$.

$$\begin{aligned}(b)\quad R_4\quad & (\hat{\boldsymbol{\rho}}^{s}_n(k);\mathbf{I}_p) - R_5(\hat{\boldsymbol{\rho}}^{s+}_n(k);\mathbf{I}_p) = \sigma^2_\varphi\gamma^{-2}(q-2)tr[R^2(k)A]\\ &\times\{E[(1-(q-2)\chi^{-2}_{q+2}(\Delta^2))^2 I(\chi^2_{q+2}(\Delta^2) \le (q-2))]\\ &+ \boldsymbol{\delta}' R^2(k)\boldsymbol{\delta}\{2E[(1-(q-2)\chi^{-2}_{q+2}(\Delta^2))I(\chi^2_{q+2}(\Delta^2) \le (q-2))]\\ &- E[(1-(q-2)\chi^{-2}_{q+4}(\Delta^2))^2 I(\chi^2_{q+4}(\Delta^2) \le (q-2))]\}\\ &+ 2k(q-2)\boldsymbol{\delta}' R^{-1}(k)R(k)\boldsymbol{\rho}E[\chi^{-2}_{q+2}(\Delta^2)] \ge 0\end{aligned}$$

uniformly in Δ^2.

Hence, $R_5(\hat{\boldsymbol{\rho}}^{s+}_n(k);\mathbf{I}_p) \le R_4(\hat{\boldsymbol{\rho}}^{s}_n(k);\mathbf{I}_p)$ uniformly in Δ^2 for fixed $k \in (0,\infty)$. Thus,

$$R_5(\hat{\boldsymbol{\rho}}^{s+}_n(k);\mathbf{I}_p) \le R_4(\hat{\boldsymbol{\rho}}^{s}_n(k);\mathbf{I}_p) \le R_1(\tilde{\boldsymbol{\rho}}_n(k);\mathbf{I}_p) \text{ for } k \in (0,k^*_s).$$

Next, we compare the amse of $\tilde{\boldsymbol{\rho}}_n(k)$ and $\hat{\boldsymbol{\rho}}^{PT}_n(k)$. Note that if $\alpha = 0$, then $\hat{\boldsymbol{\rho}}^{PT}_n(k) \equiv \hat{\boldsymbol{\rho}}_n(k)$. Thus, consider the amse-difference between $\tilde{\boldsymbol{\rho}}_n(k)$ and $\hat{\boldsymbol{\rho}}^{PT}_n(k)$ as follows:

$$R_1(\tilde{\boldsymbol{\rho}}_n(k);\mathbf{I}_p) - R_3(\hat{\boldsymbol{\rho}}^{PT}_n(k);\mathbf{I}_p) = \sigma^2_\varphi\gamma^{-2}tr[R^2(k)A]\mathcal{H}_{q+2}(\chi^2_q(\alpha);\Delta^2)$$

$$
\begin{aligned}
&- \boldsymbol{\delta}^{'} R^2(k)\boldsymbol{\delta}\{2\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2)\\
&- \mathcal{H}_{q+4}(\chi_q^2(\alpha);\Delta^2)\}\\
&- 2k\boldsymbol{\delta}R^{-1}(k)R(k)\boldsymbol{\rho}\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2)
\end{aligned}
$$

if $\alpha = 0$, the amse-difference becomes

$$\sigma_\varphi^2\gamma^{-2}tr[R^2(k)A] - \boldsymbol{\delta}^{'} R^2(k)\boldsymbol{\delta} - 2k\boldsymbol{\delta}R^{-1}(k)R(k)\boldsymbol{\rho}$$

Hence,$\hat{\boldsymbol{\rho}}_n^{PT}(k)$ is better than $\tilde{\boldsymbol{\rho}}_n(k)$ if and only if

$$\boldsymbol{\delta}^{'} R^2(k)\boldsymbol{\delta} \leq \frac{\sigma_\varphi^2\gamma^{-2}tr[R^2(k)A]\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2) - 2k\boldsymbol{\delta}^{'} R^{-1}(k)R(k)\boldsymbol{\rho}\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2)}{[2\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2) - \mathcal{H}_{q+4}(\chi_q^2(\alpha);\Delta^2)]}$$

This imply $\hat{\boldsymbol{\rho}}_n^{PT}(k)$ is superior to $\tilde{\boldsymbol{\rho}}_n(k)$ if and only if

$$\Delta^2 \leq \frac{\sigma_\varphi^2\gamma^{-2}tr[R^2(k)A]\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2) - 2k\boldsymbol{\delta}^{'} R^{-1}(k)R(k)\boldsymbol{\rho}\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2)}{ch_{max}[R^2(k)\boldsymbol{\Sigma}_{xx}^{-1}][2\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2) - \mathcal{H}_{q+4}(\chi_q^2(\alpha);\Delta^2)]}$$

Similarly,$\hat{\boldsymbol{\rho}}_n(k)$ is superior to $\tilde{\boldsymbol{\rho}}_n(k)$ if and only if

$$\Delta^2 \leq \frac{\sigma_\varphi^2\gamma^{-2}tr[R^2(k)A] - 2k\boldsymbol{\delta}R^{-1}(k)R(k)\boldsymbol{\rho}}{ch_{max}[R^2(k)\boldsymbol{\Sigma}_{xx}^{-1}]}.$$

Under $H_0 : \mathbf{H}\boldsymbol{\rho} = h$, the order the mse expressions is given by

$$R_2(\hat{\boldsymbol{\rho}}_n(k);\mathbf{I}_p) \leq R_3(\hat{\boldsymbol{\rho}}_n^{PT}(k);\mathbf{I}_p) \leq R_1(\tilde{\boldsymbol{\rho}}_n(k);\mathbf{I}_p).$$

When does $\tilde{\boldsymbol{\rho}}_n$ superior to $\hat{\boldsymbol{\rho}}_n^{PT}(k)$? Whenever

$$\Delta^2 \geq \frac{\sigma_\varphi^2\gamma^{-2}tr[R^2(k)A]\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2) - 2k\boldsymbol{\delta}^{'} R^{-1}(k)R(k)\boldsymbol{\rho}\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2)}{ch_{min}[R^2(k)\boldsymbol{\Sigma}_{xx}^{-1}][2\mathcal{H}_{q+2}(\chi_q^2(\alpha);\Delta^2) - \mathcal{H}_{q+4}(\chi_q^2(\alpha);\Delta^2)]}$$

If $k = 0$, the results coincide the five estimators in Sect. 8.3.

Similar, comments hold, tha is, $\tilde{\boldsymbol{\rho}}_n(k)$ is superior to $\hat{\boldsymbol{\rho}}_n(k)$ if and only if

$$\Delta^2 \geq \frac{\sigma_\varphi^2\gamma^{-2}tr[R^2(k)A] - 2k\boldsymbol{\delta}^{'} R^{-1}(k)R(k)\boldsymbol{\rho}}{ch_{min}[R^2(k)\boldsymbol{\Sigma}_{xx}^{-1}]}.$$

Thus, we obtain the same asymptotic properties of the R-estimators $\tilde{\boldsymbol{\rho}}_n$, $\hat{\boldsymbol{\rho}}_n^s$, $\hat{\boldsymbol{\rho}}_n^{s+}$ and $\hat{\boldsymbol{\rho}}_n^{PT}$ for the ridge estimators.

8.7 Summary and Conclusions

In this chapter, we defined a new class of R-estimators for the parameters $\boldsymbol{\rho}$ of the autoregressive model (8.1) by weighting the usual five R-estimators as in Saleh (2006) with a "ridge factor". We established that the asymptotic distributional properties of $\tilde{\boldsymbol{\rho}}_n(k)$, $\hat{\boldsymbol{\rho}}_n(k)$, $\hat{\boldsymbol{\rho}}_n^{PT}(k)$, $\hat{\boldsymbol{\rho}}_n^{s}(k)$ and $\hat{\boldsymbol{\rho}}_n^{s+}(k)$ are similar to the estimators $\tilde{\boldsymbol{\rho}}_n$, $\hat{\boldsymbol{\rho}}_n$, $\hat{\boldsymbol{\rho}}_n^{PT}$, $\hat{\boldsymbol{\rho}}_n^{s}$ and $\hat{\boldsymbol{\rho}}_n^{s+}$ based on the mse's as function of Δ^2 and k respectively. It is shown in particular that $\hat{\boldsymbol{\rho}}_n^{s+}(k)$ uniformly dominates $\hat{\boldsymbol{\rho}}_n^{s}(k)$ (when $p \geq 3$) and $\hat{\boldsymbol{\rho}}_n^{PT}(k)$ is a useful alternative to $\tilde{\boldsymbol{\rho}}_n(k)$ (when $p < 2$).

Acknowledgements The author thanks the referees for their careful reading of the paper. Also, thanks to Prof. H.L.Koul, a strong researcher for his collaborative research with me on autoregressive models which yielded several pioneering results on this topic.

References

Brockwell PJ, Davis RA (1987) Time Series. Springer, New York

Hall P, Hyde CC (1980) Martingle limit theory and its applications. Academic Press, New York

Hoerl AE, Kennard RW (1970) Ridge regression. Applications to nonorthogonal problems. Technometrics 12:55–67

Jaeckel, l. A. (1972) Estimating regression coefficients by minimizing the dispersion of residuals. Ann. Math. Stat 43:1449–1458

Koul HL. (1985) Minimum distance estimation in multiple linear regression. Sankhya ser A 47(1):57–74

Koul HL. (2002) Weighted Empirical process in Dynamic Nonlinear Models (2nd Ed). Springer

Koul HL, Saleh A K Md E (1995) R-estimation of the parameters of autoregression models. Ann Stat 21:534–551

Saleh AK Md Ehsanes (2006) Theory of preliminary test and stein-type estimation with applications. Wiley

Saleh AK Md E, Kibria BMG (2011) On some Ridge Regression Estimators: A Nonparametric Approach. Jour, Nonoparametric Statistics. 23(3):819–851

Sen PK, Saleh AK Md E (1987) On preliminary test and shrinkage M-estimation in Linear models. Ann. Statistics. 15(14):1580–1592

Chapter 9
On Hodges and Lehmann's "$6/\pi$ Result"

Marc Hallin, Yvik Swan and Thomas Verdebout

9.1 Introduction

The Pitman asymptotic relative efficiency $\mathrm{ARE}_f(\phi_1/\phi_2)$ under density f of a test ϕ_1 with respect to a test ϕ_2 is defined as the limit (when it exists), as n_1 tends to infinity, of the ratio $n_{2;f}(n_1)/n_1$ of the number $n_{2;f}(n_1)$ of observations it takes for the test ϕ_2, under density f, to match the local performance of the test ϕ_1 based on n_1 observations. That concept was first proposed by Pitman in the unpublished lecture notes (Pitman 1949) he prepared for a 1948–1949 course at Columbia University. The first published rigorous treatment of the subject was by Noether (1955). A similar definition applies to point estimation; see, for instance, Hallin (2012) for a more precise definition. An in-depth treatment of the concept can be found in Chap. 10 of Serfling (1980), Chap. 14 of van der Vaart (1998), or in the monograph by Nikitin (1995).

The study of the AREs of rank tests and R-estimators with respect to each other or with respect to their classical Gaussian counterparts has produced a number of interesting and sometimes surprising results. Considering the van der Waerden or normal-score two-sample location rank test ϕ_{vdW} and its classical normal-theory

M. Hallin (✉)
ECARES, Université libre de Bruxelles CP114/4, 50 Ave. F.D. Roosevelt B-1050, Bruxelles, Belgium
e-mail: mhallin@ulb.ac.be

ORFE, Princeton University, Sherrerd Hall, Princeton, NJ 08544, USA

Y. Swan
Faculté des Sciences, de la Technologie et de la Communication, Université de Luxembourg, Campus Kirchberg, Mathematics Research Unit, BLG, 6 rue Richard Coudenhove-Kalergi, Luxembourg 1359, Grand Duchy of Luxembourg
e-mail: yvik.swan@gmail.com

T. Verdebout
EQUIPPE and INRIA, Université Lille 3, Domaine universitaire du Pont de Bois, rue du Barreau BP 60149, Villeneuve-d'Ascq Cedex 59653, France
e-mail: thomas.verdebout@univ-lille3.fr

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_9,

competitor, the two-sample Student test $\phi_{\mathcal{N}}$, Chernoff and Savage in (1958) established the rather striking fact that, under any density f satisfying very mild regularity assumptions,

$$\mathrm{ARE}_f(\phi_{\mathrm{vdW}}/\phi_{\mathcal{N}}) \geq 1, \tag{9.1}$$

with equality holding at the Gaussian density $f = \phi$ only. That result implies that rank tests based on Gaussian scores (that is, the two-sample rank-based tests for location, but also the one-sample signed-rank ones, traditionally associated with the names of van der Waerden, Fraser, Fisher, Yates, Terry and/or Hoeffding—for simplicity, in the sequel, we uniformly call them *van der Waerden tests*)—asymptotically outperform the corresponding everyday practice Student t test; see Chernoff and Savage (1958). That result readily extends to one-sample symmetric and m-sample location, regression, and analysis of variance models with independent noise.

Another celebrated bound is the one obtained in 1956 by Hodges and Lehmann, who proved that, denoting by ϕ_{W} the Wilcoxon test (same location and regression problems as above),

$$\mathrm{ARE}_f(\phi_{\mathrm{W}}/\phi_{\mathcal{N}}) \geq 0.864, \tag{9.2}$$

which implies that the price to be paid for using rank-rank or signed-rank tests of the Wilcoxon type (that is, logistic-score-based rank tests) instead of the traditional Student ones never exceeds 13.6 % of the total number of observations. That bound moreover is sharp, being reached under the Epanechnikov density f. On the other hand, the benefits of considering Wilcoxon rather than Student can be arbitrarily large, as it is easily shown that the supremum over f of $\mathrm{ARE}_f(\phi_{\mathrm{W}}/\phi_{\mathcal{N}})$ is infinite; see Hodges and Lehmann (1956).

Both (9.1) and (9.2) created quite a surprise in the statistical community of the late 1950s, and helped dispelling the wrong idea, by then quite widespread, that rank-based methods, although convenient and robust, could not be expected to compete with the efficiency of traditional parametric procedures.

Chernoff–Savage and Hodges–Lehmann inequalities since then have been extended to a variety of more general settings. In the elliptical context, optimal rank-based procedures for location (one and m-sample case), regression, and scatter (one and m-sample cases) have been constructed in a series of papers by Hallin and Paindaveine (2002a, 2006, and 2008b), based on a multivariate concept of signed ranks. The Gaussian competitors there are of the Hotelling, Fisher, or Lagrange multiplier forms. For all those tests, Chernoff–Savage result, similar to (9.1) have been established (see also Paindaveine 2004, 2006). Hodges–Lehmann results also have been obtained, with bounds that, quite interestingly, depend on the dimension of the observation space: see Hallin and Paindaveine (2002a).

Another type of extension is into the direction of time series and linear rank statistics of the serial type. Hallin (1994) extended Chernoff and Savage's result (9.1) to the serial context by showing that the serial van der Waerden rank tests also uniformly dominate their Gaussian competitors (of the correlogram-based portmanteau, Durbin–Watson or Lagrange multiplier forms). Similarly, Hallin and Tribel

(2000) proved that the 0.864 upper bound in (9.2) no longer holds for the AREs of the Wilcoxon serial rank test with respect to their Gaussian competitors, and is to be replaced by a slightly lower 0.854 one. Elliptical versions of those results are derived in Hallin and Paindaveine (2002a, 2004, 2005).

Now, AREs with respect to Gaussian procedures such as t-tests are not always the best evaluations of the asymptotic performances of rank-based tests. Their existence indeed requires the Gaussian procedures to be valid under the density f under consideration, a condition which places restrictions on f that may not be satisfied. When the Gaussian tests are no longer valid, one rather may like to consider AREs of the form

$$\mathrm{ARE}_f(\phi_J/\phi_K) = 1/\mathrm{ARE}_f(\phi_K/\phi_J) \tag{9.3}$$

comparing the asymptotic performances (under f) of two rank-based tests ϕ_J and ϕ_K, based on score-generating functions J and K, respectively. Being distribution-free, rank-based procedures indeed do not impose any validity conditions on f, so that $\mathrm{ARE}_f(\phi_J/\phi_K)$ in general exists under much milder requirements on f; see, for instance, Hallin et al. (2011) and Hallin (2013), where AREs of the form (9.3) are provided for rank-based methods in linear models with stable errors under which Student tests are not valid.

Obtaining bounds for $\mathrm{ARE}_f(\phi_J/\phi_K)$, in general, is not as easy as for AREs of the form $\mathrm{ARE}_f(\phi_J/\phi_\mathcal{N})$. The first result of that type was established in 1961 by Hodges and Lehmann, who in (Hodges and Lehmann 1961) show that

$$0 \leq \mathrm{ARE}_f(\phi_\mathrm{W}/\phi_\mathrm{vdW}) \leq 6/\pi \approx 1.910 \tag{9.4}$$

or, equivalently,

$$0.524 \approx \pi/6 \leq \mathrm{ARE}_f(\phi_\mathrm{vdW}/\phi_\mathrm{W}) \leq \infty \tag{9.5}$$

for all f in some class $\mathcal{F}$ of density functions satisfying weak differentiability conditions. Hodges and Lehmann moreover exhibit a parametric family of densities $\mathcal{F}_\mathrm{HL} = \{f_\alpha | \alpha \in [0,\infty)\}$ for which the function $\alpha \mapsto \mathrm{ARE}_{f_\alpha}(\phi_\mathrm{W}/\phi_\mathrm{vdW})$ achieves any value in the open interval $(0, 6/\pi)$ ($\alpha \mapsto \mathrm{ARE}_{f_\alpha}(\phi_\mathrm{vdW}/\phi_\mathrm{W})$ achieves any value in the open interval $(\pi/6, \infty)$). The lower and upper bounds in (9.4) and (9.5) thus are *sharp* in the sense that they are the best possible ones. The same result was extended and generalized by Gastwirth (1970).

Note that, in case f has finite second-order moments (so that $\mathrm{ARE}_f(\phi_\mathrm{W}/\phi_\mathcal{N})$ is well defined), since $\mathrm{ARE}_f(\phi_\mathrm{vdW}/\phi_\mathcal{N}) = \mathrm{ARE}_f(\phi_\mathrm{vdW}/\phi_\mathrm{W}) \times \mathrm{ARE}_f(\phi_\mathrm{W}/\phi_\mathcal{N})$, Hodges and Lehmann's "6/π result" implies that the ARE of the van der Waerden tests with respect to the Student ones, which by the Chernoff–Savage inequality is larger than or equal to one, actually can be arbitrarily large, and that this happens for the same types of densities as for the Wilcoxon tests. This is an indication that, when Wilcoxon is quite significantly outperforming Student, that performance is shared by a broad class of rank-based tests and R-estimators, which includes the van der Waerden ones.

In Sect. 9.2, we successively consider the traditional case of *nonserial* rank statistics used in the context of location and regression models with independent observations, and the case of *serial* rank statistics; the latter involve ranks at time t and $t-k$, say, and aim at detecting serial dependence among the observations. Serial rank statistics typically involve two score functions and, instead of (9.3), yield AREs of the form

$$\mathrm{ARE}^{*}_{f}(\phi_{J_1,J_2}/\phi_{J_3,J_4}). \tag{9.6}$$

To start with, in Sect. 9.2.1, we revisit Gastwirth's classical nonserial results. More precisely, we provide (Proposition 2) a slightly different proof of the main proposition in Gastwirth (1970), with some further illustrations in the case of Student scores. In Sect. 9.2.2, we turn to the serial case, with special attention for the so-called Wilcoxon–Wald–Wolfowitz, Kendall, and van der Waerden rank autocorrelation coefficients. Serial AREs of the form (9.6) typically are the product of two factors to which the nonserial techniques of Sect. 9.2.1 separately apply; this provides bounds which, however, are not sharp. Therefore, in Sect. 9.3, we restrict to a few parametric families—the Student family (indexed by the degrees of freedom), the power-exponential family, or the Hodges–Lehmann family $\mathcal{F}_{\mathrm{HL}}$—for which numerical values are displayed.

9.2 Asymptotic Relative Efficiencies of Rank-Based Procedures

The asymptotic behavior of rank-based test statistics under local alternatives, since Hájek and Šidák (1967), is obtained via an application of Le Cam's Third Lemma (see, for instance, Chap. 13 of van der Vaart 1998). Whether the statistic is of the serial or the nonserial type, the result, under a density f with distribution function F involves integrals of the form

$$\mathcal{K}(J) := \int_0^1 J^2(u)\mathrm{d}u \qquad \mathcal{K}(J,f) := \int_0^1 J(u)\varphi_f(F^{-1}(u))\mathrm{d}u,$$

and, in the serial case,

$$\mathcal{J}(J,f) := \int_0^1 J(u)F^{-1}(u)\mathrm{d}u$$

where, assuming that f admits a weak derivative f', $\varphi_f := -f'/f$ is such that the Fisher information for location $\mathcal{I}(f) := \int_0^1 \varphi_f^2(F^{-1}(u))\mathrm{d}u$ is finite. Denote by $\mathcal{F}$ the class of such densities. If local alternatives, in the serial case, are of the ARMA type, f is further restricted to the subset $\mathcal{F}_2$ of densities $f \in \mathcal{F}$ having finite second-order moments. Differentiability in quadratic mean of $f^{1/2}$ is the standard assumption here, see Chap. 7 of van der Vaart (1998); but absolute continuity of f in the traditional sense, with a.e. derivative f', is sufficient for most purposes. We refer to Hájek and Šidák (1967) and Hallin and Puri (1994) for details in the nonserial and the serial case, respectively.

9.2.1 The Nonserial Case

In location or regression problems, or, more generally, when testing linear constraints on the parameters of a linear model (this includes ANOVA etc.), the ARE, under density $f \in \mathcal{F}$, of a rank-based test ϕ_{J_1} based on the square-summable score-generating function J_1 with respect to another rank-based test ϕ_{J_2} based on the square-summable score-generating function J_2 takes the form

$$\mathrm{ARE}_f\left(\phi_{J_1}/\phi_{J_2}\right) = \frac{\mathcal{K}(J_2)}{\mathcal{K}(J_1)} C_f^2(J_1, J_2), \quad \text{with} \quad C_f(J_1, J_2) := \frac{\mathcal{K}(J_1, f)}{\mathcal{K}(J_2, f)}, \tag{9.7}$$

provided that J_1 and J_2 are monotone, or the difference between two monotone functions. Those ARE values readily extend to the m-sample setting, and to R-estimation problems. In a time-series context with innovation density $f \in \mathcal{F}_2$, and under slightly more restrictive assumptions on the scores, they also extend to the partly rank-based tests and R-estimators considered by Koul and Saleh in (1993) and (1995).

Gastwirth (1970) has based his analysis of (9.7) on an integration by parts of the integral in the definition of $\mathcal{K}(J, f)$. If both J_1 and J_2 are differentiable, with derivatives J_1' and J_2', respectively, and provided that f is such that

$$\lim_{x\to\infty} J_1(F(x))f(x) = 0 = \lim_{x\to\infty} J_2(F(x))f(x),$$

integration by parts in those integrals yields, for (9.7),

$$\mathrm{ARE}_f\left(\phi_{J_1}/\phi_{J_2}\right) = \frac{\mathcal{K}(J_2)}{\mathcal{K}(J_1)} \left(\frac{\int_{-\infty}^{\infty} J_1'(F(x)) f^2(x)\mathrm{d}x}{\int_{-\infty}^{\infty} J_2'(F(x)) f^2(x)\mathrm{d}x} \right)^2. \tag{9.8}$$

In view of the Chernoff–Savage result (9.1), the van der Waerden score-generating function

$$J_2(u) = J_{\mathrm{vdW}}(u) = \Phi^{-1}(u) \tag{9.9}$$

(with $u \mapsto \Phi^{-1}(u)$ the standard normal quantile function) may appear as a natural benchmark for ARE computations. From a technical point of view, under this integration by parts approach, the Wilcoxon score-generating function

$$J_2(u) = J_{\mathrm{W}}(u) = u - 1/2 \tag{9.10}$$

(the Spearman–Wald–Wolfowitz score-generating function in the serial case) is more appropriate, though. Convexity arguments indeed will play an important role, and, being linear, J_{W} is both convex and concave. Since $J_{\mathrm{W}}'(u) = 1$ and $\mathcal{K}(J_{\mathrm{W}}) = 1/12$, Eq. (9.8) yields

$$12\,\mathrm{ARE}_f\left(\phi_{J_1}/\phi_{\mathrm{W}}\right) = \frac{1}{\mathcal{K}(J_1)} \left(\frac{\int_{-\infty}^{\infty} J_1'(F(x)) f^2(x)\mathrm{d}x}{\int_{-\infty}^{\infty} f^2(x)\mathrm{d}x} \right)^2. \tag{9.11}$$

Bounds on $J_1'(F(x))$ then readily yield bounds on AREs, irrespective of f.

That property of Wilcoxon scores is exploited in Propositions 2 and 3 for nonserial AREs, in Proposition 4 for the serial ones; those bounds are mainly about AREs of, or with respect to, Wilcoxon (Spearman–Wald–Wolfowitz) procedures, but not exclusively so.

Assume that $f \in \mathcal{F}_0 := \{f \in \mathcal{F} | \lim_{x\to\pm\infty} f(x) = 0\}$. Then, integration by parts is possible in the definition of $\mathcal{K}(J_{\mathrm{W}}, f)$, yielding

$$\mathcal{K}(J_{\mathrm{W}}, f) = \int_{-\infty}^{\infty} f^2(x)\mathrm{d}x.$$

Assume, furthermore, that the square-integrable score-generating function J_1 (the difference of two monotone increasing functions) is differentiable, with derivative J_1', and that

$$f \in \mathcal{F}_{J_1} := \{f \in \mathcal{F}_0 | \lim_{x\to\pm\infty} J_1(F(x))f(x) = 0\},$$

so that (9.8) holds. Finally, assume that J_1 is skew-symmetric about $1/2$. Defining the (possibly infinite) constants

$$\kappa_J^+ := \sup_{u\geq 1/2} \left|J'(u)\right| \quad \text{and} \quad \kappa_{\mathrm{J}}^- := \inf_{\mathrm{u}\geq 1/2} \left|J'(\mathrm{u})\right|,$$

we can always write

$$12\,\mathrm{ARE}_f\left(\phi_{J_1}/\phi_{\mathrm{W}}\right) \leq (\kappa_{J_1}^+)^2/\mathcal{K}(J_1) \tag{9.12}$$

while, if J_1 is non-decreasing (hence J_1' is non-negative), we further have

$$(\kappa_{J_1}^-)^2/\mathcal{K}(J_1) \leq 12\,\mathrm{ARE}_f\left(\phi_{J_1}/\phi_{\mathrm{W}}\right) \leq (\kappa_{J_1}^+)^2/\mathcal{K}(J_1). \tag{9.13}$$

The quantities appearing in (9.12) and (9.13) often can be computed explicitly, yielding ARE bounds which are, moreover, sharp under certain conditions.

For example, if J_1 is convex on $[1/2, 1)$, its derivative J_1' is non-decreasing over $[1/2, 1)$, so that

$$\kappa_{J_1}^- = J_1'(1/2) \geq 0 \quad \text{and} \quad \kappa_{\mathrm{J}_1}^+ = \lim_{\mathrm{u}\to 1} J_1'(\mathrm{u}) \leq +\infty. \tag{9.14}$$

It follows that, under the assumptions made,

$$(J_1'(1/2))^2/\mathcal{K}(J_1) \leq 12\,\mathrm{ARE}_f\left(\phi_{J_1}/\phi_{\mathrm{W}}\right) \leq (\lim_{u\to 1} J_1'(u))^2/\mathcal{K}(J_1). \tag{9.15}$$

The lower bound in (9.15) is established in Theorem 2.1 of Gastwirth (1970).

The double inequality (9.15) holds, for instance (still, under $f \in \mathcal{F}_{J_1}$), when the scores $J_1 = \varphi_g \circ G^{-1}$ are the optimal scores associated with some symmetric and *strongly unimodal* density g with distribution function G; such densities indeed are log-concave and have monotone increasing, convex over $[1/2, 1)$ score functions. Symmetric log-concave densities take the form

$$g(x) = Ke^{-\mu(x)}, \qquad K^{-1} = \int_{-\infty}^{\infty} e^{-\mu(x)}\mathrm{d}x \tag{9.16}$$

with $x \mapsto \mu(x)$ a convex, even (that is, $\mu(x) = \mu(-x)$) function; assume it to be twice differentiable, with derivatives μ' and μ''. Then, $\varphi_g(x) = \mu'(x)$, so that

$$J_1(u) := \varphi_g(G^{-1}(u)) = \mu'(G^{-1}(u)), \qquad \mathcal{K}(J_1) = \int_{-\infty}^{\infty} (\mu'(x))^2 \, g(x) \mathrm{d}x = \mathcal{I}(g)$$

where $\mathcal{I}(g)$ the Fisher information of g (which we assume to be finite), and

$$J_1'(u) = \mu''(G^{-1}(u))/g(G^{-1}(u)), \quad \text{hence} \quad J_1'(1/2) = \frac{\mu''(0)}{g(0)} = \frac{\mu''(0)}{K}.$$

Specializing (9.15) to this situation, we obtain the following proposition.

Proposition 1. *If the square-integrable score-generating function J_1 is of the form $\varphi_g \circ G^{-1}$ with g given by (9.16), μ even, convex, and twice differentiable, then, under any $f \in \mathcal{F}_{J_1}$,*

$$\left(\frac{\mu''(0)}{K}\right)^2 \le 12\,\mathcal{I}(g) A R E_f(\phi_{J_1}/\phi_{\mathrm{W}}) \le (\lim_{u \to 1} J_1'(u))^2 = (\lim_{x \to \infty} (\mu''(x)/g(x))^2. \tag{9.17}$$

With $\mu(x) = x^2/2$ (so that $K^{-1} = \sqrt{2\pi}$) in (9.16), g is the standard Gaussian density; $\mu''(0) = 1, \mathcal{I}(g) = 1$, and the lower bound in (9.17) becomes $(\mu''(0)/K)^2 = 2\pi$, whereas the upper bound is trivially infinite. This yields the Hodges–Lehmann result (9.4).

Turning back to (9.12) and (9.13), but with J_1 concave (and still nondecreasing) on $[1/2, 1)$, J_1' is nonincreasing, so that $\kappa_{J_1}^+ = J_1'(1/2)$ and

$$12\,\mathrm{ARE}_f\left(\phi_{J_1}/\phi_{\mathrm{W}}\right) \le (J_1'(1/2))^2/\mathcal{K}(J_1). \tag{9.18}$$

Not much can be said on the lower bound, though, without further assumptions on the behavior of J_1 around $u = 1$.

Replacing, for various score-generating functions J_1 and densities f, the quantities appearing in (9.12), (9.15) or (9.18) with their explicit values provides a variety of bounds of the Hodges–Lehmann type. Below, we consider the van der Waerden tests ϕ_{vdW}, based on the score-generating function (9.9) and the Cauchy-score rank tests ϕ_{Cauchy}, based on the score-generating function

$$J_{\mathrm{Cauchy}}(u) = \sin(2\pi(u - 1/2)). \tag{9.19}$$

Proposition 2. *For all symmetric densities f in $\mathcal{F}_{\mathrm{vdW}}$, $\mathcal{F}_{\mathrm{Cauchy}}$ and $\mathcal{F}_{\mathrm{vdW}} \bigcap \mathcal{F}_{\mathrm{Cauchy}}$, respectively,*

(1) $\mathrm{ARE}_f(\phi_{\mathrm{W}}/\phi_{\mathrm{vdW}}) \le 6/\pi$;
(2) $\mathrm{ARE}_f(\phi_{\mathrm{Cauchy}}/\phi_{\mathrm{W}}) \le 2\pi^2/3$;
(3) $\mathrm{ARE}_f(\phi_{\mathrm{Cauchy}}/\phi_{\mathrm{vdW}}) \le 4\pi$.

Proof. The van der Waerden score (9.9) is strictly increasing, and convex over$[1/2, 1)$. One readily obtains

$$\mathcal{K}(J_{\text{vdW}}) = 1 \quad \text{and} \quad \mathrm{J}'_{\text{vdW}}(\mathrm{u}) = \sqrt{2\pi}\exp\{(\Phi^{-1}(\mathrm{u}))^2/2\},$$

hence $\kappa^-_{\text{vdW}} = J'_{\text{vdW}}(1/2) = \sqrt{2\pi}$. Plugging this into the left-hand side inequality of (9.15) yields (1). Alternatively one can directly apply (9.17).

The Cauchy score is concave over $[1/2, 1)$, but not monotone (being of bounded variation, however, it is the difference of two monotone function). Direct inspection of (9.19) nevertheless reveals that

$$\mathcal{K}(J_{\text{Cauchy}}) = 1/2 \quad \text{and} \quad \mathrm{J}'_{\text{Cauchy}}(\mathrm{u}) = 2\pi\cos(2\pi(\mathrm{u} - 1/2)),$$

hence $\kappa^+_{\text{Cauchy}} = J'_{\text{Cauchy}}(1/2) = 2\pi$. Substituting this in (9.12) yields (2). The product of the upper bounds in (1) and (2) yields (3). □

Remarkably, those three bounds are sharp. Indeed, numerical evaluation shows that they can be approached arbitrarily well by taking extremely heavy-tails such as those of stable densities f_α with tail index $\alpha \to 0$, Student densities with degrees of freedom $\nu \to 0$, or Pareto densities with $\alpha \to 0$; see also the family $\mathcal{F}_{\text{HL}}$ of densities $f_{a,\epsilon}(x)$ defined in Eq. (9.24).

Figure 9.1 provides plots of $\text{ARE}_f(\phi_{\text{W}}/\phi_{\text{vdW}})$ and $\text{ARE}_f(\phi_{\text{Cauchy}}/\phi_{\text{vdW}})$ for various densities. Inspection of those graphs shows that both AREs are decreasing as the tails become lighter; the sharpness of bounds (1) and (3), hence also that of bound (2), is graphically confirmed.

The bounds proposed in Proposition 2 are not new, and have been obtained already in Gastwirth (1970). One would like to see similar bounds for other score functions, such as the Student ones

$$\begin{aligned} J_{t_\nu}(u) &= (\nu+1)F^{-1}_{t_\nu}(u)/(\nu + F^{-1}_{t_\nu}(u)^2) \qquad 0 < u < 1 \\ &= \frac{1+\nu}{\sqrt{\nu}}\sqrt{-1 + \frac{1}{\text{IB}_\nu(1-2u)}}\,\text{IB}_\nu(1-2u) \quad 1/2 \le u < 1 \end{aligned} \tag{9.20}$$

where $\text{IB}_\nu(\nu)$ denotes the inverse of the regularized incomplete beta function evaluated at $(1, \nu, \nu/2, 1/2)$ and $F^{-1}_{t_\nu}$ stands for the Student quantile function with ν degrees of freedom. Note that $\lim_{\nu\to-1}\text{IB}_\nu(\nu) = 0$, so that $\lim_{u\to1} J_{t_\nu}(u) = 0$. Since $J_{t_\nu}(1/2) = 0$ and $J'_{t_\nu}(1/2) > 0$, this means that, on $[1/2, 1)$, J_{t_ν} is a redescending function; in general, it is neither convex nor concave on $[1/2, 1)$.

Differentiating (9.20), we get, for $u \ge 1/2$,

$$J'_{t_\nu}(u) = \frac{\sqrt{\pi}(\nu+1)\Gamma\left(\frac{\nu}{2}\right)}{\sqrt{\nu}\Gamma\left(\frac{\nu+1}{2}\right)}\left(-1 + 2\text{IB}_\nu(1-2u)\right)\text{IB}_\nu(1-2u)^{\frac{1-\nu}{2}}, \tag{9.21}$$

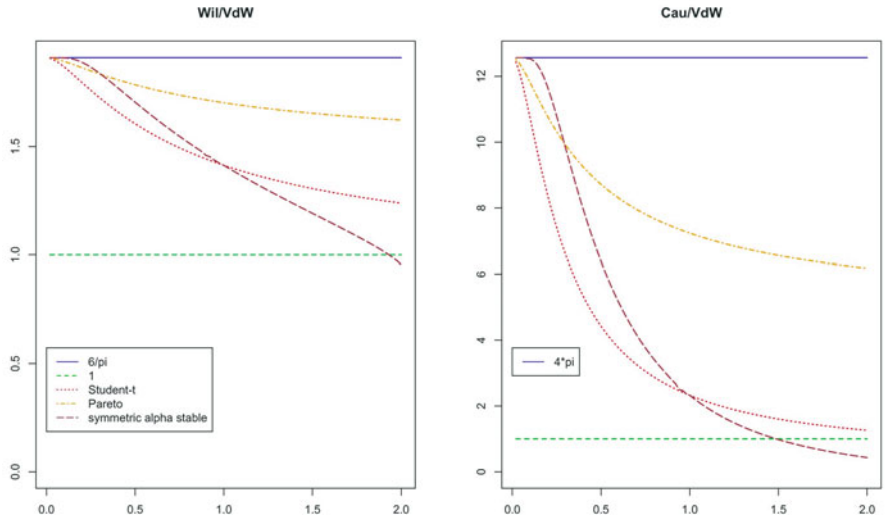


Fig. 9.1 $\mathrm{ARE}_f(\phi_{\mathrm{W}}/\phi_{\mathrm{vdW}})$ and $\mathrm{ARE}_f(\phi_{\mathrm{Cauchy}}/\phi_{\mathrm{vdW}})$ under various families of densities: symmetric stable (indexed by their tail parameter α), Student-t (indexed by their degrees of freedom ν) or Pareto (indexed by their shape parameter α)

from which we deduce that

$$\lim_{u\to 1} J'_{t_\nu}(u) = \begin{cases} 0 & 0 < \nu < 1 \\ -2\pi & \nu = 1 \\ -\infty & 1 < \nu \end{cases}.$$

Except for the $\nu = 1$ case, which is covered by (2) and (3) in Proposition 2, these values do not provide exploitable values for κ^+. For $\nu < 1$, however, one can check from (9.21) that $\max_{u\geq 1/2}|J'(x)| = J'(1/2)$, so that

$$\kappa^+_{J_{t_\nu}} = -\sqrt{\pi}(\nu+1)\Gamma\left(\frac{\nu}{2}\right)/\sqrt{\nu}\,\Gamma\left(\frac{\nu+1}{2}\right).$$

Elementary, though somewhat tedious, algebra yields

$$\mathcal{K}(J_{t_\nu}) = (\nu+1)/(\nu+3).$$

Plugging this into (9.12), we obtain, for $\nu \leq 1$, the following additional bounds.

Proposition 3. *For all* $0 < \nu \leq 1$ *and all symmetric density* f *in* $\mathcal{F}_{J_{t_\nu}}$ *and* $\mathcal{F}_{J_{t_\nu}} \bigcap \mathcal{F}_{J_{\mathrm{vdW}}}$, *respectively,*

(4) $\mathrm{ARE}_f(\phi_{t_\nu}/\phi_{\mathrm{W}}) \leq \pi\Gamma^2(\frac{\nu}{2})(\nu+3)(\nu+1)/12\nu\Gamma^2(\frac{\nu+1}{2})$, and
(5) $\mathrm{ARE}_f(\phi_{t_\nu}/\phi_{\mathrm{vdW}}) \leq \Gamma^2(\frac{\nu}{2})(\nu+3)(\nu+1)/2\nu\Gamma^2(\frac{\nu+1}{2})$.

Inequality (4) is sharp, the bound being achieved, in the limit, under very heavy tails (stable densities with $\alpha \downarrow 0$, or Student-t_μ densities with $\mu \downarrow 0$). Since this is also

the case, under the same sequences of densities, for inequality (1) in Proposition 2, inequality (5) is sharp as well. The upper bounds (4) and (5) are both decreasing functions of the tail index ν; both are unbounded at the origin, and both converge to the corresponding Cauchy values as $\nu \to 1$.

9.2.2 The Serial Case

Until the early 1980s, and despite some forerunning time-series applications such as Wald and Wolfowitz (1943) (published as early as 1943—two years before Frank Wilcoxon's pathbreaking 1945 paper), rank-based methods had been essentially limited to statistical models involving univariate independent observations. Therefore, the traditional ARE bounds (Hodges and Lehmann 1956, 1961), Chernoff–Savage (1958) or Gastwirth (1970), as well as the classical monographs (Hájek and Šidáak 1967; Randles and Wolfe 1979; Puri and Sen 1985, to quote only a few) mainly deal with univariate location and single-output linear (regression) models with independent observations. The situation since then has changed, and rank-based procedures nowadays have been proposed for a much broader class of statistical models, including time-series problems, where serial dependencies are the main features under study.

In this section, we focus on the linear rank statistics of the serial type involving two square-integrable score functions. Those statistics enjoy optimality properties in the context of linear time series (ARMA models; see Hallin and Puri 1994 for details). Once adequately standardized, those statistics yield the so-called *rank-based autocorrelation coefficients* that are denoted by ${R^{(n)}}_1, \ldots, {R^{(n)}}_n$, the ranks in a triangular array ${X^{(n)}}_1, \ldots, {X^{(n)}}_n$ of observations. *Rank autocorrelations* (with lag k) are linear serial rank statistics of the form

$$\underset{\sim}{tr}^{(n)}_{J_1 J_2;k} := \big[(n-k)^{-1} \sum_{t=k+1}^{n} J_1\big(\frac{R_t^{(n)}}{n+1}\big) J_2\big(\frac{R_{t-k}^{(n)}}{n+1}\big) - m^{(n)}_{J_1 J_2}\big](s^{(n)}_{J_1 J_2})^{-1},$$

where J_1 and J_2 are (square-integrable) score-generating functions, whereas $m^{(n)}_{J_1 J_2}$ and $s^{(n)}_{J_1 J_2} := s^{(n)}_{J_1 J_2;k}$ denote the exact mean of $J_1\big(\frac{R_t^{(n)}}{n+1}\big) J_2\big(\frac{R_{t-k}^{(n)}}{n+1}\big)$ and the exact standard error of $(n-k)^{-\frac{1}{2}} \sum_{t=k+1}^{n} J_1\big(\frac{R_t^{(n)}}{n+1}\big) J_2\big(\frac{R_{t-k}^{(n)}}{n+1}\big)$ under the assumption of i.i.d. $X_t^{(n)}$'s (more precisely, exchangeable $R_t^{(n)}$'s), respectively; we refer to pages 186 and 187 of Hallin and Puri (1994) for explicit formulas. *Signed-rank autocorrelation coefficients* are defined similarly; see Hallin and Puri (1992) or Hallin and Puri (1994).

Rank and signed-rank autocorrelations are measures of serial dependence offering rank-based alternatives to the usual autocorrelation coefficients, of the form

$$r_k^{(n)} := \sum_{t=k+1}^{n} X_t X_{t-k} \Big/ \sum_{t=1}^{n} X_t^2,$$

which consitute the Gaussian reference benchmark in this context. Of particular interest are

(i) the *van der Waerden autocorrelations* (Hallin and Puri 1988)

$$r^{(n)}_{\sim \mathrm{vdW};k} := \Big[(n-k)^{-1} \sum_{t=k+1}^{n} \Phi^{-1}\Big(\frac{R^{(n)}_t}{n+1}\Big)\Phi^{-1}\Big(\frac{R^{(n)}_{t-k}}{n+1}\Big) - m^{(n)}_{\mathrm{vdW}}\Big](s^{(n)}_{\mathrm{vdW}})^{-1},$$

(ii) the *Wald-Wolfowitz* or *Spearman autocorrelations* (Wald and Wolfowitz 1943)

$$r^{(n)}_{\sim \mathrm{SWW};k} := \Big[(n-k)^{-1} \sum_{t=k+1}^{n} R^{(n)}_t R^{(n)}_{t-k} - m^{(n)}_{\mathrm{SWW}}\Big](s^{(n)}_{\mathrm{SWW}})^{-1},$$

(iii) and the *Kendall autocorrelations* (Ferguson et al. 2000, where explicit values of $m^{(n)}_{\mathrm{K}}$ and $s^{(n)}_{\mathrm{K}}$ are provided)

$$r^{(n)}_{\sim \mathrm{K};k} := \Big[1 - \frac{4D^{(n)}_k}{(n-k)(n-k-1)} - m^{(n)}_{\mathrm{K}}\Big](s^{(n)}_{\mathrm{K}})^{-1}$$

with $D^{(n)}_k$ denoting the number of discordances at lag k, that is, the number of pairs $(R^{(n)}_t, R^{(n)}_{t-k})$ and $(R^{(n)}_s, R^{(n)}_{s-k})$ that satisfy either

$$R^{(n)}_t < R^{(n)}_s \quad \text{and} \quad R^{(n)}_{t-k} > R^{(n)}_{s-k}, \quad \text{or} \quad R^{(n)}_t > R^{(n)}_s \quad \text{and} \quad R^{(n)}_{t-k} < R^{(n)}_{s-k};$$

more specifically, $D^{(n)}_k := \sum_{t=k+1}^{n} \sum_{s=t+1}^{n} I(R^{(n)}_t < R^{(n)}_s,\ R^{(n)}_{t-k} > R^{(n)}_{s-k})$.

The van der Waerden autocorrelations are optimal—in the sense that they allow for *locally optimal* rank tests in the case of ARMA models with normal innovation densities. The Spearman and Kendall autocorrelations are serial versions of Spearman's *rho* and Kendall's *tau*, respectively, and are asymptotically equivalent under the null hypothesis of independence; although they are never optimal for any ARMA alternative, they achieve excellent overall performance. Signed rank autocorrelations are defined in a similar way.

Let J_i, $i = 1, \ldots, 4$ denote four square-summable score functions, and assume that they are monotone increasing, or the difference between two monotone increasing functions (that assumption tacitly will be made in the sequel each time AREs are to be computed). Recall that $\mathcal{F}_2$ denotes the subclass of densities $f \in \mathcal{F}$ having finite moments of order two. The asymptotic relative efficiency, under innovation density $f \in \mathcal{F}_2$, of the rank-based tests $\phi^r_{J_1 J_2}$ based on the autocorrelations $r^{(n)}_{\sim J_1 J_2;k}$ with respect to the rank-based tests $\phi^r_{J_3 J_4}$ based on the autocorrelations $r^{(n)}_{\sim J_3 J_4;k}$ is

$$\begin{aligned}
&\mathrm{ARE}^*_f(\phi^r_{J_1 J_2}/\phi^r_{J_3 J_4}) \\
&= \frac{\mathcal{K}(J_3)}{\mathcal{K}(J_1)} \left(\frac{\int_0^1 J_1(v)\varphi_f(F^{-1}(v))\mathrm{d}v}{\int_0^1 J_3(v)\varphi_f(F^{-1}(v))\mathrm{d}v}\right)^2 \frac{\mathcal{K}(J_4)}{\mathcal{K}(J_2)} \left(\frac{\int_0^1 J_2(v)F^{-1}(v)\mathrm{d}v}{\int_0^1 J_4(v)F^{-1}(v)\mathrm{d}v}\right)^2 \\
&= \frac{\mathcal{K}(J_3)}{\mathcal{K}(J_1)}\, C^2_f(J_1, J_3) \frac{\mathcal{K}(J_4)}{\mathcal{K}(J_2)}\, D^2_f(J_2, J_4) \qquad (9.22)
\end{aligned}$$

with $C_f(J_1, J_3) := \mathcal{K}(J_1, f)/\mathcal{K}(J_3, f)$ and $D_f(J_2, J_4) := \mathcal{J}(J_2, f)/\mathcal{J}(J_4, f)$.

The C_f ratios have been studied in Sect. 9.2.1, and the same conclusions apply here; as for the D_f ratios, they can be treated by similar methods.

Denote by ϕ^r_{vdW}, ϕ^r_{SWW}, ... the tests based on $\underset{\sim}{\mathcal{L}}^{(n)}_{\mathrm{vdW};k}$, $\underset{\sim}{\mathcal{L}}^{(n)}_{\mathrm{SWW};k}$, etc. The serial counterpart of $\mathrm{ARE}_f(\phi_{\mathrm{W}}/\phi_{J_1})$ is $\mathrm{ARE}^*_f(\phi^r_{\mathrm{SWW}}/\phi^r_{J_1J_2})$, for which the following result holds.

Proposition 4. *Let the score functions J_1 and J_2 be monotone increasing, skew-symmetric about $1/2$, and differentiable, with strictly positive $J_1'(1/2)$ and $J_2'(1/2)$. Suppose that $f \in \mathcal{F}_2 \bigcap \mathcal{F}_{J_1} \bigcap \mathcal{F}_{J_2}$ is a symmetric probability density function. Then,*

(1) if J_1 and J_2 are convex on $[1/2, 1)$,

$$\mathrm{ARE}^*_f(\phi^r_{\mathrm{SWW}}/\phi^r_{J_1J_2}) = \mathrm{ARE}^*_f(\phi^r_{\mathrm{K}}/\phi^r_{J_1J_2}) \leq 144\frac{\mathcal{K}(J_1)\mathcal{K}(J_2)}{(J_1'(1/2)\ J_2'(1/2))^2};$$

(2) if J_1 and J_2 are concave on $[1/2, 1)$,

$$\mathrm{ARE}^*_f(\phi^r_{J_1J_2}/\phi^r_{\mathrm{SWW}}) = \mathrm{ARE}^*_f(\phi^r_{J_1J_2}/\phi^r_{\mathrm{K}}) \leq \frac{1}{144}\frac{(J_1'(1/2)\ J_2'(1/2))^2}{\mathcal{K}(J_1)\mathcal{K}(J_2)}.$$

Proof. In view of (9.7), we have

$$\mathrm{ARE}^*_f(\phi^r_{\mathrm{SWW}}/\phi^r_{J_1J_2}) = \mathrm{ARE}_f(\phi_{\mathrm{W}}/\phi_{J_1})\frac{\mathcal{K}(J_2)}{\mathcal{K}(J_W)}\left(\frac{\int_0^1 (v-1/2)F^{-1}(v)\mathrm{d}v}{\int_0^1 J_2(v)F^{-1}(v)\mathrm{d}v}\right)^2.$$

Consider part (1) of the proposition. It follows from (9.13) that

$$\mathrm{ARE}_f(\phi_{\mathrm{W}}/\phi_{J_1}) \leq 12\,\mathcal{K}(J_1)/(J_1'(1/2))^2.$$

Since J_2 is convex over $[1/2, 1)$, $J_2(u) \geq J_2'(1/2)(u-1/2)$ for all $u \in [1/2, 1)$, so that

$$\int_0^1 J_2(v)F^{-1}(v)\mathrm{d}v = 2\int_{1/2}^1 J_2(v)F^{-1}(v)\mathrm{d}v \geq J_2'(1/2)\int_{1/2}^1 (v-1/2)F^{-1}(v)\mathrm{d}v.$$

It follows that

$$\frac{\mathcal{K}(J_2)}{\mathcal{K}(J_W)}\left(\frac{\int_0^1 (v-1/2)F^{-1}(v)\mathrm{d}v}{\int_0^1 J_2(v)F^{-1}(v)\mathrm{d}v}\right)^2 \leq \frac{12\,\mathcal{K}(J_2)}{(J_2'(1/2))^2},$$

where the assumption of finite variance is used. Part (1) of the result follows. A similar argument holds (with reversed inequalities) if J_2 is concave, yielding part (2).

Applying this result to the score functions $J_1(u) = J_2(u) = \Phi^{-1}(u)$ (convex over $[1/2, 0)$) for which $J_1'(1/2) = J_2'(1/2) = \sqrt{2\pi}$ and $\mathcal{K}(J_1) = \mathcal{K}(J_2) = 1$, we readily obtain the following serial extension of Hodges and Lehmann's "$6/\pi$ result":

$$\mathrm{ARE}^*_f(\phi^r_{\mathrm{SWW}}/\phi^r_{\mathrm{vdW}}) = \mathrm{ARE}^*_f(\phi^r_{\mathrm{K}}/\phi^r_{\mathrm{vdW}}) \leq (6/\pi)^2. \tag{9.23}$$

Table 9.1 Numerical values of C_f, D_f, $ARE_f = ARE_f(\phi_W/\phi_{vdW})$, and $ARE_f^* = ARE_f^*(\phi_{SWW}^r/\phi_{vdW}^r)$ under densities $f_{a,\epsilon}$ in the Hodges–Lehmann family $\mathcal{F}_{HL}$ (see (9.24)), for various values of ε and $a \to 0$

ε	C_f	D_f	ARE_f	ARE_f^*
0	0.398942	0.282070	1.90986	1.82346
0.2	0.396313	0.276619	1.88476	1.73062
0.4	0.388772	0.271848	1.81372	1.60844
0.6	0.377291	0.271061	1.70818	1.50608
1	0.348213	0.287973	1.45503	1.44796
2	0.294160	0.303085	1.03836	1.14461
3	0.282852	0.285646	0.960064	0.940023
10	0.282095	0.282095	0.954930	0.911891
100	0.282095	0.282095	0.954930	0.911891

An important difference, though, is that the bound in (9.23) is unlikely to be sharp. Section 9.3 provides some numerical evidence of that fact, which is hardly surprising; while the ratio $C_f(J_{vdW}, J_W)$ is maximized for densities putting all their weight about the origin, this no longer holds true for $D_f(J_{vdW}, J_W)$. In particular, the sequences of densities considered in Hodges and Lehmann (1961) or Gastwirth (1970) along which $C_f(J_{vdW}, J_W)$ tends to its upper bound typically are not the same as those along which $D_f(J_{vdW}, J_W)$ does.

9.3 Some Numerical Results

In this final section, we provide numerical values of $ARE_f(\phi_W/\phi_{vdW})$ (denoted as ARE_f in the sequel) and $ARE_f^*(\phi_{SWW}^r/\phi_{vdW}^r)$ (denoted as ARE_f^* in the sequel) under various families of distributions.

First, let us give some ARE values under Gaussian densities: if $f = \phi$, we obtain

$$C_\phi(J_W, J_{vdW}) = D_\phi(J_W, J_{vdW}) = \frac{1}{2\sqrt{\pi}} \approx 0.28209$$

so that

$$ARE_\phi(\phi_W/\phi_{vdW}) = \frac{3}{\pi} \approx 0.95493$$

and

$$ARE_\phi^*(\phi_{SWW}^r/\phi_{vdW}^r) = \frac{9}{\pi^2} \approx 0.91189.$$

Tables 9.1, 9.2, and 9.3 provide numerical values of ARE_f and ARE_f^* under

(1) (Table 9.1) The two-parameter family $\mathcal{F}_{HL}$ of densities $f_{a,\epsilon}$ associated with the distribution functions

$$F_{a,\epsilon}(x) = \begin{cases} \Phi(x) & \text{if } 0 \le x \le \epsilon \\ \Phi(\epsilon + a(x - \epsilon)) & \text{if } \epsilon < x \end{cases} \tag{9.24}$$

Table 9.2 Numerical values of C_f, D_f, $ARE_f = ARE_f(\phi_W/\phi_{vdW})$, and $ARE_f^* = ARE_f^*(\phi_{SWW}^r/\phi_{vdW}^r)$ under Student-t densities with various degrees of freedom ν

ν	C_f	D_f	ARE_f	ARE_f^*
0.1	0.394451	–	1.86710	–
1	0.343120	–	1.41277	–
2	0.321212	0.243196	1.23813	0.878736
4	0.304695	0.269173	1.11407	0.968623
6	0.297953	0.274541	1.06531	0.963551
8	0.294303	0.276784	1.03937	0.955507
10	0.292017	0.278005	1.02329	0.949042
100	0.283146	0.281737	0.962059	0.916370

Table 9.3 Numerical values of C_f, D_f, $ARE_f = ARE_f(\phi_W/\phi_{vdW})$, and $ARE_f^* = ARE_f^*(\phi_{SWW}^r/\phi_{vdW}^r)$ under Student-t densities with various degrees of freedom ν

α	C_f	D_f	ARE_f	ARE_f^*
0.1	0.393903	0.175222	1.86191	0.685991
1	0.313329	0.2720600	1.1781	1.046388
2	0.282095	0.2820950	0.954930	0.911893
10	0.222095	0.2934363	0.591916	0.611600
100	0.168549	0.2953577	0.340904	0.356871

where $F_{a,\epsilon}(x)$ is defined by symmetry for $x \leq 0$ (this family of distributions, which has been used by Hodges and Lehmann (1961), is such that the nonserial $6/\pi$ bound is achieved, in the limit, as both a and ϵ go to zero),

(2) (Table 9.2) The family $\mathcal{F}_{\text{Student}}$ of Student densities with degrees of freedom $\nu > 0$, and

(3) (Table 9.3) The family $\mathcal{F}_e$ of power-exponential densities, of the form

$$f_\alpha(x) := \frac{e^{-|x|^\alpha}}{2\Gamma(1+1/\alpha)} \qquad x \in \mathbb{R}, \ \ \alpha > 0. \tag{9.25}$$

All tables seem to confirm the same findings: both the serial and the nonerial AREs are monotone in the size of the tails, with the nonserial ARE_f attaining its maximal value ($6/\pi \approx 1.90986$) under heavy-tailed f densities, while the maximal value for the serial ARE_f^* lies somewhere around $(6/\pi)(3/\pi) \approx 1.82346$. Inspection of Table 9.1 reveals that, although the limit of C_f as $a \to 0$ is monotone in the parameter ϵ, the ratio D_f is not; from Table 9.3, the highest values of D_f under the distribution (9.24) are attained for $a \to \infty$ and $\epsilon \approx 0$.

Under Student densities $f = f_{t_\nu}$, the nonserial ARE_f is decreasing with ν, taking value 1.41277 at the Cauchy ($\nu = 1$), value one about $\nu = 15.42$ (a value of ν that is not shown in the figure; Wilcoxon is thus outperforming van der Waerden up to $\nu = 15$ degrees of freedom, with van der Waerden taking over from $\nu = 16$ on), and tending to the Gaussian value 0.95493 as $\nu \to \infty$; the serial ARE_f^* is undefined for $\nu \leq 2$, increasing for small values of ν, from an infimum of 0.878736 (obtained as $\nu \downarrow 2$) up to a maximum of 0.968852 (reached about $\nu = 4.24$), then slowly decreasing to the Gaussian value 0.911891 as $\nu \to \infty$. Sperman–Wald–Wolfowitz and Kendall thus never outperform van der Waerden autocorrelations under Student densities.

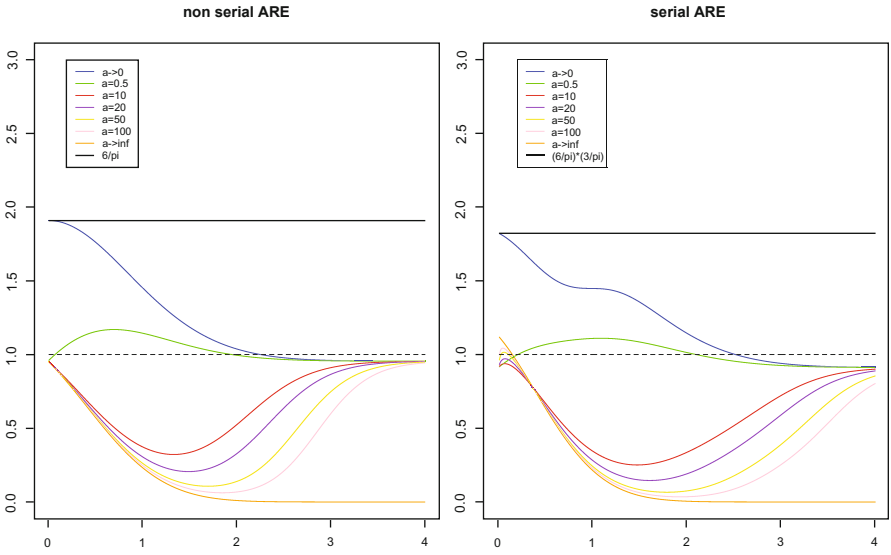


Fig. 9.2 Nonserial $\mathrm{ARE}_f = ARE_f(\phi_{\mathrm{W}}/\phi_{\mathrm{vdW}})$ (*left plot*) and serial $\mathrm{ARE}^{\star}_f = ARE^{*}_f(\phi^r_{\mathrm{SWW}}/\phi^r_{\mathrm{vdW}})$ (*right plot*) under densities $f_{a,\epsilon}$ in the Hodges–Lehmann family $\mathcal{F}_{\mathrm{HL}}$ (see 9.24), as a function of $\epsilon \in [0, 4]]$, for various choices of the parameter a

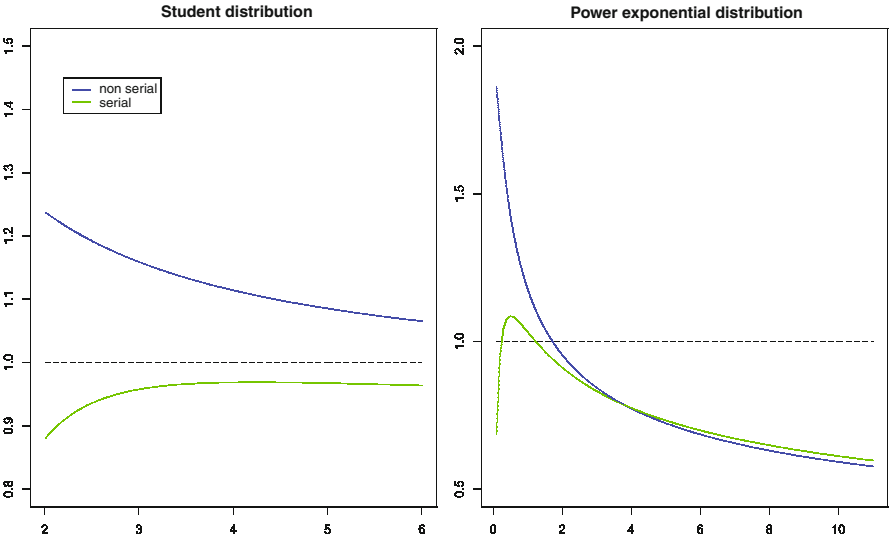


Fig. 9.3 *Left* plot: $\mathrm{ARE}_{f_\nu}(\phi_{\mathrm{W}}/\phi_{\mathrm{vdW}})$ and $\mathrm{ARE}^{\star}_{f_\nu}(\phi^r_{\mathrm{SWW}}/\phi^r_{\mathrm{vdW}})$ for f_ν the Student distribution, as a function of the degrees of freedom $\nu \in [2, 6]$. *Right* plot: ARE_{f_α} and $\mathrm{ARE}^{\star}_{f_\alpha}$ for the power exponential densities f_α (9.25), as a function of the shape parameter $\alpha \in [0, 11]$

Under the double exponential densities $f = f_\alpha$, the nonserial ARE_f is decreasing with α, with a supremum of $6/\pi$ (the Hodges–Lehmann bound, obtained as $\alpha \downarrow 0$), and reaches value one about $\alpha = 1.7206$ (similar local asymptotic performances of Wilcoxon and van der Waerden, thus, occur at power-exponentials with parameter $\alpha = 1.7206$); the serial ARE^*_f is quite bad as $\alpha \downarrow 0$, then rapidly increasing for small values of α, with a maximum of 1.08552 about $\alpha = 0.510$, then deteriorating again as $\alpha \to \infty$; for α larger than 3, the serial and nonserial AREs roughly coincide (See figs. 9.2 and 9.3).

Acknowledgments This note originates in a research visit by the last two authors to the Department of Operations Research and Financial Engineering (ORFE) at Princeton University in the Fall of 2012; ORFE's support and hospitality are gratefully acknowledged. Marc Hallin's research is supported by the Sonderforschungsbereich "Statistical modelling of nonlinear dynamic processes" (SFB 823) of the Deutsche Forschungsgemeinschaft, a Discovery Grant of the Australian Research Council, and the IAP research network grant P7/06 of the Belgian government (Belgian Science Policy). We gratefully acknowledge the pertinent comments by an anonymous referee on the original version of the manuscript, which lead to substantial improvements.

References

Chernoff H, Savage IR (1958) Asymptotic normality and efficiency of certain nonparametric tests. Ann Math Statist 29:972–994

Ferguson TS, Genest C, Hallin M (2000) Kendall's tau for serial dependence. Canad J Stat 28:587–604

Gastwirth JL (1970) On asymptotic relative efficiencies of a class of rank tests. J R Stat Soc Ser B 32:227–232

Hájek J, Šidáak Z (1967) Theory of rank tests. Academic Press, New York

Hallin M (1994) On the Pitman non-admissibility of correlogram-based methods. J Time Series Anal 15:607–611

Hallin M (2012) Asymptotic relative efficiency. In: Piegorsch W, El Shaarawi A (eds) Encyclopedia of environmetrics, 2nd edn. Wiley, New York, pp 106–110

Hallin M, Paindaveine D (2002a) Optimal tests for multivariate location based on interdirections and pseudo-Mahalanobis ranks. Ann Stat 30:1103–1133

Hallin M, Paindaveine D (2002b) Optimal procedures based on interdirections and pseudo-Mahalanobis ranks for testing multivariate elliptic white noise against ARMA dependence. Bernoulli 8:787–815

Hallin M, Paindaveine D (2004) Rank-based optimal tests of the adequacy of an elliptic VARMA model. Ann Stat 32:2642–2678

Hallin M, Paindaveine D (2005) Affine-invariant aligned rank tests for the multivariate general linear model with ARMA errors. J Multivariate Anal 93:122–163

Hallin M, Paindaveine D (2006) Semiparametrically efficient rank-based inference for shape: I Optimal rank-based tests for sphericity. Ann Stat 34:2707–2756

Hallin M, Paindaveine D (2008a) Chernoff-Savage and Hodges-Lehmann results for Wilks' test of independence. In: Balakrishnan Edsel Pena N, Silvapulle MJ (eds) Beyond parametrics in interdisciplinary research : Festschrift in honor of Professor Pranab K. Sen. I.M.S. Lecture notes—Monograph Series, pp 184–196

Hallin M, Paindaveine D (2008b) Optimal rank-based tests for homogeneity of scatter. Ann Stat 36:1261–1298

Hallin M, Puri ML (1988) Optimal rank-based procedures for time-series analysis: testing an $ARMA$ model against other $ARMA$ models. Ann Stat 16:402–432
Hallin M, Puri ML (1992) Rank tests for time series analysis. In: Brillinger D, Parzen E, Rosenblatt M (eds) New directions in time series analysis. Springer-Verlag, New York, pp 111–154
Hallin M, Puri ML (1994) Aligned rank tests for linear models with autocorrelated error terms. J Multivariate Anal 50:175–237
Hallin M, Swan Y, Verdebout T, Veredas D (2011) Rank-based testing in linear models with stable errors. J Nonparametr Stat 23:305–320
Hallin M, Swan Y, Verdebout T, Veredas D (2013) One-step R-estimation in linear models with stable errors. J Econometrics 172:195–204
Hallin M, Tribel O (2000) The efficiency of some nonparametric competitors to correlogram-based methods. In: Bruss FT, Le Cam L (eds) Game theory, optimal stopping, probability, and statistics. Papers in honor of T. S. Ferguson on the occasion of his 70th birthday. I.M.S. Lecture notes—Monograph Series, pp 249–262
Hodges JL, Lehmann EL (1956) The efficiency of some nonparametric competitors of the t-test. Ann Math Stat 2:324–335
Hodges JL, Lehmann EL (1961) Comparison of the normal scores and Wilcoxon tests. Proceedings of the fourth Berkeley symposium on mathematical statististics and probability 1:307–318
Koul HL, Saleh AKME (1993) R-estimation of the parameters of autoregressive AR(p) models. Ann Stat 21:685–701
Koul HL, Saleh AKME (1995) Autoregression quantiles and related rank-scores processes. Ann Stat 25:670–689
Nikitin Y (1995) Asymptotic efficiency of nonparametric tests. Cambridge University Press, Cambridge
Noether GE (1955) On a theorem of Pitman. Ann Math Stat 26:64–68
Paindaveine D (2004) A unified and elementary proof of serial and nonserial, univariate and multivariate, Chernoff–Savage results. Stat Methodol 1:81–91
Paindaveine D (2006) A Chernoff—Savage result for shape: on the non-admissibility of pseudo-Gaussian methods. J Multivariate Anal 97:2206–2220
Pitman EJG (1949) Notes on nonparametric statistical inference. Columbia University, mimeographed
Puri ML, Sen PK (1985) Nonparametric methods in general linear models. Wiley, New York
Randles RH, Wolfe DA (1979) Introduction to the theory of nonparametric statistics. Wiley, New York
Serfling R (1980) Approximation theorems of mathematical statistics. Wiley, New York
van der Vaart AW (1998) Asymptotic statistics. Cambridge University Press, Cambridge
Wald A, Wolfowitz J (1943) An exact test for randomness in the nonparametric case based on serial correlation. Ann Math Stat 14:378–388
Wilcoxon F (1945) Individual comparisons by ranking methods. Biom Bull 1:80–83

Chapter 10
Fiducial Theory for Free-Knot Splines

Derek L. Sonderegger and Jan Hannig

10.1 Introduction

In statistical practice, there is a tension between fitting an easily interpretable model to our data versus fitting a highly flexible model that fits the data better. One compromise between these competing ideas is a spline model. The spline model of degree p can be thought of as connected degree p polynomials with the requirement that the resulting function be "smooth" at the connection points. These connection points are usually called "knot points" and the usual smoothness requirement is that the $p-1$ derivative exists.

The simplest example is the $p = 1$ spline with one knot point, which is a linear function with some slope until the knot point, and then continues with a different slope. The smoothness requirement is that the 0th derivative exists, which is, that the function is continuous at the knot point. The resulting function is often called the hockey-stick function. A degree $p = 2$ spline with one knot point is just two quadratic curves joined together such that at the knot point the function has a 1st derivative and is therefore "smooth".

When using splines to approximate an unknown but continuous function, one important question is where to place the knots. In typical nonparametric function estimation, more knots than necessary are evenly spread along the dependent axis and a penalty based on the second derivative (also known as function "wiggliness") is introduced (Ruppert et al. 2003). An alternative approach is to use a small number of

Jan Hannig's research was supported in part by the National Science Foundation under Grant No. 1007543 and 1016441.

D. L. Sonderegger (✉)
Department of Mathematics and Statistics, Northern Arizona Univeristy,
Flagstaff, USA
e-mail: derek.sonderegger@nau.edu

J. Hannig
Department of Statistics and Operations Research, University of North Carolina,
Chapel Hill, North Carolina, USA
e-mail: jan.hannig@unc.edu

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_10,

knots but carefully place them. This problem of where to place the knots is known as the free-knot spline problem. The free-knot spline problem is primarily interested in estimating the location of the knot point and interpreting it as some sort of threshold (Toms and Lesperance 2003; Sonderegger et al. 2009).

A Bayesian solution to the arbitrary degree p problem with a fixed number of knot points is given by DiMatteo et al. (2001) and they recommend using a prior of $p(\boldsymbol{\alpha}, \boldsymbol{t}, \sigma^2) \propto \sigma^{-2}$ where σ^2 is the usual variance term, $\boldsymbol{\alpha}$ is the polynomial coefficients, and $\boldsymbol{t}$ is the vector of knot points. The maximum likelihood solution for the degree $p = 1$ free-knot spline problem is developed in Muggeo (2003) and is available in the R package segmented (Muggeo 2008).

In this chapter, we investigate the fiducial solution to the free-knot spline problem of degree $p \geq 4$. In Sect. 10.2, we first extend the univariate fiducial Bernstein-von Mises theorem to the multivariate setting, which shows that multivariate fiducial estimators have an asymptotic multivariate normal distribution under certain assumptions. In Sect. 10.3, we derive the fiducial solution to the free-knot spline problem, note that the Bernstein-von Mises assumptions are satisfied and investigate the small sample properties by conducting a simulation study of degree $p = 4$ splines comparing the fiducial solution to the Bayesian solution of DiMatteo et al. (2001). In Sect. 10.4, we give our concluding remarks.

10.1.1 Introduction to Fiducial Inference

R. A. Fisher first introduced his idea of fiducial inference (Fisher 1930) to address what he felt was the major shortcoming of Bayesian inference. His goal was to invent a posterior-like distribution without the need for a prior distribution. He did not succeed in developing a general theory for finding these fiducial distributions and his idea was met with extreme skepticism. In the 1990's, generalized confidence intervals (Weerahandi 1993) were found to have very good small sample properties and (Hannig et al. 2006) shows the connection between generalized confidence intervals and Fisher's fiducial inference. Hannig (2009) developed a general theory for developing fiducial solutions which has been used in a variety of contexts. The solution for wavelets is given by Hannig and Lee (2009). Other problems include variance components in normal mixed linear model (Hannig and Iyer 2008; Cisewski and Hannig 2012), extreme value models (Wandler and Hannig 2011), and multiple comparison issues (Wandler and Hannig 2012).

The general framework of fiducial inference assumes that the n observed data can be written as a data generating equation $\mathbb{X} = \mathbf{G}(\mathbb{U}, \boldsymbol{\xi})$, where $\boldsymbol{\xi}$ is a p length vector of parameters, and $\mathbb{U}$ is a random vector of with a completely known distribution.

Setting $\mathbb{X}_0 = (X_1, \ldots, X_p)$, $\mathbb{X}_c = (X_{p+1}, \ldots, X_n)$, $\mathbb{U}_0 = (U_1, \ldots, U_p)$ and $\mathbb{U}_c = (U_{p+1}, \ldots, U_n)$ the data generating equation can be factorized as

$$\mathbb{X}_0 = \mathbf{G_0}(\mathbb{U}_0, \boldsymbol{\xi}) \quad \text{and} \quad \mathbb{X}_c = \mathbf{G_c}(\mathbb{U}_c, \boldsymbol{\xi}).$$

Assuming that for each $\boldsymbol{\xi} \in \Xi$ that $\mathbf{G}_0(\boldsymbol{\xi}, \cdot)$ and $\mathbf{G}_c(\boldsymbol{\xi}, \cdot)$ are one-to-one and differentiable and that $\mathbf{G}_0(\boldsymbol{\xi}, \cdot)$ also invertible, then Hannig (2009) shows that the generalized fiducial distribution is

$$r\left(\boldsymbol{\xi}|\mathbf{x}_0\right) = \frac{f_{\mathbf{x}}\left(\boldsymbol{x}|\boldsymbol{\xi}\right) J_0(\boldsymbol{x}_0, \boldsymbol{\xi})}{\int_{\Xi} f_{\mathbf{x}}\left(\boldsymbol{x}|\boldsymbol{\xi}'\right) J_0\left(\boldsymbol{x}_0, \boldsymbol{\xi}'\right) d\boldsymbol{\xi}'}$$

where

$$J_0(\boldsymbol{x}_0, \boldsymbol{\xi}) = \left| \frac{det\left(\frac{d}{d\boldsymbol{\xi}} \mathbf{G}_0^{-1}\left(\boldsymbol{x}_0, \boldsymbol{\xi}\right)\right)}{det\left(\frac{d}{d\boldsymbol{x}_0} \mathbf{G}_0^{-1}\left(\boldsymbol{x}_0, \boldsymbol{\xi}\right)\right)} \right|$$

and $f_{\mathbf{x}}(x|\boldsymbol{\xi})$ is the density function. Since the choice to use the first p observations in the definition of $\mathbf{G}_0$ was arbitrary, we could select any p observations that satisfy the one-to-one, differentiable, and invertible conditions. Hannig (2009, 2013) suggests letting the Jacobian $J\left(\boldsymbol{x}, \boldsymbol{\xi}\right)$ be the average of all possible values of J_0 and using

$$r(\boldsymbol{\xi}|\mathbf{x}) = \frac{f_{\mathbf{x}}\left(\boldsymbol{x}|\boldsymbol{\xi}\right) J(\boldsymbol{x}, \boldsymbol{\xi})}{\int_{\Xi} f_{\mathbf{x}}(\boldsymbol{x}|\boldsymbol{\xi}') J(\boldsymbol{x}, \boldsymbol{\xi}') d\boldsymbol{\xi}'}. \tag{10.1}$$

This distribution is similar to a Bayesian posterior distribution with the Jacobian taking the role of the prior. This can be seen in the standard regression problem where the Jacobian simplifies to $J\left(\boldsymbol{x}, \boldsymbol{\xi}\right) = \sigma^{-2} h\left(\boldsymbol{x}\right)$. Since $h\left(\boldsymbol{x}\right)$ is in Jacobians in both the numerator and denominator, it will cancel and the fiducial distribution is the same as the Bayesian posterior with commonly used reference prior distribution σ^{-2}.

Two numerical issues commonly arise in the evaluation of the fiducial density. First, it is often not feasible to take the average of all possible values of J_0 because the number of possible permutations grows as n^p. This is often solved by taking a random selection of possible J_0 and using the sample mean as an approximation to $J\left(\boldsymbol{x}, \boldsymbol{\xi}\right)$. A second challenge comes in evaluating the denominator, which is often intractable due to the high number of dimensions. To address this issue, we use the standard Markov Chain Monte Carlo (MCMC) techniques to take a random sample from the fiducial density and all subsequent inference is based on that sample.

10.2 Asymptotic Consistency of the Multivariate Fiducial Estimators

Many estimators have an asymptotic normal distribution and fiducial estimators are no exception. Conditions A0–A6 in Appendix A are the standard conditions sufficient to prove that the maximum likelihood estimators to have an asymptotic normal distribution (Lehmann and Casella 1998). That is, the maximum likelihood estimators $\hat{\boldsymbol{\xi}}_n$ are consistent and $\sqrt{n}(\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi})$ is asymptotically normal with mean 0 and covariance matrix $\left[I(\boldsymbol{\xi})\right]^{-1}$, where $I\left(\boldsymbol{\xi}\right)$ is the Fisher information matrix.

The Bernstein-von Mises theorem gives conditions (B1–B2 in appendix A.10.1) under which the Bayesian posterior distribution is asymptotically normal (van der Vaart 1998, 2003). In brief, the proof can be thought of as showing that the posterior distribution becomes close to the distribution of the MLE. Hannig (2009) gives sufficient conditions (C1–C2) for the univariate fiducial distribution to converge to the Bayesian posterior which is in turn close to the MLE distribution. Hannig (2009) defines the following assumptions:

(C1) For any $\delta > 0$

$$\inf_{\xi \notin B(\xi_0,\delta)} \frac{\min_{i=1\ldots n} \log f(\xi, X_i)}{\left|L_n(\xi) - L_n(\xi_0)\right|} \xrightarrow{P_{\xi_0}} 0$$

where $L_n(\xi) = \sum_{i=1}^{n} \log f(x_i|\xi)$ and $B\left(\xi_0, \delta\right)$ is a neighborhood of diameter δ centered at ξ_0.

(C2) Let $\pi(\xi) = E_{\xi_0} J_0(X_0, \xi)$. The Jacobian function $J(X, \xi) \xrightarrow{a.s.} \pi(\xi)$ uniformly on compacts in ξ. In the single variable case, this reduces to assumptions that $J(X, \xi)$ is continuous in ξ, $\pi(\xi)$ is finite and $\pi(\xi_0) > 0$, and for some δ_0

$$E_{\xi_0}\left(\sup_{\xi \in B(\xi_0,\delta)} J_0(X, \xi)\right) < \infty.$$

The extension to the multiparameter case follows Yeo and Johnson (2001) and replaces assumption C2 with C2a, b, and c. Let $\omega \in \Omega$ be a collection of indices in $\{1, 2, \ldots p\}$ and $\bar{\omega} = \{1, 2, \ldots p\} \backslash \omega$. Define

$$J_\omega(x_\omega; \xi) = E_{\xi_0}\left[J_0(x_\omega, X_{\bar{\omega}}; \xi)\right].$$

(C2.a) There exists an integrable and symmetric function $g(\cdot)$ and compact space $\bar{B}\left(\xi_0, \delta\right)$ such that for $\xi \in \bar{B}\left(\xi_0, \delta\right)$ and $x \in \mathbb{R}^p$ then $|J(x; \xi)| \leq g(x)$.

(C2.b) There exists a sequence of measurable sets S_M^p such that

$$P\left(\mathbb{R}^p - \cup_{M=1}^{\infty} S_M^p\right) = 0.$$

(C2.c) For each M and for all $\omega \in \Omega$, $J_\omega(x_\omega; \xi)$ is equicontinuous in ξ for $\{x_\omega\} \in S_M^\omega$ where $S_M^p = S_M^\omega \times S_M^{\bar{\omega}}$.

Let $\mathcal{R}_\xi$ be an observation from the fiducial distribution $r(\xi|x)$ and denote the density of $s = \sqrt{n}\left(\mathcal{R}_\xi - \hat{\xi}_n\right)$ by $\pi^*(\xi, x)$.

Theorem 1 *Given a random sample of independent observations $X_1, \ldots, X_n$, then under assumptions A0–A6, B1–B2, and C1–C2.c*

$$\int_{\mathbb{R}^p} \left| \pi^*(s, x) - \frac{\sqrt{det\left|I\left(\xi_0\right)\right|}}{\sqrt{2\pi}} e^{-s^T I(\xi_0) s/2} \right| ds \xrightarrow{P_{\theta_0}} 0. \tag{10.2}$$

Due to its technical nature, we relegate the proof to Appendix A, Sect. 10.2.

10.3 Fiducial Free-Knot Splines

We consider the fiducial free-knot spline solution for splines of degree $p \geq 4$. We first derive the fiducial distribution using a simple set of spline basis functions so that the derivatives necessary derivatives can be calculated for the Jacobian. We then address the asymptotic behavior of the solution by applying Theorem 1 to this solution. We next consider the practical issue of creating a proposal distribution for the MCMC simulation. Finally, we conduct a simulation study to compare the fiducial method to the Bayesian solution with reference prior $\propto \sigma^{-2}$ in four scenarios.

10.3.1 Deriving the Fiducial Free-Knot Spline

Suppose data $\{x_i, y_i\}$ for $i \in [1, \ldots, n]$ are generated from

$$y_i = g\left(x_i | \boldsymbol{\alpha}, \boldsymbol{t}\right) + \sigma \epsilon_i$$

where $\epsilon_i \overset{iid}{\sim} N\left(0, 1\right)$ and $g\left(x | \boldsymbol{\alpha}, \boldsymbol{t}\right)$ is a degree $p \geq 4$ spline with κ knot points denoted $\boldsymbol{t}$ and $p + \kappa + 1$ polynomial coefficients $\boldsymbol{\alpha}$. We assume that κ is known, but the knot locations $\boldsymbol{t}$ are unknown and are the primary target of investigation. The spline can be written using many different basis functions, but computational ease, we consider the piecewise truncated polynomial basis

$$g(x_i | \boldsymbol{\alpha}, \boldsymbol{t}) = \sum_{j=0}^{p} \alpha_j x_i^j + \sum_{k=1}^{\kappa} \alpha_{p+k} \left(x_i - t_k\right)_+^p$$

where

$$\left(u\right)_+ = \begin{cases} 0 & \text{if } u < 0 \\ u & \text{otherwise} \end{cases}$$

is the truncation operator and has higher precedence than the exponentiation. This representation makes it clear that the response function changes form at each knot point. The following derivation of the fiducial solution could, in principle, be done using more numerically stable basis functions, but the derivatives become more complicated. Our early work on this problem implemented a purely numerical solution using the b-spline basis, but the lack of closed form representation prevented showing that Theorem 1 holds.

We derive the fiducial solution the to free-knot spline solution by first inverting the data generating equation and subsequently solving for ϵ_i. The Jacobian is then found by taking the derivative (with respect to the parameters of interest) of the inversion result.

Specifically, we denote the inverse by $\boldsymbol{G}_0^{-1}(y_i, \boldsymbol{\xi})$ and let $\boldsymbol{\xi} = \left\{\boldsymbol{\alpha}, \boldsymbol{t}, \sigma^2\right\}^T$. We recognize that

$$\epsilon_i = \boldsymbol{G}_0^{-1}(y_i, \boldsymbol{\xi}) = \frac{1}{\sigma}\left(y_i - g(x_i|\boldsymbol{\theta})\right)$$

and therefore the partial derivatives with respect to the parameters are

$$\begin{aligned}
\frac{\partial \boldsymbol{G}_0^{-1}(y_i, \boldsymbol{\xi})}{\partial \boldsymbol{\alpha}} &= -\frac{1}{\sigma}\left(1, x_i, \ldots, x_i^p, (x_i - t_1)_+^p, \ldots, (x_i - t_\kappa)_+^p\right) \\
\frac{\partial \boldsymbol{G}_0^{-1}(y_i, \boldsymbol{\xi})}{\partial \boldsymbol{t}} &= \frac{p}{\sigma}\left(\alpha_{p+1}(x_i - t_1)_+^{p-1}, \ldots, \alpha_{p+\kappa}(x_i - t_\kappa)_+^{p-1}\right) \\
\frac{\partial \boldsymbol{G}_0^{-1}(y_i, \boldsymbol{\xi})}{\partial \sigma^2} &= -\frac{1}{2\sigma^3}\left(y_i - g(x_i|\boldsymbol{\theta})\right) \\
\frac{\partial \boldsymbol{G}_0^{-1}(y_i, \boldsymbol{\xi})}{\partial y_i} &= \frac{1}{\sigma}
\end{aligned}$$

where we define $0^0 = 0$ for notational convenience. Let $\boldsymbol{y}_0 = \left\{y_{(1)}, \ldots, y_{(l)}\right\}$ where $l = p + \kappa + 2$ be any selection of data points that satisfies the necessary invertability criteria. The Jacobian using these data points $\boldsymbol{y}_0$ is therefore

$$J_0\left(\boldsymbol{y}_0, \boldsymbol{\xi}\right) = \left|\frac{1}{\sigma^2} p^\kappa \det\left[\begin{array}{ccc} \boldsymbol{B}_{\boldsymbol{\alpha}} & \boldsymbol{B}_{\boldsymbol{t}} & \boldsymbol{B}_{\sigma^2} \end{array}\right]\right|$$

where

$$\boldsymbol{B}_{\boldsymbol{\alpha}} = \begin{bmatrix} 1 & x_{(1)} & \ldots & x_{(1)}^p & (x_{(1)} - t_1)_+^p & \ldots & (x_{(1)} - t_\kappa)_+^p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{(l)} & \ldots & x_{(l)}^p & (x_{(l)} - t_1)_+^p & \ldots & (x_{(l)} - t_\kappa)_+^p \end{bmatrix},$$

$$\boldsymbol{B}_{\boldsymbol{t}} = \begin{bmatrix} \alpha_{1+p+1}\left(x_{(1)} - t_1\right)_+^{p-1} & \ldots & \alpha_{1+p+\kappa}\left(x_{(1)} - t_\kappa\right)_+^{p-1} \\ \vdots & \ddots & \vdots \\ \alpha_{1+p+1}\left(x_{(l)} - t_1\right)_+^{p-1} & \ldots & \alpha_{1+p+\kappa}\left(x_{(l)} - t_\kappa\right)_+^{p-1} \end{bmatrix},$$

and

$$\boldsymbol{B}_{\sigma^2} = \begin{bmatrix} -\frac{1}{2}\left(y_{(1)} - g\left(x_{(1)}|\boldsymbol{\theta}\right)\right) \\ \vdots \\ -\frac{1}{2}\left(y_{(l)} - g\left(x_{(l)}|\boldsymbol{\theta}\right)\right) \end{bmatrix}.$$

Because $\boldsymbol{B}_{\sigma^2}$ contains a subtraction of a linear combination of columns of $\boldsymbol{B}_\alpha$ and $\boldsymbol{B}_{\boldsymbol{t}}$, the subtraction does not change the determinant and therefore

$$\left|\frac{1}{\sigma^2} p^\kappa \det\left[\begin{array}{ccc} \boldsymbol{B}_{\boldsymbol{\alpha}} & \boldsymbol{B}_{\boldsymbol{t}} & \boldsymbol{B}_{\sigma^2} \end{array}\right]\right| = \left|\frac{1}{2\sigma^2} p^\kappa \det\left[\begin{array}{ccc} \boldsymbol{B}_{\boldsymbol{\alpha}} & \boldsymbol{B}_{\boldsymbol{t}} & \tilde{\boldsymbol{B}}_{\sigma^2} \end{array}\right]\right|$$

where

$$\tilde{\boldsymbol{B}}_{\sigma^2} = \begin{bmatrix} y_{(1)} \\ \vdots \\ y_{(l)} \end{bmatrix}.$$

However, the question of which sets of indices satisfy the one-to-one and invertibility requirements is not obvious. A sufficient condition is that the set of indices includes at least two observations from each interknot region. As we are primarily interested in cases where the number of observations is much larger than the number of knots, this condition is not onerous.

Theorem 2 *Given $g(x|\boldsymbol{\alpha}, \boldsymbol{t})$, a free-knot spline of degree 4 or greater with parameters $\boldsymbol{\alpha}$ and $\boldsymbol{t}$ with truncated polynomial basis functions and observations with x_i a randomly selected element on some contiguous interval $[a, b]$ of $\mathbb{R}$ and $y_i = g(x_i|\boldsymbol{\alpha}, \boldsymbol{t}) + \sigma\epsilon_i$ where $\epsilon_i \overset{iid}{\sim} N(0, 1)$, define $\boldsymbol{\xi} = (\boldsymbol{\alpha}, \boldsymbol{t}, \sigma^2)$. Let $\pi^*(\boldsymbol{\xi}, \boldsymbol{y})$ be the fiducial distribution of $\mathcal{R}_{\boldsymbol{\xi}}$. Then,*

$$\int_{\mathbb{R}^p} \left| \pi^*(\boldsymbol{s}, \boldsymbol{y}) - \frac{\sqrt{det\left|I\left(\boldsymbol{\xi}_0\right)\right|}}{\sqrt{2\pi}} e^{-s^T I(\boldsymbol{\xi}_0)\mathbf{s}/2} \right| d\boldsymbol{s} \overset{P_{\boldsymbol{\theta}_0}}{\to} 0$$

Proof. It suffices to show that the free-knot spline satisfies assumptions A0–A6, B1–B2, C1–C2.c. These are shown in Appendix B, which is available at the author's website. □

A shortcoming of this proof is the requirement that $p \geq 4$, while many free-knot spline applications are concerned with degree $p = 1$ or 2 splines.

10.3.2 Numerical Evaluation of the Fiducial Density

There are two substantial challenges to numerical evaluation of the fiducial density. The first is that the Jacobian does not simplify to a "nice" expression utilizing all of the data. We use the suggestion of Hannig (2009) to use the mean of randomly selected Jacobians as an estimate of $J(\boldsymbol{x}, \boldsymbol{\xi})$. The second challenge is that the scaling constant in the denominator of Eq.10.1 is intractable and we only know the fiducial distribution up to a scaling constant. This is the same numerical challenge found in evaluating a Bayesian posterior distribution and we use MCMC methods to select a random sample from the fiducial distribution. The key step of the MCMC is to produce good proposal values, which is often difficult when model parameters are highly correlated. Unfortunately, our choice to use the analytically convenient truncated polynomial basis functions results in numerically inconvenient correlated parameters.

If the knot point locations were known, then the fiducial distribution of the $\boldsymbol{\alpha}$ and σ^2 terms is known and is same as the Bayesian posterior distribution with reference prior distribution $\propto \sigma^{-2}$. More formally, letting $\boldsymbol{X} = [\boldsymbol{B}_\alpha, \boldsymbol{B}_t]$ be the design matrix with fixed and known knot points, the fiducial distribution is $\boldsymbol{\alpha}|\sigma^2, y \sim N\left(\hat{\boldsymbol{\alpha}}, V_{\boldsymbol{\alpha}}\right)$ where $\hat{\boldsymbol{\alpha}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{y}$ and $V_{\boldsymbol{\alpha}} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}$. Similarly, the marginal distribution of $\sigma^2|\boldsymbol{y}$ is a scaled inverse-χ^2 distribution, $\sigma^2|\boldsymbol{y} \sim \text{Inv-}\chi^2\left(n - p - \kappa - 1, s^2\right)$ where s^2 is the usual mean squared error term $s^2 = \left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}\right)^T\left(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\alpha}}\right)/(n-p-\kappa-1)$. We denote the product of these distributions as the *fixed fiducial distribution.*

Unfortunately, the fiducial distribution of σ^2 and $\boldsymbol{\alpha}$ conditioned on the knot point locations $\boldsymbol{t}$ is not the earlier fixed fiducial distribution because the Jacobian term cannot be factored into terms that contain only $\boldsymbol{t}$ parameters or only $\boldsymbol{\alpha}$ terms. However, the fixed fiducial distribution does provide a useful ... proposal distribution in a MCMC estimation.

The procedure for creating a proposed value in the Markov chain is to take the current knot locations and perturb them by adding a small amount of noise. The proposed knots are $\boldsymbol{t}^* = \boldsymbol{t} + u^*$ where $u^* \sim MVN\left(0, \sigma_k^2\boldsymbol{I}_k\right)$, $\boldsymbol{I}_k$ is the identity matrix and σ_k^2 is the tuning parameter for the MCMC and reflects how much each knot point is "jittered." We then take these proposed knot points and consider them as known and use the aforementioned fixed fiducial distributions to produce proposed values for σ^2 and then $\boldsymbol{\alpha}$.

These three proposal distributions are multiplied to create the total proposal distribution $T\left(\boldsymbol{\xi}^*|\boldsymbol{\xi}\right)$. For the given proposed set of parameters, if the ratio

$$r = \frac{f\left(\boldsymbol{y}|\boldsymbol{\xi}^*\right)}{f\left(\boldsymbol{y}|\boldsymbol{\xi}\right)}\frac{T\left(\boldsymbol{\xi}|\boldsymbol{\xi}^*\right)}{T\left(\boldsymbol{\xi}^*|\boldsymbol{\xi}\right)}$$

is greater than a *Uniform*(0,1) random deviate, we accept the proposed value as the next value in the Markov chain, otherwise the current vector of parameters is used.

The use of the fixed fiducial distribution is similar in spirit to the method of DiMatteo et al. (2001) where they integrate out the $\boldsymbol{\alpha}$ and σ^2 parameters and consider only the distribution of the knot points $\boldsymbol{t}$. The difference is that their prior factored nicely whereas the Jacobian does not.

10.3.3 Simulation Study for Degree Four Splines

The simulation study will compare the fiducial method to the Bayesian method on four different degree four splines, all defined on domain $x \in [0, 1]$ and with a similar range of y values.

The software we used to evaluate the performance of the fiducial solution compared to the Bayesian method with prior $\propto \sigma^{-2}$ used the same software for implementing the MCMC and generating proposed values, with the only difference in the software being whether the likelihood was multiplied by the Bayesian prior distribution or the calculated Jacobian.

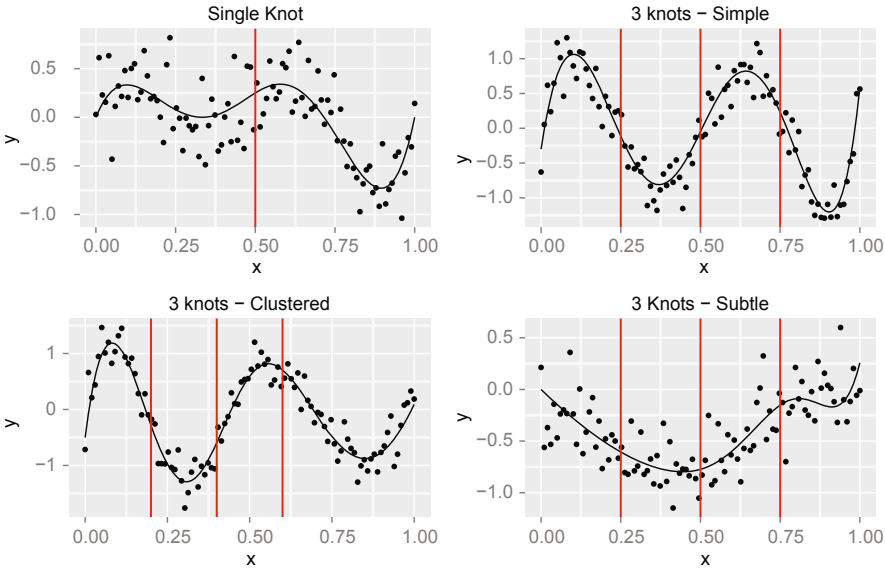


Fig. 10.1 Degree four examples—the examples shown are the high sample size and high variability case. *Upper left panel, the "Single" knot case;* upper right panel, three knots evenly spread across the x-axis which we refer to as the "simple" three knot case. *Lower left panel, three knots "clustered" to the left side of the x-axis;* lower right panel, 3 evenly spaced knots with with a "subtle" effect initially but with increasing effect size from left to right

Table 10.1 Coefficients defining the four different simulation scenarios

Scenario	Knot point(s)	Spline coefficients
Single knot	0.5	$0, 8, -60, 144, -108, 256$
Three knots-simple	0.25, 0.50, 0.75	$0, 30, -203, 386, -179, -276, 854, 270$
Three knots-clustered	0.20, 0.40, 0.60	$-1, 47, -397, 967, -640, -510, 2002, -1043$
Three knots-subtle	0.25, 0.50, 0.75	$0, -3, 2, 1, 1, 10, -100, 600$

The first spline has a single knot point at the center of the range of x values. The second has three knot points even spread through the x values. The third function also has three knot points, but the knots are not evenly distributed across the x values, instead they are clustered toward the left. The final function has three knot points evenly spread on the x-axis, but has a subtle change to the function at the first knot point, a larger change at the middle knot point and a large change at final knot. These functions are shown in Fig. 10.1 and are defined in table 10.1.

For each scenario, we compared the methods using two different levels of variance and two samples sizes. The sample sizes $n = \{40, 100\}$ were chosen to reflect real world cases of data scarcity and moderate abundance. The two variance levels reflect an idealistically low level of variance ($\sigma = 0.1$) and a more realistic "signal-to-noise" level ($\sigma = 0.25$) commonly seen in the authors' applied work.

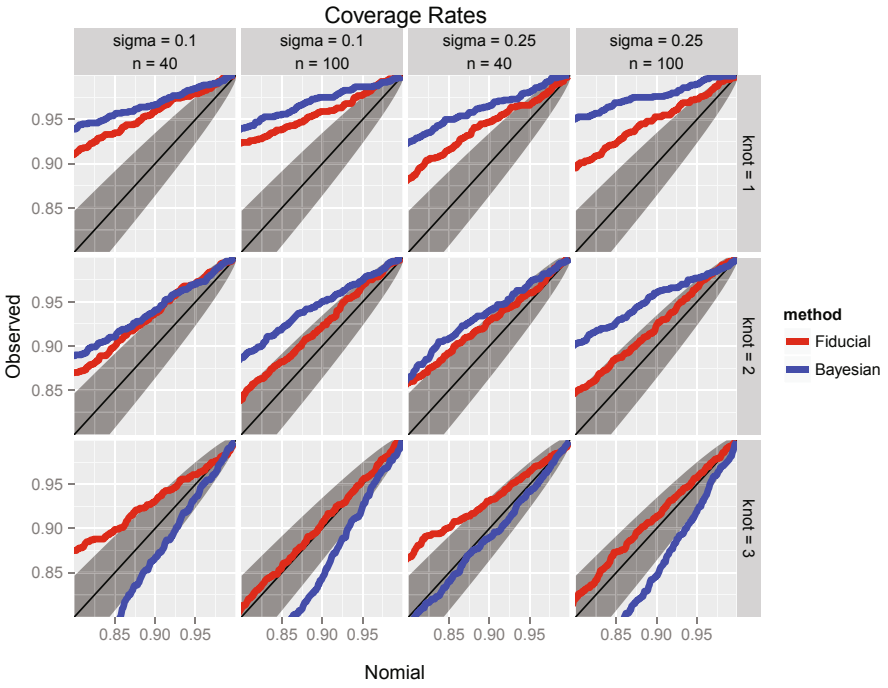


Fig. 10.2 Coverage rates for the "Three Knot-Clustered" simulation. The color (*red, blue*) represents the method (*fiducial, Bayesian*). The topmost panel is the coverage of knot one in the $sigma = 0.1, n = 40$ simulation. Graphs of the coverage rate for the other scenarios was similar and can be found in Appendix B

We consider coverage rates (Fig. 10.2) of the fiducial credible intervals of the true knot point values. In the "coverage plots" presented, the X-axis denotes the desired confidence level and the Y-axis is the observed coverage rate in the experiment. If the observed coverage rate is below the equivalence line ($y = x$), then the method is considered *liberal* and if the observed rate is above the equivalence line then the method is *conservative*. Ideally, a method would lie exactly on the equivalence line but a conservative method is more preferable to a liberal because claiming a 95% coverage rate, when, in truth, the coverage rate is less is a more serious error than having the true coverage rate being larger than claimed. The only complaint against a conservative method is that the lengths of confidence intervals are larger than necessary to achieve the desired confidence level.

In the coverage plots presented, the oval lines around the equivalence line are the region in which we would expect the coverage rates to lie in due to stochastic variation in the simulation. For each simulation, the α-level necessary for the inclusion of the true parameter value in a confidence interval was calculated. Since, the data is actually generated from the model we are fitting, then these α-levels should follow a uniform distribution if the coverage rates are correct. The jth ordered statistic of

these, therefore, follows a $Beta\,(j, n-j+1)$ distribution and appropriate 95 % point-wise confidence region can be calculated from this.

For each of the 16 combinations of function type, sample size, and variance, 1,000 simulations were performed and took approximately 4 days to run on a desktop computer. For the three knot simple case, a fiducial analysis took $\approx$ 100 s while the Bayesian solution took $\approx$ 10 s. The reason for this drastic difference is that for every evaluation of the fiducial density, the jacobian at that point must be estimated from averaging repeated samples of $J_0\,(\boldsymbol{x}_0, \boldsymbol{\xi})$.

10.3.4 Simulation Results

We display only the results of the "Three Knot-Clustered" function here and graphics of the other functions to the appendix because the results were similar.

The coverage rates (Fig. 10.2) for the for the fiducial method was typically slightly higher than the desired level, but was generally within the expected coverage region given the sample size. The Bayesian method was also generally consistent with the desired rate, but was liberal in a few instances. For the "single knot case," both the fiducial and Bayesian methods were neither conservative nor liberal. In the "simple three knot case", the Bayesian method was liberal for all knots and sample sizes in the high variance cases, while the fiducial method was liberal for only the first knot in the high variance high sample case. In the "three knot clustered" case, the Bayesian method is conservative for knots one and two, but liberal for the third. In contrast, the fiducial intervals were conservative for knot one. In the "three knot subtle" case, the Bayesian method was conservative for knot one and two. The fiducial method was conservative for knot one in the small variance case. Overall, the fiducial estimator tends to have a coverage rate that is closer to the nominal rate than the Bayesian.

The lengths of the *95 %* confidence (or credible) interval lengths showed a consistent trend across our simulation (Fig. 10.3). The Bayesian intervals were longer in every scenario we examined, however, the difference was the smallest in the single knot case.

10.4 Conclusions

Free-knot splines are computationally challenging to fit, but in instances where inference on the knot points is desired, we believe that the fiducial method is a viable method for analysis. Simulation shows that the fiducial method is an effective method for the high degree free-knot spline problem and is superior to the Bayesian solution with prior $\propto \sigma^{-2}$. This is consistent with our previous experience of the fiducial method being equivalent to or better than the standard Bayesian solution derived using the default prior (Cisewski and Hannig 2012).

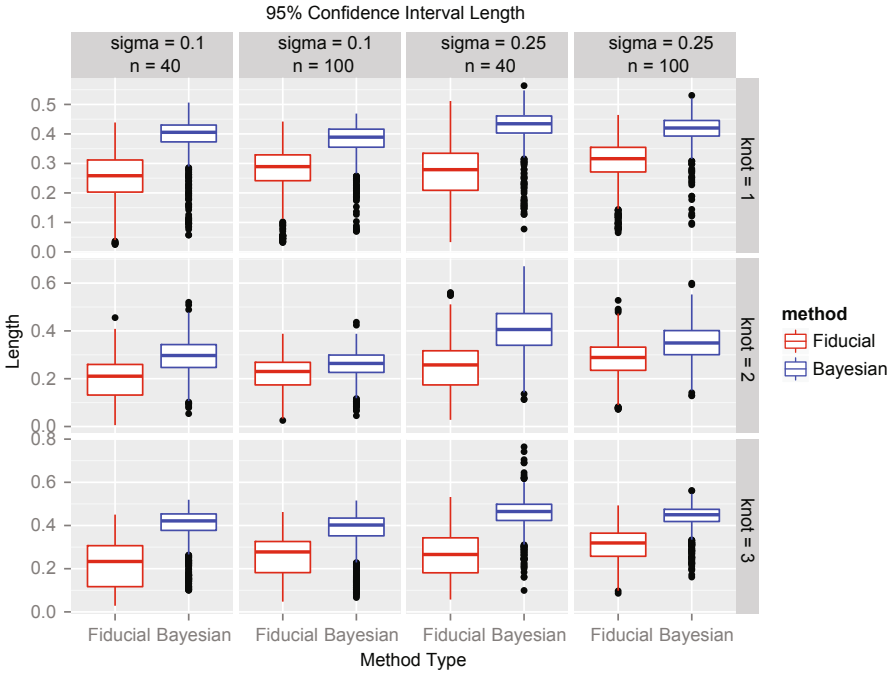


Fig. 10.3 Confidence interval lengths for the "Three Knot-Clustered" simulation. The color (*red, blue*) represents the method (*fiducial, Bayesian*). The topmost panel is the coverage of knot one in the $sigma = 0.1, n = 40$ simulation. Graphs of the interval lengths for the other scenarios was similar and can be found in Appendix B

The foundational theory for fiducial inference is given in Hannig (2009) and this chapter expands the fiducial Bernstein-von Mises theorem to the multivariate setting. However, this result is not the most general result possible due to the restrictive assumption of continuous fourth derivatives. In particular, we believe that replacing the standard differentiability conditions used in the proof of Theorem 2, with Le Cam's continuity in quadratic mean assumptions (van der Vaart 1998) would allow us to relax the differentiability assumptions to obtain the most general Bernstein-von Mises type theorem for fiducial distributions. This is a subject of future work.

One case where continuous derivatives do not exist is the case of free-knot splines of degree one. These are of great interest due to the interpretability of the knot point as a change point. Based on our simulations results, we conjecture that asymptotic normality holds even in this case. Further investigation into the the behavior of the fiducial method in this case relative to both the Bayesian solution and segmented regression (Muggeo 2003) are of interest.

For this chapter, we assume that the number of knot points to be fit is known. In some cases, the physical system under investigation provides insight into the number of knots. In the cases where the number of knots is not known, a reversible jump

MCMC algorithm could allow for model selection, but would require some penalty term on models of increasing complexity.

Perhaps, the largest reason for practitioners to not use new methodologies is the lack of accessible software packages. If a new methodology has no freely available software, or requires expensive software packages (such as Matlab and its associated toolboxes), applied researchers tend to not adopt a method. To alleviate this issue, we have provided the R package "FiducialFreeKnotSplines" that contains the software used in the simulation studies conducted for this chapter and is freely available on the Comprehensive R Archive Network (CRAN).

Acknowledgements Dr Hannig thanks Prof. Hira Koul for his encouragement and help ever since he was a graduate student at Michigan State University. A young researcher cannot ask for a better role model. The authors also thank the two anonymous referees that made several useful suggestions for improving the manuscript.

Appendix A: Proof of Asymptotic Normality of Fiducial Estimators

We start with several assumptions. The assumptions A0–A6 are sufficient for the maximum likelihood estimate to converge asymptotically to a normal distribution and can be found in Lehmann and Casella (1998) as 6.3 (A0)–(A2) and 6.5 (A)–(D). The assumption B2 shows that the Jacobian converges to a prior (Hannig 2009) and B1 is the assumption necessary for the Bayesian solution to converge to that of the MLE (Ghosh and Ramamoorthi 2003, Theorem 1.4.1).

A.1 Assumptions

A.1.1 Conditions for Asymptotic Normality of the MLE

(A0) The distributions $P_{\boldsymbol{\xi}}$ are distinct.
(A1) The set $\{x : f(x|\boldsymbol{\xi}) > 0\}$ is independent of the choice of $\boldsymbol{\xi}$.
(A2) The data $\boldsymbol{X} = \{X_1, \ldots, X_n\}$ are independent identically distributed (i.i.d.) with probability density $f(\cdot\,|\boldsymbol{\xi})$.
(A3) There exists an open neighborhood about the true parameter value $\boldsymbol{\xi}_0$ such that all third partial derivatives $\left(\partial^3/\partial\xi_i\partial\xi_j\partial\xi_k\right) f(\boldsymbol{x}|\boldsymbol{\xi})$ exist in the neighborhood, denoted by $B(\boldsymbol{\xi}_0, \delta)$.
(A4) The first and second derivatives of $L(\boldsymbol{\xi}, x) = \log f(x|\boldsymbol{\xi})$ satisfy

$$E_{\boldsymbol{\xi}}\left[\frac{\partial}{\partial\xi_j}L(\boldsymbol{\xi}, x)\right] = 0$$

and

$$I_{j,k}(\boldsymbol{\xi}) = E_{\boldsymbol{\xi}}\left[\frac{\partial}{\partial \xi_j} L(\boldsymbol{\xi}, x) \cdot \frac{\partial}{\partial \xi_k} L(\boldsymbol{\xi}, x)\right] = -E_{\boldsymbol{\xi}}\left[\frac{\partial^2}{\partial \xi_j \partial \xi_k} L(\boldsymbol{\xi}, x)\right].$$

(A5) The information matrix $I(\boldsymbol{\xi})$ is positive definite for all $\boldsymbol{\xi} \in B(\boldsymbol{\xi}_0, \delta)$

(A6) There exists functions $M_{jkl}(\boldsymbol{x})$ such that

$$\sup_{\boldsymbol{\xi} \in B(\boldsymbol{\xi}_0, \delta)} \left| \frac{\partial^3}{\partial \xi_j \partial \xi_k \partial \xi_l} L(\boldsymbol{\xi}, x) \right| \leq M_{j,k,l}(x) \quad \text{and} \quad E_{\boldsymbol{\xi}_0} M_{j,k,l}(x) < \infty$$

A.1.2 Conditions for the Bayesian Posterior Distribution to be Close to That of the MLE.

Let $\pi(\boldsymbol{\xi}) = E_{\boldsymbol{\xi}_0} J_0(X_0, \boldsymbol{\xi})$ and $L_n(\boldsymbol{\xi}) = \sum L(\boldsymbol{\xi}, X_i)$

(B1) For any $\delta > 0$ there exists $\epsilon > 0$ such that

$$P_{\boldsymbol{\xi}_0}\left\{ \sup_{\boldsymbol{\xi} \notin B(\boldsymbol{\xi}_0, \delta)} \frac{1}{n}\left(L_n(\boldsymbol{\xi}) - L_n(\boldsymbol{\xi}_0)\right) \leq -\epsilon \right\} \to 1$$

(B2) $\pi\,(\boldsymbol{\xi})$ is positive at $\boldsymbol{\xi}_0$

A.1.3 Conditions for Showing That the Fiducial Distribution is Close to the Bayesian Posterior

(C1) For any $\delta > 0$

$$\inf_{\boldsymbol{\xi} \notin B(\boldsymbol{\xi}_0, \delta)} \frac{\min_{\mathrm{i}=1\ldots\mathrm{n}} \mathrm{L}(\boldsymbol{\xi}, \mathrm{X_i})}{\left|L_n(\boldsymbol{\xi}) - L_n(\boldsymbol{\xi}_0)\right|} \xrightarrow{P_{\boldsymbol{\xi}_0}} 0$$

(C2) Let $\pi(\boldsymbol{\xi}) = E_{\boldsymbol{\xi}_0} J_0(X_0, \boldsymbol{\xi})$. The Jacobian function $J\,(\boldsymbol{X}, \boldsymbol{\xi}) \xrightarrow{a.s.} \pi\,(\boldsymbol{\xi})$ uniformly on compacts in $\boldsymbol{\xi}$. In the single variable case, this reduces to $J\,(\boldsymbol{X}, \xi)$ is continuous in ξ, $\pi\,(\xi)$ is finite and $\pi\,(\xi_0) > 0$, and for some δ_0

$$E_{\xi_0}\left(\sup_{\xi \in B(\xi_0, \delta)} J_0\,(\boldsymbol{X}, \xi) \right) < \infty.$$

In the multivariate case, we follow Yeo and Johnson (2001). Let

$$J_j\left(x_1, \ldots, x_j; \boldsymbol{\xi}\right) = E_{\boldsymbol{\xi}_0}\left[J_0\left(x_1, \ldots, x_j, X_{j+1}, \ldots, X_k; \boldsymbol{\xi}\right)\right].$$

(C2.a) There exists a integrable and symmetric functions $g\left(x_1,\ldots,x_j\right)$ and compact space $\bar{B}\left(\boldsymbol{\xi}_0,\delta\right)$ such that for $\boldsymbol{\xi}\in\bar{B}\left(\boldsymbol{\xi}_0,\delta\right)$ then $\left|J_j\left(x_1,\ldots,x_j;\boldsymbol{\xi}\right)\right|\leq g\left(x_1,\ldots,x_j\right)$ for $j=1,\ldots,k$.

(C2.b) There exists a sequence of measurable sets S_M^k such that

$$P\left(\mathbb{R}^k-\cup_{M=1}^{\infty}S_M^k\right)=0$$

(C2.c) For each M and for all $j\in 1,\ldots,k$, $J_j\left(x_1,\ldots,x_j;\boldsymbol{\xi}\right)$ is equicontinuous in $\boldsymbol{\xi}$ for $\{x_1,\ldots,x_j\}\in S_M^j$ where $S_M^k=S_M^jS_M^{k-j}$.

A.2 *Proof of Asymptotic Normality of Multivariate Fiducial Estimators*

We now prove the asymptotic normality (Theorem 1) for multivariate fiducial estimators.

Proof. Assume without loss of generality that $\xi\in\Xi=\mathbb{R}^p$. We denote $J_n\left(\boldsymbol{x}_n,\boldsymbol{\xi}\right)$ as the average of all possible Jacobians over a sample of size n and $\pi\left(\boldsymbol{\xi}\right)=E_{\xi_0}J_0\left(\boldsymbol{x},\boldsymbol{\xi}\right)$. Assumption C2 and the uniform strong law of large numbers for U-statistics imply that $J_n\left(\boldsymbol{x},\boldsymbol{\xi}\right)\overset{a.s.}{\to}\pi\left(\boldsymbol{\xi}\right)$ uniformly in $\boldsymbol{\xi}\in\bar{B}\left(\boldsymbol{\xi}_0,\delta\right)$ and that $\pi\left(\boldsymbol{\xi}\right)$ is continuous. Therefore,

$$\sup_{\boldsymbol{\xi}\in\bar{B}\left(\boldsymbol{\xi}_0,\delta\right)}\left|J_n\left(x_n,\boldsymbol{\xi}\right)-\pi\left(\boldsymbol{\xi}\right)\right|\to 0\ \ P_{\boldsymbol{\xi}_0}\,a.s.$$

The multivariate proof now proceeds in a similar fashion as the univariate case. Let

$$\begin{aligned}\pi^*\left(\boldsymbol{s},\boldsymbol{x}\right)&=\frac{J_n\left(\boldsymbol{x}_n,\hat{\boldsymbol{\xi}}_n+\frac{\boldsymbol{s}}{\sqrt{n}}\right)f\left(\boldsymbol{x}_n|\hat{\boldsymbol{\xi}}_n+\frac{\boldsymbol{s}}{\sqrt{n}}\right)}{\int_{\mathbb{R}^p}J_n\left(\boldsymbol{x}_n,\hat{\boldsymbol{\xi}}_n+\frac{\boldsymbol{t}}{\sqrt{n}}\right)f\left(\boldsymbol{x}_n|\hat{\boldsymbol{\xi}}_n+\frac{\boldsymbol{t}}{\sqrt{n}}\right)d\boldsymbol{t}}\\&=\frac{J_n\left(\boldsymbol{x}_n,\hat{\boldsymbol{\xi}}_n+\frac{\boldsymbol{s}}{\sqrt{n}}\right)\exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n+\frac{\boldsymbol{s}}{\sqrt{n}}\right)\right]}{\int_{\mathbb{R}^p}J_n\left(\boldsymbol{x}_n,\hat{\boldsymbol{\xi}}_n+\frac{\boldsymbol{t}}{\sqrt{n}}\right)\exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n+\frac{\boldsymbol{t}}{\sqrt{n}}\right)\right]d\boldsymbol{t}}\\&=\frac{J_n\left(\boldsymbol{x}_n,\hat{\boldsymbol{\xi}}_n+\frac{\boldsymbol{s}}{\sqrt{n}}\right)\exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n+\frac{\boldsymbol{s}}{\sqrt{n}}\right)-L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right]}{\int_{\mathbb{R}^p}J_n\left(\boldsymbol{x}_n,\hat{\boldsymbol{\xi}}_n+\frac{\boldsymbol{t}}{\sqrt{n}}\right)\exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n+\frac{\boldsymbol{t}}{\sqrt{n}}\right)-L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right]d\boldsymbol{t}}\end{aligned}$$

and just as Ghosh and Ramamoorthi (2003), we let $H=-\frac{1}{n}\frac{\partial}{\partial\boldsymbol{\xi}\partial\boldsymbol{\xi}}L_n\left(\hat{\boldsymbol{\xi}}_n\right)$ and we notice that $H\to I\left(\boldsymbol{\xi}_0\right)\ a.s.\ P_{\boldsymbol{\xi}_0}$. It will be sufficient to prove

$$\int_{\mathbb{R}^p} \left| J_n\left(x_n, \hat{\xi}_n + \frac{t}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\xi}_n + \frac{t}{\sqrt{n}}\right) - L_n\left(\hat{\xi}_n\right)\right] \right.$$
$$\left. -\pi\left(\boldsymbol{\xi}_0\right) \exp\left[\frac{-\boldsymbol{t}^T I\left(\boldsymbol{\xi}_0\right)\boldsymbol{t}}{2}\right] \right| d\boldsymbol{t} \xrightarrow{P_{\xi_0}} 0 \tag{10.3}$$

Let t_i represent the ith component of vector $\boldsymbol{t}$. By Taylor's Theorem, we can compute

$$\begin{aligned} L_n\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) &= L_n\left(\hat{\boldsymbol{\xi}}_n\right) + \sum_{i=1}^{p}\left(\frac{t_i}{\sqrt{n}}\right)\frac{\partial}{\partial\xi_i}L_n\left(\hat{\boldsymbol{\xi}}_n\right) \\ &\quad + \frac{1}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}\left(\frac{t_i t_j}{\left(\sqrt{n}\right)^2}\frac{\partial}{\partial\xi_i\partial\xi_j}L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right) \\ &\quad + \frac{1}{6}\sum_{i=1}^{p}\sum_{j=1}^{p}\sum_{k=1}^{p}\left(\frac{t_i t_j t_k}{\left(\sqrt{n}\right)^3}\frac{\partial}{\partial\xi_i\partial\xi_j\partial\xi_k}L_n\left(\boldsymbol{\xi}'\right)\right) \\ &= L_n\left(\hat{\boldsymbol{\xi}}_n\right) - \frac{\boldsymbol{t}^T H\boldsymbol{t}}{2} + R_n \end{aligned}$$

for some $\boldsymbol{\xi}' \in \left[\hat{\boldsymbol{\xi}}_n, \hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right]$. Notice that $R_n = O_p\left(\|\boldsymbol{t}\|/n^{3/2}\right)$.

Given any $0 < \delta < \delta_0$ and $c > 0$, we break $\mathbb{R}^p$ into three regions:

$$\begin{aligned} A_1 &= \left\{\boldsymbol{t} :\ \|\boldsymbol{t}\| < c\log\sqrt{n}\right\} \\ A_2 &= \left\{\boldsymbol{t} :\ c\log\sqrt{n} < \|\boldsymbol{t}\| < \delta\sqrt{n}\right\} \\ A_3 &= \left\{\boldsymbol{t} :\ \delta\sqrt{n} < \|\boldsymbol{t}\|\right\} \end{aligned}$$

On $A_1 \cup A_2$ we compute

$$\begin{aligned} &\int_{A_1\cup A_2} \left| J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] \right. \\ &\qquad \left. -\pi\left(\boldsymbol{\xi}_0\right)\exp\left[-\frac{1}{2}\boldsymbol{t}' I\left(\boldsymbol{\xi}_0\right)\boldsymbol{t}\right] \right| d\boldsymbol{t} \\ &\leq \int_{A_1\cup A_2} \left| J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) - \pi\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) \right| \\ &\qquad \cdot \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] d\boldsymbol{t} \\ &+ \int_{A_1\cup A_2} \left| \pi\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] \right. \\ &\qquad \left. -\pi\left(\boldsymbol{\xi}_0\right)\exp\left[-\frac{1}{2}\boldsymbol{t}' I\left(\boldsymbol{\xi}_0\right)\boldsymbol{t}\right] \right| d\boldsymbol{t} \end{aligned}$$

Since $\pi(\cdot)$ is a proper prior on $A_1 \cup A_2$, then the second term goes to 0 by the Bayesian Bernstein-von Mises theorem. Next we notice that

$$\begin{aligned}
&\int_{A_1 \cup A_2} \left| J_n\left(x, \hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) - \pi\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right)\right| \\
&\qquad \cdot \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] d\boldsymbol{t} \\
&\leq \sup_{\boldsymbol{t} \in A_1 \cup A_2} \left| J_n\left(x, \hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) - \pi\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right)\right| \\
&\qquad \cdot \int_{A_1 \cup A_2} \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] d\boldsymbol{t}
\end{aligned}$$

Since $\sqrt{n}\left(\hat{\boldsymbol{\xi}}_n - \boldsymbol{\xi}_0\right) \xrightarrow{\mathcal{D}} N\left(0, I\left(\boldsymbol{\xi}_0\right)^{-1}\right)$, then

$$P_{\boldsymbol{\xi}_0}\left[\left\{\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n};\ \boldsymbol{t} \in A_1 \cup A_2\right\} \subset B\left(\boldsymbol{\xi}_0, \delta_0\right)\right] \to 1.$$

Furthermore, since $L_n\left(\hat{\boldsymbol{\xi}}_n + \boldsymbol{t}/\sqrt{n}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right) = -\frac{\boldsymbol{t}^T H \boldsymbol{t}}{2} + R_n$ then the integral converges in probability to 1. Since $\max_{\mathbf{t} \in A_1 \cup A_2} \left\|\boldsymbol{t}/\sqrt{\mathrm{n}}\right\| \leq \delta$ and $J_n \to \pi$, then the term $\to 0$ in probability.

Next, we turn to

$$\begin{aligned}
&\int_{A_3} \left| J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right]\right. \\
&\qquad \left. -\pi\left(\boldsymbol{\xi}_0\right) \exp\left[\frac{-\boldsymbol{t}^T I\left(\boldsymbol{\xi}_0\right) \boldsymbol{t}}{2}\right]\right| d\boldsymbol{t} \\
&\leq \int_{A_3} J_n\left(\boldsymbol{x}_i, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] d\boldsymbol{t} \\
&\qquad + \int_{A_3} \pi\left(\boldsymbol{\xi}_0\right) \exp\left[\frac{-\boldsymbol{t}^T I\left(\boldsymbol{\xi}_0\right) \boldsymbol{t}}{2}\right] d\boldsymbol{t}
\end{aligned}$$

The second integral goes to 0 in $P_{\boldsymbol{\xi}_0}$ probability because $\min_{\mathrm{A}_3} \|\boldsymbol{t}\| \to \infty$. As for the first integral,

$$\begin{aligned}
&\int_{A_3} J_n\left(\boldsymbol{x}, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] d\boldsymbol{t} \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_{A_3} J\left(\boldsymbol{x}_i, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] d\boldsymbol{t} \\
&= \frac{1}{n} \sum_{i=1}^{n} \int_{A_3} J\left(\boldsymbol{x}_i, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) f\left(x_i | \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) \\
&\qquad \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right) - \log f\left(x_i | \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right)\right] d\boldsymbol{t}
\end{aligned}$$

Because $J(\cdot)$ is a probability measure, then so is $J(\cdot)\, f(\cdot)$. Assumption C1 assures that the exponent goes to $-\infty$ and therefore the integral converges to 0 in probability.

Having shown Eq. 10.3, we now follow Ghosh and Ramamoorthi (2003) and let

$$C_n = \int_{\mathbb{R}^p} \left| J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{t}}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{t}}{\sqrt{n}}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] \right| d\boldsymbol{t}$$

then the main result to be proved (Eq. 10.2) becomes

$$C_n^{-1} \left\{ \int_{\mathbb{R}^p} \left| J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] \cdot \right.\right.$$
$$\left.\left. - C_n \frac{\sqrt{det\left|I\left(\boldsymbol{\xi}_0\right)\right|}}{\sqrt{2\pi}} e^{-\boldsymbol{s}^T I(\boldsymbol{\xi}_0)\boldsymbol{s}/2} \right| \right\} d\boldsymbol{s} \overset{P_{\boldsymbol{\xi}_0}}{\to} 0 \quad (10.4)$$

Because

$$\int_{\mathbb{R}^p} J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n\right) \exp\left[-\frac{\boldsymbol{s}^T H \boldsymbol{s}}{2}\right] d\boldsymbol{s} = J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n\right) \int_{\mathbb{R}^p} \exp\left[-\frac{\boldsymbol{s}^T H \boldsymbol{s}}{2}\right] d\boldsymbol{s}$$
$$= J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n\right) \frac{\sqrt{2\pi}}{\sqrt{\det(H)}}$$
$$\overset{a.s.}{\to} \pi\left(\boldsymbol{\xi}_0\right) \sqrt{\frac{2\pi}{\det\left(I\left(\boldsymbol{\xi}_0\right)\right)}}$$

and Eq. 10.3 imply that $C_n \overset{P}{\to} \pi\left(\boldsymbol{\xi}_0\right) \sqrt{\frac{2\pi}{\det(I(\boldsymbol{\xi}_0))}}$ it is enough to show that the integral in Eq:10.4 goes to 0 in probability. This integral is less than $I_1 + I_2$ where

$$I_1 = \int_{\mathbb{R}^P} \left| J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) \exp\left[L_n\left(\hat{\boldsymbol{\xi}}_n + \frac{\boldsymbol{s}}{\sqrt{n}}\right) - L_n\left(\hat{\boldsymbol{\xi}}_n\right)\right] \right.$$
$$\left. - J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n\right) \exp\left[\frac{-\boldsymbol{s}^T H \boldsymbol{s}}{2}\right] \right| d\boldsymbol{s}$$

and

$$I_2 = \int_{\mathbb{R}^P} \left| J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n\right) \exp\left[\frac{-\boldsymbol{s}^T H \boldsymbol{s}}{2}\right] - C_n \frac{\sqrt{det\left|I\left(\boldsymbol{\xi}_0\right)\right|}}{\sqrt{2\pi}} e^{-\boldsymbol{s}^T I(\boldsymbol{\xi}_0)\boldsymbol{s}/2} \right| d\boldsymbol{s}.$$

Eq. 10.3 shows that $I_1 \to 0$ in probability and I_2 is

$$I_2 = \left| J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n\right) - C_n \frac{\sqrt{det\left|I\left(\boldsymbol{\xi}_0\right)\right|}}{\sqrt{2\pi}} \right| \int_{\mathbb{R}^P} \exp\left[\frac{-\boldsymbol{s}^T H \boldsymbol{s}}{2}\right] d\boldsymbol{s}$$
$$\overset{P}{\to} 0$$

because $J_n\left(\boldsymbol{x}_n, \hat{\boldsymbol{\xi}}_n\right) \overset{P}{\to} \pi\left(\boldsymbol{\xi}_0\right)$ and $C_n \overset{P}{\to} \pi\left(\boldsymbol{\xi}_0\right) \sqrt{\frac{2\pi}{\det(I(\boldsymbol{\xi}_0))}}$. □

Appendix B: Proof of Assumptions for Free-Knot Splines Using a Truncated Polynomial Basis

We now consider the free-knot spline case. Suppose we are interested in a p degree (order $m = p + 1$) polynomial spline with κ knot points, $\boldsymbol{t} = \{t_1, \ldots, t_\kappa\}^T$ where $t_k \in (a + \delta, b - \delta)$ and $|t_i - t_j| \leq \delta$ for $i \neq j$ and some $\delta > 0$. Furthermore, we assume that the data points $\{x_i, y_i\}$ independent with the distribution of the x_i having positive density on $[a, b]$.

Denote the truncated polynomial spline basis functions as

$$\begin{aligned} N(x, \boldsymbol{t}) &= \{N_1(x, \boldsymbol{t}), \ldots, N_{\kappa+m}(x, \boldsymbol{t})\}^T \\ &= \left\{1, x, \ldots, x^p, (x - t_1)_+^p, \ldots, (x - t_\kappa)_+^p\right\}^T \end{aligned}$$

and let $y_i = N(x_i, \boldsymbol{t})^T \boldsymbol{\alpha} + \sigma \epsilon_i$ where $\epsilon_i \overset{iid}{\sim} N(0, 1)$ and thus the density function is

$$f(y, \boldsymbol{\xi}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}\left(y - N(x, \boldsymbol{t})^T \boldsymbol{\alpha}\right)^2\right]$$

where $\boldsymbol{\xi} = \{\boldsymbol{t}, \boldsymbol{\alpha}, \sigma^2\}$ and the log-likelihood function is

$$L(\boldsymbol{\xi}, y) = \frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\left(y - N(x, \boldsymbol{t})^T \boldsymbol{\alpha}\right)^2$$

B.1 Assumptions A0–A4

Assumptions A0–A2 are satisfied. We now consider assumption A3 and A4. We note that if $p \geq 4$ then the necessary three continuous derivatives exist and now examine the derivatives. Let $\boldsymbol{\theta} = \{\boldsymbol{t}, \boldsymbol{\alpha}\}$ and thus

$$\begin{aligned} E_\xi\left[\frac{\partial}{\partial\theta_j} L(\boldsymbol{\xi}, y)\right] &= E_\xi\left[-\frac{1}{2\sigma^2} 2\left(y - N(x, \boldsymbol{t})^T \boldsymbol{\alpha}\right)\left(-\frac{\partial}{\partial\theta_j} N(x, \boldsymbol{t})^T \boldsymbol{\alpha}\right)\right] \\ &= -\frac{1}{2\sigma^2} 2\left(E_\xi[y] - N(x, \boldsymbol{t})^T \boldsymbol{\alpha}\right)\left(-\frac{\partial}{\partial\theta_j} N(x, \boldsymbol{t})^T \boldsymbol{\alpha}\right) \\ &= 0 \end{aligned}$$

and

$$\begin{aligned} E_\xi\left[\frac{\partial}{\partial\sigma^2} L(\boldsymbol{\xi}, \boldsymbol{y})\right] &= E_\xi\left[-\frac{1}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\left(y - N(x, \boldsymbol{t})^T \boldsymbol{\alpha}\right)^2\right] \\ &= -\frac{1}{2\sigma^2} + \frac{1}{2\left(\sigma^2\right)^2}\left(\sigma^2\right) \\ &= 0. \end{aligned}$$

Next, we consider information matrix. First, we consider the $\boldsymbol{\theta}$ terms.

$$\begin{aligned} E_{\boldsymbol{\xi}}\left[\frac{\partial}{\partial\theta_j}L(\boldsymbol{\xi},\boldsymbol{y})\,\frac{\partial}{\partial\theta_k}L(\boldsymbol{\xi},\boldsymbol{y})\right] &= E_{\boldsymbol{\xi}}\left[\frac{1}{\sigma^4}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2\left(\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial}{\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\right] \\ &= \frac{1}{\sigma^4}E_{\boldsymbol{\xi}}\left[\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2\right]\left(\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial}{\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right) \\ &= \frac{1}{\sigma^2}\left(\frac{\partial}{\partial\theta_j}N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial}{\partial\theta_k}N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}\right) \end{aligned}$$

The j,k partials for the second derivative are

$$\begin{aligned} \frac{\partial^2}{\partial\theta_j\partial\theta_k}L(\boldsymbol{\xi},y) &= \frac{\partial}{\partial\theta_j}\left[-\frac{1}{2\sigma^2}2\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(-\frac{\partial}{\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\right] \\ &= \frac{\partial}{\partial\theta_j}\left[-\frac{1}{\sigma^2}\left(-y_i\left(\frac{\partial}{\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)+N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\left(\frac{\partial}{\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\right)\right] \\ &= -\frac{1}{\sigma^2}\left[-y\frac{\partial^2}{\partial\theta_j\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}+\left(\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial}{\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\right. \\ &\quad \left.+N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\frac{\partial^2}{\partial\theta_j\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right] \end{aligned}$$

which have expectation

$$\begin{aligned} E_{\boldsymbol{\xi}}\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}L(\boldsymbol{\xi},y)\right] &= -\frac{1}{\sigma^2}\left(\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial}{\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right) \\ &= -E_{\boldsymbol{\xi}}\left[\frac{\partial}{\partial\theta_j}L(\boldsymbol{\xi},y)\,\frac{\partial}{\partial\theta_k}L(\boldsymbol{\xi},\boldsymbol{y})\right] \end{aligned}$$

as necessary. Next, we consider

$$\begin{aligned} &E_{\boldsymbol{\xi}}\left[\frac{\partial}{\partial\theta_j}L(\boldsymbol{\xi},y)\,\frac{\partial}{\partial\sigma^2}L(\boldsymbol{\xi},y)\right] \\ &\quad= E_{\boldsymbol{\xi}}\left[\frac{1}{\sigma^2}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\left[-\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2\right]\right] \\ &\quad= E_{\boldsymbol{\xi}}\left[-\frac{1}{2\sigma^4}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}+\frac{1}{2\sigma^6}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^3\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right] \\ &\quad= 0 \end{aligned}$$

which is equal to

$$\begin{aligned} E_{\boldsymbol{\xi}}\left[\frac{\partial}{\partial\theta_j\partial\sigma^2}L(\boldsymbol{\xi},\boldsymbol{y})\right] &= E_{\boldsymbol{\xi}}\left[\frac{2}{2\sigma^4}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right] \\ &= 0. \end{aligned}$$

Finally,

$$
\begin{aligned}
&E_{\xi}\left[\frac{\partial}{\partial\sigma^2}L(\boldsymbol{\xi},y)\,\frac{\partial}{\partial\sigma^2}L(\boldsymbol{\xi},y)\right]\\
&\quad=E_{\xi}\left[\left\{-\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2\right\}\left\{-\frac{1}{2\sigma^2}+\frac{1}{2\sigma^4}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2\right\}\right]\\
&\quad=E_{\xi}\left[\frac{1}{4\sigma^4}-\frac{2}{4\sigma^6}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2+\frac{1}{4\sigma^8}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^4\right]\\
&\quad=\frac{1}{4\sigma_0^4}-\frac{2}{4\sigma_0^6}\sigma_0^2+\frac{1}{4\sigma_0^8}3\sigma_0^4\\
&\quad=\frac{2}{4\sigma_0^4}
\end{aligned}
$$

which is equal to

$$
\begin{aligned}
-E_{\xi}\left[\frac{\partial}{\partial\sigma^2\partial\sigma^2}L(\boldsymbol{\xi},y)\right]&=-E_{\xi}\left[\frac{1}{2}\sigma^{-4}-\frac{2}{2}\sigma^{-6}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2\right]\\
&=-\frac{1}{2}\sigma_0^{-4}+\frac{2}{2}\sigma_0^{-4}.
\end{aligned}
$$

Therefore, the interchange of integration and differentiation is justified.

B.2 Assumptions A5

To address whether the information matrix is positive definite, we notice that since $E_{\xi}\left[\frac{\partial}{\partial\sigma^2}L(\boldsymbol{\xi},y)\,\frac{\partial}{\partial\sigma^2}L(\boldsymbol{\xi},y)\right]>0$ and $E_{\xi}\left[\frac{\partial}{\partial\theta_j}L(\boldsymbol{\xi},y)\,\frac{\partial}{\partial\sigma^2}L(\boldsymbol{\xi},y)\right]=0$, we only need to be concerned with the submatrix

$$
\begin{aligned}
I_{j,k}(\boldsymbol{\theta})&=\sum_{i=1}^{n}E_{\xi}\left[\frac{\partial}{\partial\theta_j}L(\boldsymbol{\xi},y_i)\,\frac{\partial}{\partial\theta_k}L(\boldsymbol{\xi},y_i)\right]\\
&=\frac{1}{\sigma^2}\sum_{i=1}^{n}\left(\frac{\partial}{\partial\theta_j}N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial}{\partial\theta_k}N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}\right).
\end{aligned}
$$

where the σ^{-2} term can be ignored because it does not affect the positive definiteness. First, we note

$$
\begin{aligned}
\frac{\partial}{\partial t_j}N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}&=-p\left(x_i-t_j\right)_+^{p-1}\alpha_{p+j+1}\\
\frac{\partial}{\partial\alpha_j}N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}&=N_j(x_i,\boldsymbol{t}).
\end{aligned}
$$

If we let

$$X = \begin{bmatrix} N_1(x_1,\boldsymbol{t}) & \cdots & N_{m+\kappa}(x_1,\boldsymbol{t}) & \frac{\partial}{\partial t_1}N(x_1,\boldsymbol{t})^T\boldsymbol{\alpha} & \cdots & \frac{\partial}{\partial t_\kappa}N(x_1,\boldsymbol{t})^T\boldsymbol{\alpha} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ N_1(x_n,\boldsymbol{t}) & \cdots & N_{m+\kappa}(x_n,\boldsymbol{t}) & \frac{\partial}{\partial t_1}N(x_n,\boldsymbol{t})^T\boldsymbol{\alpha} & \cdots & \frac{\partial}{\partial t_\kappa}N(x_n,\boldsymbol{t})^T\boldsymbol{\alpha} \end{bmatrix}$$

then $I(\boldsymbol{\theta}) = X^T X$. Then, $I(\boldsymbol{\theta})$ is positive definite if the columns of X are linearly independent. This is true under the assumptions that $t_j \neq t_k$ and that $\alpha_{m+j} \neq 0$.

B.3 Assumptions A6

We next consider a bound on the third partial derivatives. We start with the derivatives of the basis functions.

$$\frac{\partial^2}{\partial t_j \partial t_k} N(x,\boldsymbol{t})^T\boldsymbol{\alpha} = 0 \qquad \text{if } j \neq k$$

$$\frac{\partial^2}{\partial t_j \partial t_j} N(x,\boldsymbol{t})^T\boldsymbol{\alpha} = p(p-1)\left(x - t_j\right)_+^{p-2} \alpha_{p+j+1}$$

$$\frac{\partial^2}{\partial \alpha_j \partial \alpha_k} N(x,\boldsymbol{t})^T\boldsymbol{\alpha} = 0$$

$$\frac{\partial^2}{\partial t_j \partial \alpha_{p+j+1}} N(x,\boldsymbol{t})^T\boldsymbol{\alpha} = -p\left(x - t_j\right)_+^{p-1}$$

$$\frac{\partial^3}{\partial t_j \partial t_j \partial t_j} N(x,\boldsymbol{t})^T\boldsymbol{\alpha} = -p(p-1)(p-2)\left(x - t_j\right)_+^{p-3} \alpha_{p+j+1}$$

$$\frac{\partial^3}{\partial t_j \partial t_j \partial \alpha_{p+j+1}} N(x,\boldsymbol{t})^T\boldsymbol{\alpha} = p(p-1)\left(x - t_j\right)_+^{p-2}$$

Since, x is an element of a compact set, then for $\boldsymbol{\xi} \in B(\boldsymbol{\xi}_0, \delta)$ all of the earlier partials are bounded as is $N(x,\boldsymbol{t})^T\boldsymbol{\alpha}$. Therefore

$$\frac{\partial^3}{\partial \theta_j \partial \theta_k \partial \theta_l} L(\boldsymbol{\xi}, x)$$
$$= -\frac{1}{\sigma^2}\left[-y\frac{\partial^3}{\partial \theta_j \partial \theta_k \partial \theta_l} N(x,\boldsymbol{t})^T\boldsymbol{\alpha} + \left(\frac{\partial^2}{\partial \theta_j \partial \theta_k} N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial^2}{\partial \theta_l} N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\right.$$

$$+\left(\frac{\partial^2}{\partial\theta_j\partial\theta_l}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial^2}{\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)$$
$$+\left(\frac{\partial^2}{\partial\theta_l\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial^2}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)$$
$$+N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\left(\frac{\partial^3}{\partial\theta_j\partial\theta_k\partial\theta_l}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\Bigg]$$

and

$$\frac{\partial^3}{\partial\theta_j\partial\theta_k\partial\sigma^2}L(\boldsymbol{\xi},x)$$
$$=\frac{1}{\sigma^4}\left[-y\frac{\partial^2}{\partial\theta_j\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}+\left(\frac{\partial}{\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\right.$$
$$\left.+N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\frac{\partial^2}{\partial\theta_j\partial\theta_k}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right]$$

and

$$\frac{\partial^3}{\partial\theta_j\partial\sigma^2\partial\sigma^2}L(\boldsymbol{\xi},y)=-\frac{2}{\sigma^6}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)\left(-\frac{\partial}{\partial\theta_j}N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)$$

and

$$\frac{\partial^3}{\partial\sigma^2\partial\sigma^2\partial\sigma^2}L(\boldsymbol{\xi},\boldsymbol{y})=-\frac{1}{\sigma^6}+\frac{3}{\sigma^8}\left(y-N(x,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2$$

are also bounded $\boldsymbol{\xi}\in B(\boldsymbol{\xi}_0,\delta)$ since $\sigma_0^2>0$ by assumption. The expectation of the bounds also clearly exists.

B.4 Lemmas

To show that the remaining assumptions are satisfied, we first examine the behavior of

$$g(\boldsymbol{\theta}_0,\boldsymbol{\theta},x_i)=N(x_i,\boldsymbol{t}_0)^T\boldsymbol{\alpha}_0-N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}.$$

Notice that for x_i chosen on a uniform grid over $[a,b]$ then

$$\frac{1}{n}\sum_{i=1}^{n}(g(\boldsymbol{\theta}_0,\boldsymbol{\theta},x_i))^2\to\frac{1}{b-a}\int_a^b(g(\boldsymbol{\theta}_0,\boldsymbol{\theta},x))^2\,dx.$$

Furthermore we notice that $g\,(\boldsymbol{\theta}_0,\boldsymbol{\theta},x)$ is also a spline. The sum of the two splines is also a spline. Consider the degree p case of $g\,(x|\boldsymbol{\alpha},t)+g\,(x|\boldsymbol{\alpha}^*,t^*)$ where $t<t^*$.

Then the sum is a spline with knot points $\{t, t^*\}$ and whose first $p+1$ coefficients are $\boldsymbol{\alpha} + \boldsymbol{\alpha}^*$ and last two coefficients are $\left\{\alpha_{p+1}, \alpha^*_{p+1}\right\}$.

At this point, we also notice

$$\begin{aligned} E\left[n^{-1}\sum g\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i\right)\epsilon_i\right] &= n^{-1}\sum g\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i\right) E\left[\epsilon_i\right] \\ &= 0 \end{aligned}$$

$$\begin{aligned} V\left[n^{-1}\sum g\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i\right)\epsilon_i\right] &= n^{-2} V\left[\sum g\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i\right)\epsilon_i\right] \\ &= n^{-2}\sum V\left[g\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i\right)\epsilon_i\right] \\ &= n^{-2}\sum g\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i\right)^2 V\left[\epsilon_i\right] \\ &= n^{-2}\sum g\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i\right)^2 \\ &\to 0 \end{aligned}$$

and that $\sum \epsilon_i^2 \sim \chi_n^2$ and thus $n^{-1}\sum \epsilon_i^2$ converges in probability to the constant 1. Therefore, by the SLLN,

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n}\left[g\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i\right) + \sigma_0\epsilon_i\right]^2 &= \frac{1}{n}\sum_{i=1}^{n}\left[g\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i\right)\right]^2 + \frac{2\sigma_0}{n}\sum_{i=1}^{n}\epsilon_i g\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i\right) + \frac{\sigma_0^2}{n}\sum_{i=1}^{n}\epsilon_i^2 \\ &= \frac{1}{n}\sum_{i=1}^{n}\left[g\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i\right)\right]^2 + O_p\left(n^{-1}\right) + \frac{\sigma_0^2}{n}\sum_{i=1}^{n}\epsilon_i^2 \\ &\xrightarrow{a.s.} \frac{1}{b-a}\int_a^b \left(g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x)\right)^2 dx + \sigma_0^2. \end{aligned}$$

Lemma 1. *Given a degree* p *polynomial* $g(x|\boldsymbol{\alpha})$ *on* $[a, b]$ *with coefficients* $\boldsymbol{\alpha}$*, then* $\exists\ \lambda_{n,m}, \lambda_{n,M} > 0$ *such that* $||\boldsymbol{\alpha}||^2\lambda_{n,m}^2 \le \frac{1}{n}\sum_{i=1}^{n}\left[g(x_i|\boldsymbol{\alpha})\right]^2 \le ||\boldsymbol{\alpha}||^2\lambda_{n,M}^2$.

Proof. If $\boldsymbol{\alpha} = 0$, then $g\left(x|\boldsymbol{\alpha}\right) = 0$ and the result is obvious. If $g\left(x|\boldsymbol{\alpha}\right)$ is a polynomial with at least one non-zero coefficient, it therefore cannot be identically zero on $[a, b]$ and therefore for $n > p$ then $\frac{1}{n}\sum\left[g(x_i|\boldsymbol{\alpha})\right]^2 > 0$ since the polynomial can only have at most p zeros. We notice that

$$\begin{aligned} \int_a^b \left[g(x|\boldsymbol{\alpha})\right]^2 dx &= \int_a^b \left[\sum_{i=0}^{p}\alpha_i^2 x^{2i} + 2\sum_{i=0}^{p-1}\sum_{j=i+1}^{p}\alpha_i\alpha_j x^{i+j}\right] dx \\ &= \sum_{i=0}^{p}\frac{\alpha_i^2}{i+1}x^{2i+1} + 2\sum_{i=0}^{p-1}\sum_{j=i+1}^{p}\frac{\alpha_i\alpha_j}{i+j+1}x^{i+j+1}\Bigg|_{x=a}^{b} \\ &= \boldsymbol{\alpha}^T X\boldsymbol{\alpha} \end{aligned}$$

where the matrix $\boldsymbol{X}$ has i, j element $\left(b^{i+j} - a^{i+j}\right)/(i+j)$. Since $\int_a^b [g(x|\boldsymbol{\alpha})]^2\, dx > 0$ for all $\boldsymbol{\alpha}$ then the matrix X must be positive definite. Next we notice that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}[g(x_i|\boldsymbol{\alpha})]^2 =& \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\alpha}^T\boldsymbol{X}_i\boldsymbol{\alpha}\\ =&\boldsymbol{\alpha}^T\left(\frac{1}{n}\sum\boldsymbol{X}_i\right)\boldsymbol{\alpha}\\ =&\boldsymbol{\alpha}^T\boldsymbol{X}_n\boldsymbol{\alpha}\end{aligned}$$

and therefore $\boldsymbol{X}_n \to \boldsymbol{X}$ and therefore, denoting the eigenvalues of $\boldsymbol{X}_n$ as $\boldsymbol{\lambda}_n$ and the eigenvalues of $\boldsymbol{X}$ as $\boldsymbol{\lambda}$, we have $\boldsymbol{\lambda}_n \to \boldsymbol{\lambda}$

Letting $\lambda_{n,m}$ and $\lambda_{n,M}$ be the minimum and maximum eigenvalues of $\boldsymbol{X}_n$ be the largest, then $\lambda_{n,m}^2 \|\boldsymbol{\alpha}\|^2 \le \frac{1}{n}\sum[g(x|\boldsymbol{\alpha})]^2 \le \lambda_{n,M}^2 \|\boldsymbol{\alpha}\|^2$. □

The values $\lambda_{n,m}, \lambda_{n,M}$ depend on the interval that the polynomial is integrated/summed over and that if $a = b$, then the integral is zero. In the following lemmas, we assume that there is some minimal distance between two knot-points and between a knot-point and the boundary values a, b.

Lemma 2. *Given a degree* p *spline* $g(x|\boldsymbol{\theta})$ *with* κ *knot points on* $[a,b]$, *let* $\tau = (|a| \vee |b|)^\kappa$. *Then* $\forall\ \delta > 2\tau,\ \exists\ \lambda_n > 0$ *such that if* $\|\boldsymbol{\theta}\| > \delta$ *then* $\frac{1}{n}\sum[g(x_i|\boldsymbol{\theta})]^2 > \left(\delta^2+\tau^2\right)\lambda_n$.

Proof. Notice that $||\boldsymbol{\theta}||^2 > \delta^2 > 4\tau^2$ implies $||\boldsymbol{\alpha}||^2 > \delta^2 - \tau^2$. First we consider the case of $\kappa = 1$. If $\alpha_0^2 + \cdots + \alpha_p^2 > \left(\delta^2+\tau^2\right)/9$ then $\frac{1}{n}\sum[g(x_i|\boldsymbol{\theta})]^2\, 1_{[a,t]}(x_i) > \lambda_n\left(\delta^2+\tau^2\right)$ for some $\lambda_n > 0$. If $\alpha_0^2 + \cdots + \alpha_p^2 \le \left(\delta^2+\tau^2\right)/9$ then $\alpha_{p+1}^2 \ge 3\left(\delta^2+\tau^2\right)/4$. Therefore $(\alpha_p + \alpha_{p+1})$, the coefficient of the x^p term of the polynomial on $[t_1, b]$ is

$$\begin{aligned}\left\|\alpha_p + \alpha_{p+1}\right\|^2 >& \left\|\alpha_{p+1}\right\|^2 - \left\|\alpha_p\right\|^2\\ >& \frac{3\left(\delta^2+\tau^2\right)}{4} - \frac{\left(\delta^2+\tau^2\right)}{4}\\ >& \frac{1}{2}\left(\delta^2+\tau^2\right)\end{aligned}$$

and thus the squared norm of the coefficients of the polynomial on $[t_1, b]$ must also be greater than $\frac{1}{2}\left(\delta^2+\tau^2\right)$ and thus $\frac{1}{n}\sum[g(x_i|\boldsymbol{\theta})]^2\, 1_{[t,b]}(x_i) > \lambda_n\left(\delta^2+\tau^2\right)$ for some $\lambda_n > 0$. The proof for multiple knots is similar, only examining all $\kappa+1$ polynomial sections for one with coefficients with squared norm larger than some fraction of $\left(\delta^2+\tau^2\right)$. □

Lemma 3. *For all* $\delta > 0$, *there exists* $\lambda_n > 0$ *such that for all* $\boldsymbol{\theta} \notin B(\boldsymbol{\theta}_0, \delta)$ *then* $\frac{1}{n}\sum(g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i))^2 > \lambda_n\delta$.

Proof. By the previous lemma, for all $\Delta > 2\tau$ there exists $\exists\, \Lambda_n > 0$ such that for all $\boldsymbol{\theta} \notin B(\boldsymbol{\theta}_0, \Delta)$ then $\frac{1}{n}\sum \left(g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right)^2 > \Lambda_n \Delta$. We now consider the region

$$\mathcal{C} = \text{closure}\left[B\left(\boldsymbol{\theta}_0, \Delta\right) B\left(\boldsymbol{\theta}_0, \delta\right)\right]$$

Assume to the contrary that there exists $\delta > 0$ such that $\forall\, \lambda_n > 0$, $\exists\, \boldsymbol{\theta} \in \mathcal{C}$ such that $\frac{1}{n}\sum \left(g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right)^2 \leq \lambda_n \delta$ and we will seek a contradiction. By the negation, there exists a sequence $\boldsymbol{\theta}_n \in \mathcal{C}$ such that $\frac{1}{n}\sum \left(g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right)^2 \to 0$. But since $\boldsymbol{\theta}_n$ is in a compact space, there exists a subsequence $\boldsymbol{\theta}_{n_k}$ that converges to $\boldsymbol{\theta}_\infty \in \mathcal{C}$ and $\frac{1}{n}\sum \left(g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right)^2 = 0$. But since $\boldsymbol{\theta}_0 \notin \mathcal{C}$ this is a contradiction. □

Corollary 4. *There exists* λ *such that for any* $\delta > 0$ *and* $\boldsymbol{\theta} \notin B(\boldsymbol{\theta}_0, \delta)$

$$\frac{1}{n}\sum_{i=1}^{n} \left[g\left(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i\right) + \sigma_0 \epsilon_i\right]^2 \geq \lambda_n^2 \delta^2 + O_p\left(n^{-1/2}\right) + \sigma_0^2.$$

We now focus our attention on the ratio of the maximum value of a polynomial and its integral.

Lemma 5. *Given a degree* p *polynomial* $g\left(x|\boldsymbol{\alpha}\right)$ *on* $[a, b]$*, then*

$$\frac{\max_{\mathrm{i}\in\{1,\ldots,\mathrm{n}\}}\left[\mathrm{g}\left(\mathrm{x_i}|\boldsymbol{\alpha}\right)\right]^2}{\frac{1}{n}\sum_{i=1}^{n}\left[g\left(x_i|\boldsymbol{\alpha}\right)\right]^2 dx} \leq \frac{\lambda_M^2}{\lambda_{n,m}^2} \to \frac{\lambda_M^2}{\lambda_m^2}$$

for some $\lambda_M, \lambda_m > 0$.

Proof. Since we can write $\left[g\left(x|\boldsymbol{\alpha}\right)\right]^2 = \boldsymbol{\alpha}^T W_x \boldsymbol{\alpha}$ for some nonnegative definite matrix W_x which has a maximum eigenvalue $\lambda_{M,x}$, and because the the maximum eigenvalue is a continuous function in x, let $\lambda_M = \sup \lambda_{M,x}$. Then the maximum of $\left[g\left(x|\boldsymbol{\alpha}\right)\right]^2$ over $x \in [a, b]$ is less than $\|\boldsymbol{\alpha}\|^2 \lambda_M^2$. The denominator is bounded from below by $\|\boldsymbol{\alpha}\|^2 \lambda_{n,m}^2$. □

Lemma 6. *Given a degree* p *spline* $g\left(x|\boldsymbol{\theta}\right)$ *on* $[a, b]$*, then*

$$\frac{\max\left[\mathrm{g}\left(\mathrm{x}|\boldsymbol{\theta}\right)\right]^2}{\int_a^b \left[g\left(x|\boldsymbol{\theta}\right)\right]^2 dx} \leq \frac{\lambda_M^2}{\lambda_m^2}$$

for some $\lambda_M, \lambda_m > 0$.

Proof. Since a degree p spline is a degree p polynomial on different regions defined by the knot-points, and because the integral over the whole interval $[a, b]$ is greater than the integral over the regions defined by the knot-points, we can use the previous lemma on each section and then chose the largest ratio. □

Lemma 7. *Given a degree* p *spline* $g\left(x|\boldsymbol{\theta}\right)$ *on* $[a, b]$ *then*

$$\frac{n^{-1/2}\max_{\mathrm{i}}\left[\epsilon_{\mathrm{i}}\sigma_0 + \mathrm{g}\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0, \mathrm{x_i}\right)\right]^2}{n^{-1}\sum_{i=1}^{n}\left[\epsilon_i\sigma_0 + g\left(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i\right)\right]^2} = O_p\left(1\right) \tag{10.5}$$

uniformly over $\boldsymbol{\theta}$.

Proof. Notice

$$\begin{aligned}\frac{n^{-1/2}\mathrm{max_i}\left[\epsilon_\mathrm{i}\sigma_0 + \mathrm{g}\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,\mathrm{x_i}\right)\right]^2}{n^{-1}\sum_{i=1}^n\left[\epsilon_i\sigma_0 + g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right)\right]^2} &\leq \frac{2n^{-1/2}\mathrm{max_i}\left[\epsilon_\mathrm{i}^2\sigma_0^2\right] + 2\mathrm{n}^{-1/2}\mathrm{max_i}\left[\mathrm{g}\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,\mathrm{x_i}\right)\right]^2}{n^{-1}\sum_{i=1}^n\left[\epsilon_i\sigma_0 + g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right)\right]^2} \\ &= \frac{2\sigma_0^2 n^{-1/2}\mathrm{max_i}\epsilon_\mathrm{i}^2 + \mathrm{max_i}\left[\mathrm{g}\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,\mathrm{x_i}\right)\right]^2}{n^{-1}\sum_{i=1}^n\left[\epsilon_i\sigma_0 + g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right)\right]^2} \\ &= \frac{O_p\left(\frac{\log n}{\sqrt{n}}\right) + \mathrm{max_i}\left[\mathrm{g}\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,\mathrm{x_i}\right)\right]^2}{n^{-1}\sum_{i=1}^n\left[\epsilon_i\sigma_0 + g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right)\right]^2}\end{aligned}$$

and since $n^{-1}\sum_{i=1}^n\left[\epsilon_i\sigma_0 + g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right)\right]^2 \xrightarrow{P} \frac{1}{b-a}\int_a^b\left(g(\boldsymbol{\theta}_0,\boldsymbol{\theta},x)\right)^2\,dx + \sigma_0^2$, and lemma 8 bounds the ratio of the terms that involve $\boldsymbol{\theta}$, this ratio is bounded in probability uniformly over $\boldsymbol{\theta}$. □

B.5 Assumptions B1

Returning to assumption B1, we now consider $\boldsymbol{\xi} \notin B(\boldsymbol{\xi}_0,\delta)$ and

$$\begin{aligned}L_n\left(\boldsymbol{\xi}\right) &= \sum\log\left\{\frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{-1}{2\sigma}\sum\left(y_i - N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}\right)^2\right]\right\} \\ &= -\frac{n}{2}\log\left(2\pi\right) - n\log\sigma - \frac{1}{2\sigma}\sum\left[y_i - N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}\right]^2 \\ &= -\frac{n}{2}\log\left(2\pi\right) - n\log\sigma - \frac{1}{2\sigma}\sum\left[N(x_i,\boldsymbol{t}_0)^T\boldsymbol{\alpha}_0 + \sigma_0\epsilon_i - N(x_i,\boldsymbol{t})^T\boldsymbol{\alpha}\right]^2 \\ &= -\frac{n}{2}\log\left(2\pi\right) - n\log\sigma - \frac{1}{2\sigma}\sum\left[g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right) + \sigma_0\epsilon_i\right]^2\end{aligned}$$

and therefore

$$\begin{aligned}&\frac{1}{n}\left(L_n(\boldsymbol{\xi}) - L_n(\boldsymbol{\xi}_0)\right) \\ &\quad= -\log\sigma - \frac{1}{2n\sigma^2}\sum\left[g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right) + \sigma_0\epsilon_i\right]^2 + \log\sigma_0 + \frac{1}{2n\sigma_0}\sum\left[g\left(\boldsymbol{\theta}_0,\boldsymbol{\theta}_0,x_i\right) + \sigma_0\epsilon_i\right]^2 \\ &\quad= \log\frac{\sigma_0}{\sigma} - \frac{1}{2n\sigma^2}\sum\left[g\left(\boldsymbol{\theta},\boldsymbol{\theta}_0,x_i\right) + \sigma_0\epsilon_i\right]^2 + \frac{1}{2n\sigma_0^2}\sum\left[\sigma_0\epsilon_i\right]^2 \\ &\quad= \log\frac{\sigma_0}{\sigma} - \frac{\left(\lambda_n\left(\boldsymbol{\theta},\boldsymbol{\theta}_0\right)\right)^2}{2\sigma^2} - \frac{\sigma_0^2}{2\sigma^2} + \frac{1}{2n}\sum\left[\epsilon_i\right]^2\end{aligned}$$

where

$$[\lambda_n(\boldsymbol{\theta}, \boldsymbol{\theta}_0)]^2 = \frac{1}{n}\sum [g(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i) + \sigma_0 \epsilon_i]^2 - \sigma_0^2$$

which converges in probability to $\frac{1}{b-a}\int_a^b [g(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x)]^2 \, dx$. The function goes to $-\infty$ as $\sigma \to 0$ and $\sigma \to \infty$. Taking the derivative

$$\frac{d}{d\sigma}\left[\log\frac{\sigma_0}{\sigma} - \frac{1}{2\sigma^2}\left[(\lambda_n)^2 + \sigma_0^2\right] + \frac{1}{2n}\sum \epsilon_i^2\right] = -\frac{1}{\sigma} + \frac{1}{\sigma^3}\left[(\lambda_n)^2 + \sigma_0^2\right]$$

and setting it equal to zero yields a single critical point of at $\sigma^2 = \left[(\lambda_n)^2 + \sigma_0^2\right]$ which results in a maximum of

$$\log\left(\frac{\sigma_0}{\sqrt{(\lambda_n)^2 + \sigma_0^2}}\right) - \frac{1}{2} + \frac{1}{2}n^{-1}\sum \epsilon_i^2 \tag{10.6}$$

which bounded away from zero in probability for $\boldsymbol{\xi} \notin B(\boldsymbol{\xi}_0, \delta)$

B.6 Assumption C1

Assumption C1 is

$$\inf_{\boldsymbol{\xi} \notin B(\boldsymbol{\xi}_0, \delta)} \frac{\min_{\mathrm{i}=1\ldots\mathrm{n}} \mathrm{L}(\boldsymbol{\xi}, \mathrm{X_i})}{\left|L_n(\boldsymbol{\xi}) - L_n(\boldsymbol{\xi}_0)\right|} \xrightarrow{P_{\boldsymbol{\xi}_0}} 0$$

First notice

$$\begin{aligned} L(\boldsymbol{\xi}, Y_i) &= -\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}\left(Y_i - N(x_i, \boldsymbol{t})^T\boldsymbol{\alpha}\right)^2 \\ &= -\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}\left(\epsilon_i\sigma_0 + N(x_i, \boldsymbol{t}_0)^T\boldsymbol{\alpha}_0 - N(x_i, \boldsymbol{t})^T\boldsymbol{\alpha}\right)^2 \\ &= -\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}\left(\epsilon_i\sigma_0 + g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right)^2 \end{aligned}$$

and we consider $\mathcal{C} = \left\{\boldsymbol{\xi} : \boldsymbol{\xi} \notin B(\boldsymbol{\xi}_0, \delta)\right\}$. Define

$$\begin{aligned} f_n(\boldsymbol{\xi}) &= \frac{\min \mathrm{L}(\boldsymbol{\xi}, \mathrm{Y_i})}{\left|L_n(\boldsymbol{\xi}) - L_n\left(\boldsymbol{\xi}_0\right)\right|} \\ &= \frac{-\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}\max\left[\epsilon_\mathrm{i}\sigma_0 + \mathrm{g}(\boldsymbol{\theta}_0, \boldsymbol{\theta}, \mathrm{x_i})\right]^2}{n \cdot \frac{1}{n}\left|L_n(\boldsymbol{\xi}) - L_n\left(\boldsymbol{\xi}_0\right)\right|} \end{aligned}$$

and notice that the denominator is bounded away from 0 by 10.6.

$$f_n(\boldsymbol{\xi}) = \frac{-\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}\max\left[\epsilon_i\sigma_0 + g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right]^2}{-n \cdot \frac{1}{n}\left(L_n(\boldsymbol{\xi}) - L_n(\boldsymbol{\xi}_0)\right)}$$

$$= \frac{\frac{1}{\sqrt{n}}\left[-\frac{1}{2}\log(2\pi) - \log\sigma - \frac{1}{2\sigma^2}\max\left[\epsilon_i\sigma_0 + g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right]^2\right]}{-\sqrt{n} \cdot \frac{1}{n}\left[n\log\frac{\sigma_0}{\sigma} - \frac{1}{2\sigma^2}\sum[g(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i) + \sigma_0\epsilon_i]^2 + \frac{1}{2}\sum\epsilon_i^2\right]}$$

$$= \frac{1}{\sqrt{n}} \cdot \frac{-\frac{1}{2\sqrt{n}}\log(2\pi) - \frac{1}{\sqrt{n}}\log\sigma - \frac{1}{2\sqrt{n}\sigma^2}\max\left[\epsilon_i\sigma_0 + g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right]^2}{-\log\frac{\sigma_0}{\sigma} + \frac{1}{2n\sigma^2}\sum[g(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i) + \sigma_0\epsilon_i]^2 - \frac{1}{2n}\sum\epsilon_i^2}$$

$$= \frac{1}{\sqrt{n}}\left[\frac{-\frac{1}{2\sqrt{n}}\log(2\pi)}{-\log\frac{\sigma_0}{\sigma} + \frac{1}{2n\sigma^2}\sum[g(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i) + \sigma_0\epsilon_i]^2 - \frac{1}{2n}\sum\epsilon_i^2}\right.$$

$$\left. + \frac{-\frac{1}{\sqrt{n}}\log\sigma - \frac{1}{2\sqrt{n}\sigma^2}\max\left[\epsilon_i\sigma_0 + g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right]^2}{-\log\frac{\sigma_0}{\sigma} + \frac{1}{2n\sigma^2}\sum[g(\boldsymbol{\theta}, \boldsymbol{\theta}_0, x_i) + \sigma_0\epsilon_i]^2 - \frac{1}{2n}\sum\epsilon_i^2}\right]$$

We consider the infimums of the terms inside the brackets separately.

For the first term, since the denominator is bounded in probability above 0 uniformly in $\boldsymbol{\theta}$, and the numerator goes to zero, the infimum of the first term goes to 0 in probability.

The second term is uniformly bounded over $\boldsymbol{\theta}$ by lemma 9. Notice that the numerator is

$$-\frac{1}{\sqrt{n}}\log\sigma - \frac{1}{2\sqrt{n}\sigma^2}\max\left[\epsilon_i\sigma_0 + g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right]^2$$

$$\geq -\frac{1}{\sqrt{n}}\log\sigma - \frac{\max[\epsilon_i\sigma_0]^2}{\sqrt{n}\sigma^2} - \frac{\max\left[g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right]^2}{\sqrt{n}\sigma^2}$$

$$= -\frac{1}{\sqrt{n}}\log\sigma - \frac{\sigma_0^2\, O_p(\log n)}{\sqrt{n}\sigma^2} - \frac{\max\left[g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right]^2}{\sqrt{n}\sigma^2}$$

$$\geq \frac{-\log n}{\sqrt{n}}\log\sigma - \frac{\sigma_0^2\, O_p(\log n)}{\sqrt{n}\sigma^2} - \frac{\max\left[g(\boldsymbol{\theta}_0, \boldsymbol{\theta}, x_i)\right]^2}{\sqrt{n}\sigma^2}$$

and all three terms of the numerator converge to 0 for every σ. Therefore, for $\sigma \in [0, d]$ for some large d, the infimum converges to 0. For $\sigma > d$, the $\log\sigma$ terms dominate and the infimum occurs at $\sigma = d$ which also converges to 0. Therefore

$$\inf_{\boldsymbol{\xi} \notin B(\boldsymbol{\xi}_0, \delta)} \frac{\min L(\boldsymbol{\xi}, Y_i)}{\left|L_n(\boldsymbol{\xi}) - L_n(\boldsymbol{\xi}_0)\right|} \xrightarrow{P} 0.$$

B.7 Assumptions C2

Finally we turn our attention to the Jacobian. Recall that the Jacobian is

$$J_0\left(\boldsymbol{y}_0, \boldsymbol{\xi}\right) = \left| \frac{1}{\sigma^2} p^{\kappa} \det \left[\begin{array}{ccc} \boldsymbol{B}_{\boldsymbol{\alpha}} & \boldsymbol{B}_{\boldsymbol{t}} & \boldsymbol{B}_{\sigma^2} \end{array} \right] \right|$$

where

$$\boldsymbol{B}_{\boldsymbol{\alpha}} = \begin{bmatrix} 1 & x_{(1)} & \dots & x_{(1)}^p & (x_{(1)} - t_1)_+^p & \dots & (x_{(1)} - t_\kappa)_+^p \\ \vdots & \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{(l)} & \dots & x_{(l)}^p & (x_{(l)} - t_1)_+^p & \dots & (x_{(l)} - t_\kappa)_+^p \end{bmatrix},$$

$$\boldsymbol{B}_{\boldsymbol{t}} = \begin{bmatrix} \alpha_{1+p+1} \left(x_{(1)} - t_1\right)_+^{p-1} I\left(x_{(1)} - t_1\right) & \dots & \alpha_{1+p+\kappa} \left(x_{(1)} - t_\kappa\right)_+^{p-1} I\left(x_{(1)} - t_\kappa\right) \\ \vdots & \ddots & \vdots \\ \alpha_{1+p+1} \left(x_{(l)} - t_1\right)_+^{p-1} I\left(x_{(l)} - t_1\right) & \dots & \alpha_{1+p+\kappa} \left(x_{(l)} - t_\kappa\right)_+^{p-1} I\left(x_{(l)} - t_\kappa\right) \end{bmatrix},$$

and

$$\boldsymbol{B}_{\sigma^2} = \begin{bmatrix} -\frac{1}{2}\left(y_{(1)} - g(x_{(1)}|\boldsymbol{\theta})\right) \\ \vdots \\ -\frac{1}{2}\left(y_{(l)} - g(x_{(l)}|\boldsymbol{\theta})\right) \end{bmatrix}.$$

Following the notation of Yeo and Johnson, we suppress parenthesis and 0 subscripts. We consider the $\boldsymbol{\xi}$ in compact space $\bar{B}(\boldsymbol{\xi}_0, \delta)$. We notice that for $\delta < \sigma^{-2}$ that $J(\boldsymbol{y}; \boldsymbol{\xi}) \leq \delta^{\kappa+1} p^{\kappa} g(\boldsymbol{y})$ for some $g(\boldsymbol{y})$ because $\boldsymbol{B}_{\boldsymbol{\alpha}}$ and $\boldsymbol{B}_{\boldsymbol{t}}$ are functions of $\boldsymbol{x}, \boldsymbol{t}$ which are bounded.

We let S_M^l be the unit square in $\mathbb{R}^l$ of radius M.

Finally, we notice that $J_j(y_1, \dots, y_j; \boldsymbol{\xi}) = E\left[J\left(y_1, \dots, y_j, Y_{j+1}, \dots, Y_l; \boldsymbol{\xi}\right)\right]$ is a polynomial in $\boldsymbol{\theta}$ scaled by σ^2, which is equicontinuous on compacts of $\boldsymbol{\xi}$ where σ is bounded away from 0.

Appendix C: Full Simulation Results

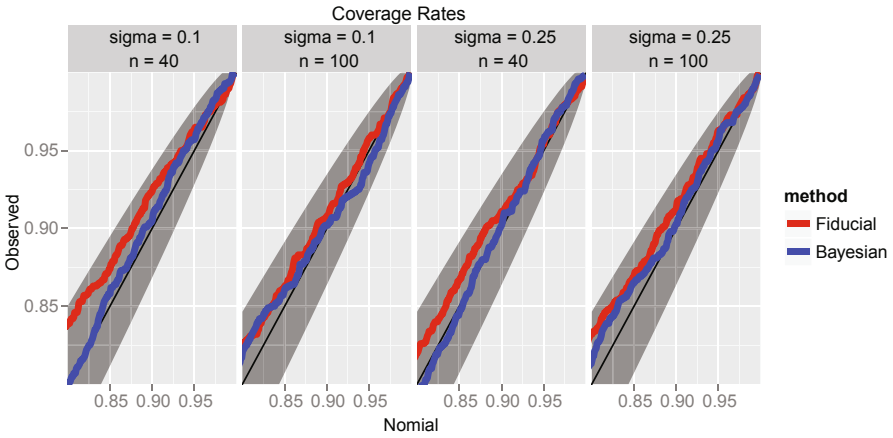


Fig. 10.4 Coverage rates for the single knot scenario. The color (*red, blue*) represents the method (*fiducial, Bayesian)*

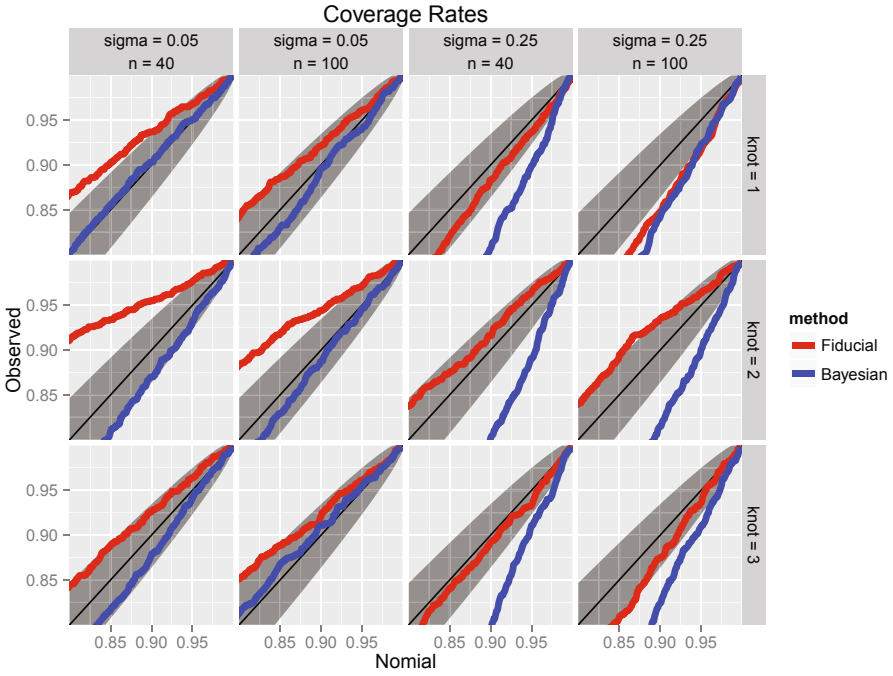


Fig. 10.5 Coverage rates for the three knot "Simple" scenario. The color (*red, blue)* represents the method *(fiducial, Bayesian).* The topmost panel is the coverage of knot one in the $sigma = 0.1, n = 40$ simulation

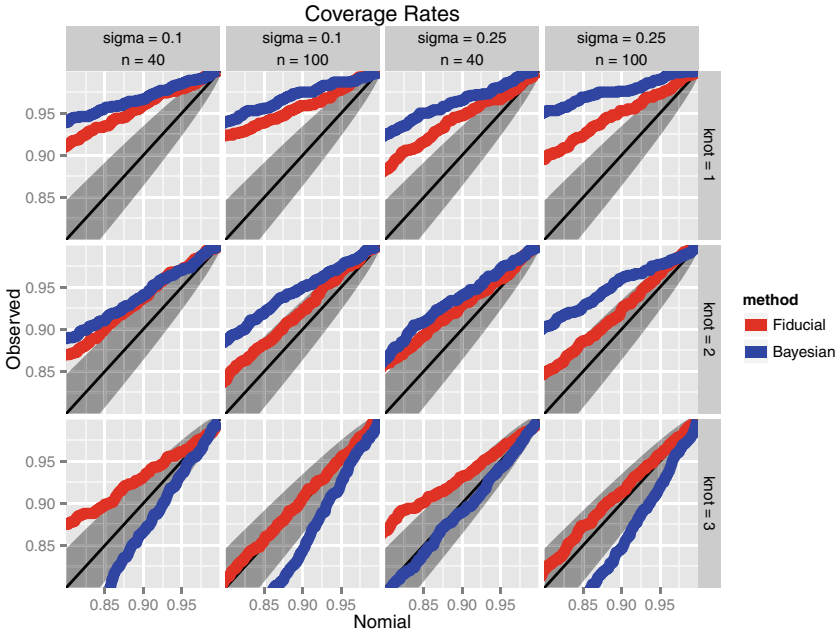


Fig. 10.6 Coverage rates for the three knot "Clustered" scenario. The color *(red, blue)* represents the method *(fiducial, Bayesian)*. The topmost panel is the coverage of knot one in the $sigma = 0.1, n = 40$ simulation

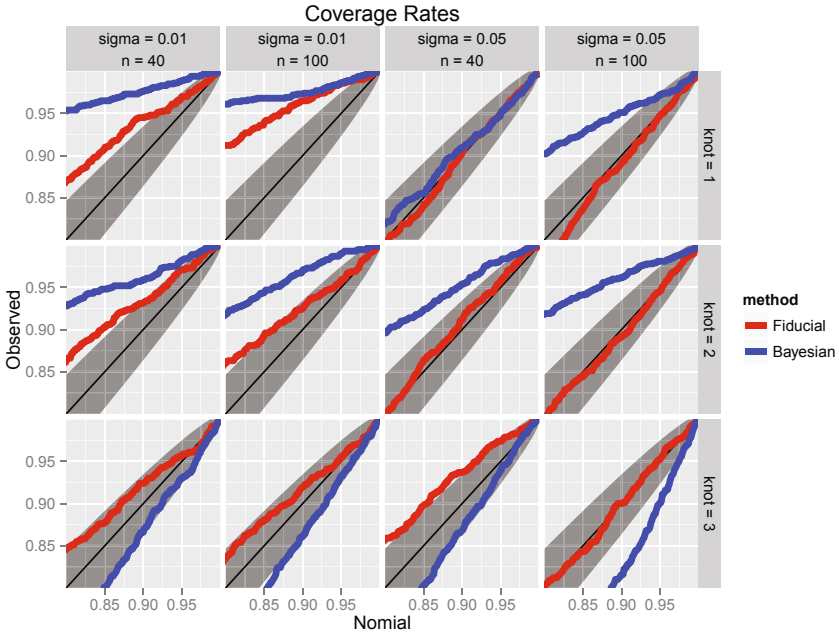


Fig. 10.7 Coverage rates for the three knot "Subtle" scenario. The color *(red, blue)* represents the method *(fiducial, Bayesian)*. The topmost panel is the coverage of knot one in the $sigma = 0.1, n = 40$ simulation

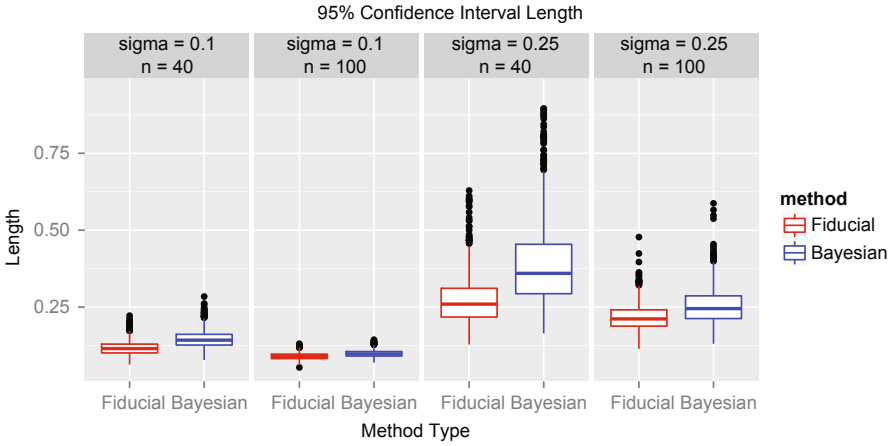


Fig. 10.8 Confidence interval lengths for the single knot scenario. The color *(red, blue)* represents the method *(fiducial, Bayesian)*

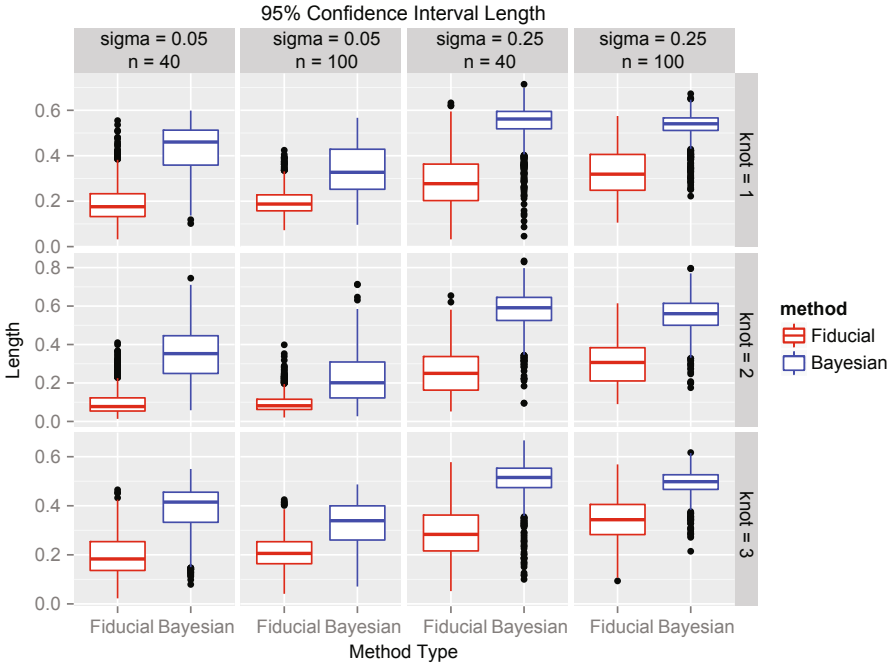


Fig. 10.9 Confidence interval lengths for the three knot "Simple" scenario. The color *(red, blue)* represents the method *(fiducial, Bayesian)*. The topmost panel is the coverage of knot one in the $sigma = 0.1, n = 40$ simulation

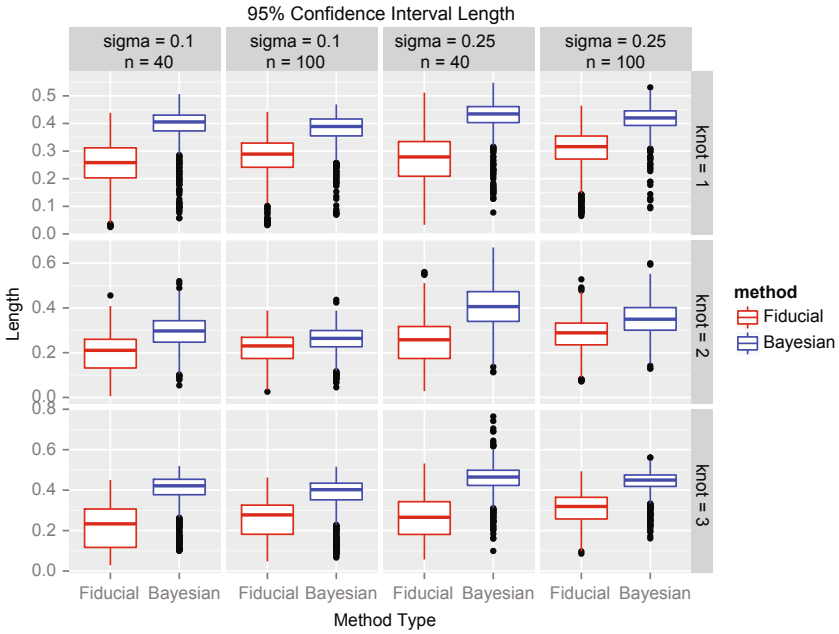


Fig. 10.10 Confidence interval lengths for the three knot "Clustered" scenario. The color (*red, blue)* represents the method *(fiducial, Bayesian)*. The topmost panel is the coverage of knot one in the $sigma = 0.1, n = 40$ simulation

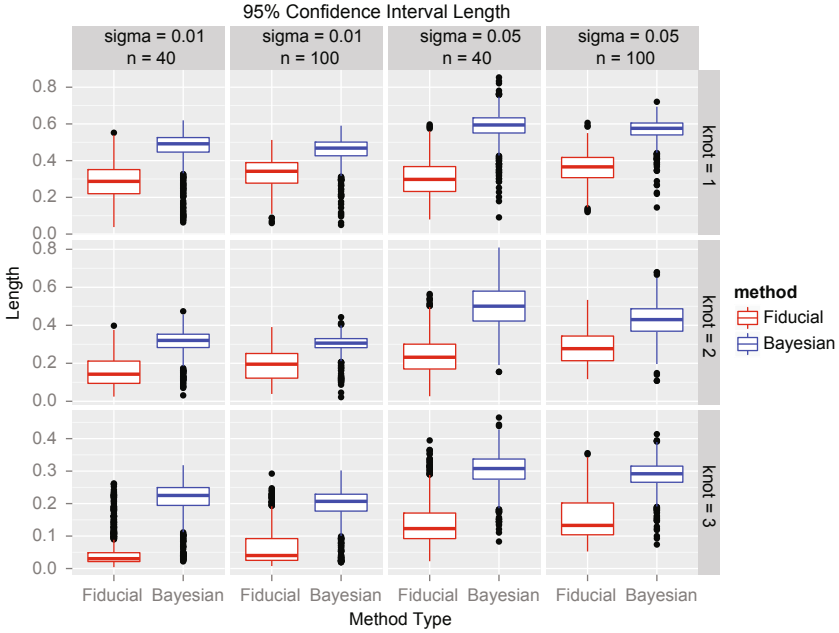


Fig. 10.11 Confidence interval lengths for the three knot "Subtle" scenario. The color (*red, blue*) represents the method (*fiducial, Bayesian*). The topmost panel is the coverage of knot one in the $sigma = 0.1, n = 40$ simulation

References

Cisewski J, Cisewski J, Hannig J (2012) Generalized fiducial inference for normal linear mixed models Ann Stat 40:2102–2127

DiMatteo I, Genovese C, Kass R, Robert E (2001) Bayesian curve-fitting with free-knot splines. Biometrika 88:1055–1071

Lidong E, Hannig J, Iyer H (2008) Fiducial intervals for variance components in an un-balanced two-component normal mixed linear model. J Am Stat Assoc 103:854–865

Fisher RA (1930) Inverse probability. Proc Camb Philos Soc xxvi:528–535

Ghosh JK, Ramamoorthi RV (2003). Bayesian Nonparametrics. Springer-Verlag, New York

Hannig J (2009) On generalized fiducial inference. Statist Sinica 19:491–544

Hannig J (2013) Generalized fiducial inference via discretization. Stat Sinica 23:489–514

Hannig J, Iyer H, Patterson P (2006) Fiducial generalized confidence intervals. J Am Stat Assoc 101:254–269. 10.1198/016214505000000736

Hannig J, Lee TCM (2009). Generalized fiducial inference for wavelet regression. Biometrika 96:847–860. 10.1093/biomet/asp050

Lehmann EL, George C, (1998) Theory of point estimation. Springer, New York

Muggeo VMR (2003) Estimating regression models with unknown break-points. Stat Med 22:3055–3071

Muggeo VMR (2008) Segmented: an R package to fit regression models with broken-line relationships. R News, 8, 1: 20–25.

Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression. Cambridge University Press, Cambridge

Sonderegger DL, Wang H, Clements WH, Noon BR (2009) Using SiZer to detect thresholds in ecological data. Front Ecol Environ 7:190–195 doi:10.1890/070179

Toms JD, Lesperance ML (2003) Piecewise regression: a tool for identifying ecological thresholds. Ecology 84:2034–2041

van der Varrt AW (1998) Asymptotic statistics. Cambridge University Press, Cambridge

Wandler DV, Hannig J (2011) Generalized fiducial confidence intervals for extremes. Extremes 15:67–87. 10.1007/s10687-011-0127-9

Wandler DV, Hannig J (2012) A fiducial approach to multiple comparisons. J Stat Plan Infer 142:878–895. 10.1016/j.jspi.2011.10.011

Weerahandi S (1993) Generalized confidence intervals. J Am Stat Assoc 88(423):899–905

Yeo IK, Johnson RA (2001) A uniform strong law of large numbers for U-statistics with application to transforming to near symmetry. Stat Probab Lett 51 63–69

Chapter 11
An Empirical Characteristic Function Approach to Selecting a Transformation to Symmetry

In-Kwon Yeo and Richard A. Johnson

11.1 Introduction

Many statistical techniques are based on assumption about the form of population distribution. The validity of those results may depend on the assumed conditions being satisfied. When the observed data seriously violate these assumptions, transformation of data can improve the agreement with the assumption about underlying distribution. As an objective way of determining a transformation was introduced by Box and Cox (1964), transformation of data has widely used in applied statistics as well as theoretical statistics.

Generally, a main goal of transforming data is to enhance the normality and homoscedasticity of data. Box and Cox (1964) discussed estimating transformation parameter by the maximum likelihood approach and by a Bayesian method. It is well known that, under the normality assumption, the maximum likelihood estimator of the Box–Cox transformation parameter is very sensitive to outliers, see (Andrews 1971). Carroll (1980) proposed a robust method for selecting a power transformation to achieve approximate normality in a linear model.

Robust techniques sometimes require symmetry rather than normality of data. Hinkley (1975) and Taylor (1985) suggested methods for estimating the transformation parameter in the Box–Cox transformation when the goal is to obtain approximate symmetry. Yeo and Johnson (2001) and Yeo (2001) introduced an M-estimator which is obtained by minimizing the integrated square of the imaginary part of the empirical characteristic function of Yeo–Johnson transformed data.

Many authors including Koutrouvelis (1980); Koutrouvelis and Kellermeier (1981); Fan (1997); Klar and Meintanis (2005), and Jimenez-Gamero et al. (2009)

I.-Kwon Yeo (✉)
Department of Statistics, Sookmyung women's University,
Seoul 140-742, Seoul, Korea
e-mail: inkwon@sm.ac.kr

R. A. Johnson
Department of Statistics, University of Wisconsin-Madison,
53706 Madison, WI, USA
e-mail: rich@stat.wisc.edu

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_11,

have proposed the goodness-of-fit test statistics based on measuring differences between the empirical characteristic function and the characteristic function in the null hypothesis.

Our estimators are obtained by minimizing a squared distance between the empirical characteristic function of the transformed data and the target characteristic function. Specifically, we minimize the integral of the squared modulus of the difference of the two characteristic functions multiplied by a weight function. This estimation procedure for a vector-valued parameter can be viewed as solving estimating equations based on a U-statistic, see Lee (1990). According to Yeo et al. (2013), the estimator by the empirical characteristic function approach is still sensitive, but less sensitive than the maximum likelihood estimate, to an outlier when the target distribution is normal.

11.2 Estimation

Let $\psi(x, \lambda)$ be a general class of transformations which are indexed by the transformation parameter λ. Examples include the families introduced by (Box and Cox 1964; John and Draper 1980; Burbidge et al. 1988, and Yeo and Johnson 2000). Based on calculations of the relative skewness by van Zwet (1964), the Box–Cox transformation and the Yeo–Johnson transformation can improve the symmetry of data and be applied to skewed data. By contrast, the modulus transformation by John and Draper (1980) and the inverse hyperbolic sine transformation by Johnson (1949) and Burbidge et al. (1988) are useful to reduce the kurtosis of heavy-tailed data. Hence, we focus on the Box–Cox transformation, for $x > 0$,

$$\psi(x, \lambda) = \begin{cases} \left(x^{\lambda} - 1\right)/\lambda, & \lambda \neq 0 \\ \log(x), & \lambda = 0 \end{cases}$$

and the Yeo–Johnson transformation

$$\psi(x, \lambda) = \begin{cases} \left\{(x+1)^{\lambda} - 1\right\}/\lambda, & \lambda \neq 0,\ x \geq 0 \\ \log(x+1), & \lambda = 0,\ x \geq 0 \\ -\left\{(-x+1)^{2-\lambda} - 1\right\}/(2-\lambda), & \lambda \neq 2,\ x < 0 \\ -\log(-x+1), & \lambda = 2,\ x < 0 \end{cases}$$

and theorems derived below are based on these transformations. Note that, for these transformations, $\partial^k \psi(x, \lambda)/\partial x^k$ and $\partial^l \psi(x, \lambda)/\partial \lambda^l$ are continuous in (x, λ) for $k = 0, 1, 2$ and $l = 0, 1, \ldots$ and $\psi(x, \lambda)$ is increasing in both x and λ.

Let $X_1, \ldots, X_n$ be independent and identically distributed random variables with distribution function $F(\cdot)$.

Assumption 1 *There exists a λ for which the distribution of $\psi(X, \lambda)$ is a location-scale family with parameters μ and σ and symmetric about μ.*

Usually, in a given example, it may not be possible to select λ so that Assumption 1 holds. Nevertheless, we make that assumption similar to the assumption of normality in Box and Cox (1964).

Let $\phi(t)$ be the characteristic function of the standardized target distribution and let $\phi_n(\boldsymbol{\theta}, t)$ be the empirical characteristic function of standardized transformed variables $Z_j(\boldsymbol{\theta}) = \left\{\psi(X_j, \lambda) - \mu\right\}/\sigma$, $j = 1, \ldots, n$, that is,

$$\phi_n(\boldsymbol{\theta}, t) = \frac{1}{n}\sum_{j=1}^{n} \exp\left(itZ_j(\boldsymbol{\theta})\right) = \phi_{cn}(\boldsymbol{\theta}, t) + i\phi_{sn}(\boldsymbol{\theta}, t),$$

where $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3)^T = (\lambda, \mu, \sigma)^T$ denotes the vector of parameters of interest and $\phi_{cn}(\boldsymbol{\theta}, t) = n^{-1}\sum_{j=1}^{n} \cos\left(tZ_j(\boldsymbol{\theta})\right)$ and $\phi_{sn}(\boldsymbol{\theta}, t) = n^{-1}\sum_{j=1}^{n} \sin\left(tZ_j(\boldsymbol{\theta})\right)$.

Yeo and Johnson (2001) and Yeo (2001) studied selecting transformation so that the transformed variable is nearly symmetrically distributed about μ. They selected λ and μ to make the integrated square of the imaginary part of the empirical characteristic function of $\psi(X_1, \lambda), \ldots, \psi(X_n, \lambda)$ with factor $\exp(-it\mu)$ minimized,

$$\int \mathrm{Im}\{\exp(-it\mu)\phi_n(\lambda, t)\}^2\, dG(t) = \int \left\{\frac{1}{n}\sum_{j=1}^{n} \sin\left(t(\psi(\lambda, X_j) - \mu)\right)\right\}^2 dG(t),$$

where $\phi_n(t) = n^{-1}\sum_{j=1}^{n} \exp\left(it\psi(X_j, \lambda)\right)$ and $G(\cdot)$ is a symmetric distribution function.

In this chapter, we propose to transform X according to $Z(\boldsymbol{\theta})$ and then to select $\boldsymbol{\theta}$ to minimize an integrated weighted version of the distance between the empirical characteristic function and a real-valued target characteristic function, $\phi(t)$. Specifically, we minimize,

$$\begin{aligned}\varphi_n(\boldsymbol{\theta}) &= ||\phi_n(\boldsymbol{\theta}) - \phi||_w^2 \\ &= \int_{-\infty}^{\infty} \{\phi_n(\boldsymbol{\theta}, t) - \phi(t)\}\,\overline{\{\phi_n(\boldsymbol{\theta}, t) - \phi(t)\}} w(t)\, dt,\end{aligned}$$

where $\overline{\{\phi_n(\boldsymbol{\theta}, t) - \phi(t)\}}$ denotes the complex conjugate and $w(t)$ is a nonnegative real-valued weight function. We assume that $w(t)$ is nonnegative and symmetric about zero and $\int w(t)\, dt < \infty$. Since the target distribution is assumed to be symmetric about zero, the characteristic function $\phi(t)$ is real-valued so that $\overline{\phi(t)} = \phi(t)$. Therefore,

$$\begin{aligned}\varphi_n(\boldsymbol{\theta}) &= \int w(t)\phi_n(\boldsymbol{\theta}, t)\overline{\phi_n(\boldsymbol{\theta}, t)}\, dt \\ &\quad - \int w(t)\phi(t)\left\{\phi_n(\boldsymbol{\theta}, t) + \overline{\phi_n(\boldsymbol{\theta}, t)}\right\} dt + \int w(t)\phi(t)^2\, dt \\ &\propto \frac{1}{n}\sum_{j<k} \int w(t) \cos\left(t\left\{Z_j(\boldsymbol{\theta}) - Z_k(\boldsymbol{\theta})\right\}\right) dt \\ &\quad - \sum_{j=1}^{n} \int w(t)\phi(t) \cos\left(tZ_j(\boldsymbol{\theta})\right) dt. \qquad (11.1)\end{aligned}$$

The behavior in neighborhood of zero is important for characteristic functions. As in Szekely et al. (2007), we may choose $w(t)$ equal to t^{-2} on some interval containing

zero and define integrals as the principal values. The integral on 0 to ∞ is the limit as $\epsilon \to 0$ of the integral over $(\epsilon, \epsilon^{-1})$. Under this preferred weight function, the estimation procedure involves some difficult numerical integrations and the proof of the asymptotic results is somewhat cumbersome. Instead, we impose moment conditions on $w(t)$ below.

Let $\varphi(\boldsymbol{\theta})$ be the integrated distance between the true characteristic function of $Z(\boldsymbol{\theta})$ and $\phi(t)$, that is,

$$\varphi(\boldsymbol{\theta}) = ||\phi(\boldsymbol{\theta}) - \phi||_w^2 = \int_{-\infty}^{\infty} \{\phi(\boldsymbol{\theta}, t) - \phi(t)\}\, \overline{\{\phi(\boldsymbol{\theta}, t) - \phi(t)\}} w(t)\, dt,$$

where $\phi(\boldsymbol{\theta}, t) = E[\exp(itZ(\boldsymbol{\theta}))]$ denotes the characteristic function of the standardized transformed variable $Z(\boldsymbol{\theta})$. The distribution of $Z(\boldsymbol{\theta})$ is equivalent to the target distribution if and only if $\varphi(\boldsymbol{\theta})$ is zero. Hence, a reasonable approach to estimation is to select the value $\hat{\boldsymbol{\theta}} = (\hat{\lambda}, \hat{\mu}, \hat{\sigma})^T$ which minimizes $\varphi_n(\boldsymbol{\theta})$, that is $\hat{\boldsymbol{\theta}} = \arg\min \varphi_n(\boldsymbol{\theta})$.

11.3 Asymptotic Theory

Assume that the parameter space $\boldsymbol{\Theta}$ is a compact set of the form

$$\boldsymbol{\Theta} = \{\boldsymbol{\theta} \mid a_i \le \theta_i \le b_i \text{ where } 0 < a_3 \text{ and } |a_i|, |b_i| < \infty \text{ for } i = 1, 2, 3\}. \tag{11.2}$$

Theorem 1. *Suppose that the parameter space $\boldsymbol{\Theta}$ is a compact set such as (11.2) and $w(t)$ is nonnegative and symmetric about zero and $\int w(t)\, dt < \infty$. Then, $\varphi_n(\boldsymbol{\theta}) \xrightarrow{a.s.} \varphi(\boldsymbol{\theta})$ uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\varphi(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$.*

Proof. Since $|\phi_{cn}(\boldsymbol{\theta}, t)| \le 1$, $|\phi_{sn}(\boldsymbol{\theta}, t)| \le 1$, and $|\phi(t)| \le 1$, it is clear that

$$\begin{aligned} \varphi_n(\boldsymbol{\theta}) &= \int_{-\infty}^{\infty} \{\phi_n(\boldsymbol{\theta}, t) - \phi(t)\}\, \overline{\{\phi_n(\boldsymbol{\theta}, t) - \phi(t)\}} w(t)\, dt \\ &= \int_{-\infty}^{\infty} \left\{(\phi_{cn}(\boldsymbol{\theta}, t) - \phi(t))^2 + \phi_{sn}(\boldsymbol{\theta}, t)^2\right\} w(t)\, dt \qquad (11.3) \\ &\le 5 \int_{-\infty}^{\infty} w(t)\, dt < \infty. \end{aligned}$$

We begin by defining

$$\begin{aligned} \eta(z_1, z_2; \boldsymbol{\theta}) = \int_{-\infty}^{\infty} &\{(\cos(tz_1(\boldsymbol{\theta})) - \phi(t))\,(\cos(tz_2(\boldsymbol{\theta})) - \phi(t)) \qquad (11.4) \\ &+ \sin(tz_1(\boldsymbol{\theta}))\sin(tz_2(\boldsymbol{\theta}))\}\, w(t)\, dt, \end{aligned}$$

and then have, from (11.3),

$$\varphi_n(\boldsymbol{\theta}) = \frac{1}{n^2} \sum_{j=1}^{n} \sum_{k=1}^{n} \eta(Z_j, Z_k; \boldsymbol{\theta}) = \frac{n-1}{n} U_n(\boldsymbol{\theta}) + \frac{1}{n^2} \sum_{j=1}^{n} \eta(Z_j, Z_j; \boldsymbol{\theta}) \tag{11.5}$$

where

$$U_n(\boldsymbol{\theta}) = \binom{n}{2}^{-1} \sum_{j<k} \eta(Z_j, Z_k; \boldsymbol{\theta}).$$

Letting $S_M = [-M, M]$ and $\Omega_M = S_M \times S_M \times \boldsymbol{\theta}$, we can conclude that, since $\eta(z_1, z_2; \boldsymbol{\theta})$ is bounded and continuous in $(z_1, z_2; \boldsymbol{\theta}) \in \Omega_M$, $\eta_1(z_1, z_2; \boldsymbol{\theta})$ is equicontinuous in $\boldsymbol{\Theta}$. Furthermore, the uniform strong law of large numbers of U-statistics in Yeo and Johnson (2001) ensures that $U_n(\boldsymbol{\theta}) \xrightarrow{a.s.} E[\eta(Z_1, Z_2; \boldsymbol{\theta})] = \eta(\boldsymbol{\theta})$ uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and $\eta(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Since, by the uniform strong law of large numbers in Rubin (1956),

$$\frac{1}{n} \sum_{j=1}^{n} \eta(Z_j, Z_j; \boldsymbol{\theta}) \xrightarrow{a.s.} E\left[\eta(Z_j, Z_j; \boldsymbol{\theta})\right] \tag{11.6}$$

uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and this limit function in (11.6) is continuous in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$, the last term in (11.5) can be neglected. Therefore, as claimed

$$\varphi_n(\boldsymbol{\theta}) \xrightarrow{a.s.} \eta(\boldsymbol{\theta})$$

uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ and the limit is continuous in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$.

Finally we note that

$$\varphi(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} \left\{ (E[\cos(t Z_1(\boldsymbol{\theta}))] - \phi(t))^2 + E[\sin(t Z_1(\boldsymbol{\theta}))]^2 \right\} w(t)\, dt = \eta(\boldsymbol{\theta}) \tag{11.7}$$

because Z_1 and Z_2 are independent and identically distributed. □

Lemma 1. *Let $\{g_n(\boldsymbol{\theta})\}$ be a sequence of random functions defined on a probability space and depend on a compact set, $\boldsymbol{\theta} \in \boldsymbol{\Theta}$. Suppose that*

(i) There exists a continuous function $g(\boldsymbol{\theta})$ defined on $\boldsymbol{\Theta}$ such that $g_n(\boldsymbol{\theta}) \xrightarrow{a.s.} g(\boldsymbol{\theta})$ uniformly in $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,
(ii) $T(\boldsymbol{\theta})$ has a unique minimum at $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$.

Then, $\hat{\boldsymbol{\theta}}_n = \arg\min_n(\boldsymbol{\theta})$ is a strongly consistent estimator of $\boldsymbol{\theta}_0$.
Since it is a standard result, we omit the proof.

Theorem 2. *Suppose the conditions of Theorem 1 hold and $\varphi(\boldsymbol{\theta})$ has a unique global minimum at $\boldsymbol{\theta}_0 = (\lambda_0, \mu_0, \sigma_0)^T$. Then, $\hat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}_0$.*

Proof. Since, according to Theorem 1, $\varphi_n(\boldsymbol{\theta}) \xrightarrow{a.s.} \varphi(\boldsymbol{\theta})$ uniformly in $\boldsymbol{\theta}$ and $\varphi(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$ and, by assumption, $\boldsymbol{\theta}_0$ is unique minimizer of $\varphi(\boldsymbol{\theta})$, Lemma 1 allows us to conclude that $\hat{\boldsymbol{\theta}} \xrightarrow{a.s.} \boldsymbol{\theta}_0$. □

Before stating asymptotic normality, we introduce some notations. For any function $g(\boldsymbol{\theta})$,

$$\nabla g(\boldsymbol{\theta}_*) = \left(\left. \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_j} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*} \right) \quad \text{and} \quad \nabla^2 g(\boldsymbol{\theta}_*) = \left(\left. \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*} \right)$$

are the gradient and the Hessian of g evaluated at $\boldsymbol{\theta}_*$, respectively, for $j, k = 1, 2, 3$. We also write

$$\nabla_j g(\boldsymbol{\theta}_*) = \left. \frac{\partial g(\boldsymbol{\theta})}{\partial \theta_j} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*} \quad \text{and} \quad \nabla^2_{jk} g(\boldsymbol{\theta}_*) = \left. \frac{\partial^2 g(\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \right|_{\boldsymbol{\theta}=\boldsymbol{\theta}_*}.$$

Theorem 3. *Suppose the conditions of Theorem 2 hold and that* $\int |t| w(t) dt < \infty$. *Furthermore, assume that* $\psi_1(x, \lambda) = \partial \psi(x, \lambda)/\partial \lambda$ *is continuous in* (x, λ), *and that there exist functions* $h(x)$ *and* $h_1(x)$ *that satisfy* $|\psi(x, \lambda)| \le h(x)$ *and* $|\psi_1(x, \lambda)| \le h_1(x)$ *for all* λ *in* $\boldsymbol{\Theta}$ *and* $E\left[h^2(X)\right] < \infty$ *and* $E\left[h_1^2(X)\right] < \infty$, *respectively. Then, for* $\boldsymbol{\theta}_0$ *an interior point of* $\boldsymbol{\Theta}$, $n^{1/2} \nabla \varphi_n(\boldsymbol{\theta}_0)$ *is asymptotically distributed with* $N(\boldsymbol{\theta}, \boldsymbol{\theta}(\boldsymbol{\theta}_0))$, *where* $\boldsymbol{\theta}(\boldsymbol{\theta}_0)$ *is specified in the proof.*

Proof. We need to obtain an expression $\nabla \eta(z_1, z_2; \boldsymbol{\theta})$ where $\eta(z_1, z_2; \boldsymbol{\theta})$ is defined in (11.4). Note that

$$\begin{aligned} \nabla \eta(z_1, z_2; \boldsymbol{\theta}) &= \int_{-\infty}^{\infty} \left\{ A(z_1, z_2, t, \boldsymbol{\theta}) + B(z_1, z_2, t, \boldsymbol{\theta}) \right\} t w(t)\, dt \\ &\quad + \int_{-\infty}^{\infty} \left\{ A(z_2, z_1, t, \boldsymbol{\theta}) + B(z_2, z_1, t, \boldsymbol{\theta}) \right\} t w(t)\, dt, \end{aligned}$$

where

$$\begin{aligned} A(z_1, z_2, t, \boldsymbol{\theta}) &= \{\phi(t) - \cos(t z_2(\boldsymbol{\theta}))\} \sin(t z_1(\boldsymbol{\theta}))\, \nabla z_1(\boldsymbol{\theta}) \\ B(z_1, z_2, t, \boldsymbol{\theta}) &= \cos(t z_1(\boldsymbol{\theta})) \sin(t z_2(\boldsymbol{\theta})) \nabla z_1(\boldsymbol{\theta}) \end{aligned}$$

and these involve the factor $\nabla z(\boldsymbol{\theta})$. Since $|\psi(x, \lambda)|$ and $|\psi_1(x, \lambda)|$ are bounded and $\boldsymbol{\Theta}$ is compact, each entry of

$$\nabla z(\boldsymbol{\theta}) = (\psi_1(x, \theta_1)/\theta_3, -1/\theta_3, -\left(\psi(x, \theta_1) - \theta_2\right)/\theta_3^2)^T$$

is bounded. We can now verify that $\nabla \eta(z_1, z_2; \boldsymbol{\theta})$ can be obtained by differentiating under the integral sign in (11.4). The result is, for $j = 1, 2, 3$,

$$\nabla_j \eta(z_1, z_2; \boldsymbol{\theta}) \le 4\left\{ \left| \nabla_j z_1(\boldsymbol{\theta}) \right| + \left| \nabla_j z_2(\boldsymbol{\theta}) \right| \right\} \int_{-\infty}^{\infty} |t|\, w(t)\, dt < \infty.$$

Since $\nabla \eta(z_1, z_2; \boldsymbol{\theta})$ is bounded and continuous in $(z_1, z_2; \boldsymbol{\theta}) \in \Omega_M$, $\nabla \eta(z_1, z_2; \boldsymbol{\theta})$ is equicontinuous in $\boldsymbol{\Theta}$. The random quantity $\psi(X, \lambda)$ and $\psi_1(X, \lambda)$ are each assumed to be dominated for all $\boldsymbol{\theta}$ by a function with finite expectation. The same is clearly true all of the entries of $\nabla \eta(Z_j, Z_k; \boldsymbol{\theta})$.

We now turn to the main proof. From (11.5), we see that

$$\nabla\varphi_n(\boldsymbol{\theta}) = \frac{n-1}{n}\nabla U_n(\boldsymbol{\theta}) + \frac{1}{n^2}\sum_{j=1}^{n}\nabla\eta(Z_j, Z_j; \boldsymbol{\theta})$$

$$= \frac{n-1}{n}\binom{n}{2}^{-1}\sum_{j<k}\nabla\eta(Z_j, Z_k; \boldsymbol{\theta}) + \frac{1}{n^2}\sum_{j=1}^{n}\nabla\eta(Z_j, Z_j; \boldsymbol{\theta}) \qquad (11.8)$$

Again, by the uniform strong law of large numbers, the second term in (11.8) can be neglected.

Note that, since the sine function is odd and the cosine function even, $\nabla\eta(z_1, z_2; \boldsymbol{\theta})$ is a symmetric kernel and so $\nabla U_n(\boldsymbol{\theta}) = \binom{n}{2}^{-1}\sum_{j<k}\nabla\eta(Z_j, Z_k; \boldsymbol{\theta})$ is also a U-statistics. Thus, the multivariate central limit theorems for random samples and U-statistics ensure the asymptotic normality of $\nabla\varphi_n(\boldsymbol{\theta}_0)$ with the mean vector $\nabla\varphi(\boldsymbol{\theta}_0) = \boldsymbol{\theta}$ and the covariance matrix $\mathbf{W}_n(\boldsymbol{\theta}_0)$, where the (j,k)-th element of $\mathbf{W}_n(\boldsymbol{\theta}_0)$ is

$$W_n^{(j,k)}(\boldsymbol{\theta}_0) = \frac{(n-1)^2}{n^2}\binom{n}{2}^{-1}\left\{2(n-2)E[\nabla_j\eta(Z_1, Z_2; \boldsymbol{\theta}_0)\nabla_k\eta(Z_1, Z_3; \boldsymbol{\theta}_0)]\right.$$
$$\left. + E[\nabla_j\eta(Z_1, Z_2; \boldsymbol{\theta}_0)\nabla_k\eta(Z_1, Z_2; \boldsymbol{\theta}_0)]\right\}$$

Therefore, $n^{1/2}\nabla\varphi_n(\boldsymbol{\theta}_0)$ is asymptotically normally distributed as $N(0, \Sigma(\boldsymbol{\theta}_0))$, where the (j,k)-th element of $\Sigma(\boldsymbol{\theta}_0)$ is

$$\Sigma^{(j,k)}(\boldsymbol{\theta}_0) = 4E[\nabla_j\eta(Z_1, Z_2; \boldsymbol{\theta}_0)\nabla_k\eta(Z_1, Z_3; \boldsymbol{\theta}_0)] \quad \square$$

Theorem 4. *Suppose the conditions of Theorem 3 hold and* $\int t^2 w(t)dt < \infty$. *Furthermore, assume that* $\psi_2(x,\lambda) = \partial^2\psi(x,\lambda)/\partial\lambda^2$ *is continuous in* (x,λ), *and that there exists a function* $h_2(x)$ *that satisfies* $|\psi_2(x,\lambda)| \le h_2(x)$ *for all* λ *in* $\boldsymbol{\Theta}$ *and* $E\left[h_2(X)^2\right] < \infty$. *Then,* $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ *is asymptotically distributed with* $N(\boldsymbol{\theta}, \mathbf{V}(\boldsymbol{\theta}_0)\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{V}(\boldsymbol{\theta}_0)^T)$, *where* $\mathbf{V}(\boldsymbol{\theta}_0) = \left(\nabla^2\varphi(\boldsymbol{\theta}_0)\right)^{-1}$.

Proof. Expanding $n^{1/2}\nabla\varphi_n(\hat{\boldsymbol{\theta}})$ about $\boldsymbol{\theta}_0$, we obtain that

$$n^{1/2}\nabla\varphi_n(\hat{\boldsymbol{\theta}}) = n^{1/2}\nabla\varphi_n(\boldsymbol{\theta}_0) + \nabla^2\varphi_n(\tilde{\boldsymbol{\theta}})n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0),$$

where $\tilde{\boldsymbol{\theta}} = \alpha_n\hat{\boldsymbol{\theta}} + (1-\alpha_n)\boldsymbol{\theta}_0$ for $\alpha_n \in [0,1]$. Since $n^{1/2}\nabla\varphi(\hat{\boldsymbol{\theta}}) = \boldsymbol{\theta}$ at the minimum when $\hat{\boldsymbol{\theta}}$ lies in the interior of $\boldsymbol{\Theta}$, $n^{1/2}\nabla\varphi_n(\boldsymbol{\theta}_0) + \nabla^2\varphi_n(\tilde{\boldsymbol{\theta}})n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in probability to $\boldsymbol{\theta}$. From (11.8), $\nabla^2\varphi_n$ can be written as

$$\nabla^2\varphi_n(\boldsymbol{\theta}) = \frac{n-1}{n}\binom{n}{2}^{-1}\sum_{j<k}\nabla^2\eta(Z_j, Z_k; \boldsymbol{\theta}) + \frac{1}{n^2}\sum_{j=1}^{n}\nabla^2\eta(Z_j, Z_j; \boldsymbol{\theta}). \qquad (11.9)$$

Since $|\psi(x,\theta_1)|$, $|\psi_1(x,\theta_1)|$, and $|\psi_2(x,\theta_1)|$ are bounded and $\boldsymbol{\Theta}$ is compact, each element of

$$\nabla^2 z(\boldsymbol{\theta}) = \begin{bmatrix} \psi_2(x,\theta_1)/\theta_3 & 0 & -\psi_1(x,\theta_1)/\theta_3^2 \\ 0 & 0 & 1/\theta_3^2 \\ -\psi_1(x,\theta_1)/\theta_3^2 & 1/\theta_3^2 & 2\left(\psi(x,\theta_1)-\theta_2\right)/\theta_3^3 \end{bmatrix}$$

is bounded and, after some manipulation, we can show, for $j,k = 1,2,3$,

$$\begin{aligned} \nabla^2_{jk}\eta(z_1,z_2;\boldsymbol{\theta}) \;\le\; & 3\left\{\left|\nabla_j z_1(\boldsymbol{\theta})\nabla_k z_1(\boldsymbol{\theta})\right| + \left|\nabla_j z_2(\boldsymbol{\theta})\nabla_k z_2(\boldsymbol{\theta})\right|\right\}\int_{-\infty}^{\infty} t^2\, w(t)\, dt \\ & +2\left\{\left|\nabla^2_{jk} z_1(\boldsymbol{\theta})\right| + \left|\nabla^2_{jk} z_2(\boldsymbol{\theta})\right|\right\}\int_{-\infty}^{\infty} |t|\, w(t)\, dt < \infty. \end{aligned}$$

By the uniform strong law of large numbers, the last term in (11.9) can be neglected. Since $\nabla^2\eta(z_1,z_2;\boldsymbol{\theta})$ is a symmetric kernel,

$$\nabla^2 U_n(\boldsymbol{\theta}) = \binom{n}{2}^{-1}\sum_{j<k}\nabla^2\eta(Z_j,Z_k;\boldsymbol{\theta})$$

is a U-statistic. Applying the uniform strong law of large numbers for U-statistic by Yeo and Johnson (2001) to $\nabla^2 U_n$, we conclude that $\nabla^2\varphi_n(\boldsymbol{\theta})$ converges almost surely to $\nabla^2\varphi(\boldsymbol{\theta})$ uniformly in $\boldsymbol{\theta}\in\boldsymbol{\Theta}$. Further, the limit function $\nabla^2\varphi(\boldsymbol{\theta})$ is continuous in $\boldsymbol{\theta}$. Hence, using the uniform convergence of $\nabla^2\varphi_n$ and the continuity of $\nabla^2\varphi$ with almost sure convergence of $\hat{\boldsymbol{\theta}}$ to $\boldsymbol{\theta}_0$, it is easy to show that

$$\nabla^2\varphi_n(\tilde{\boldsymbol{\theta}}) \text{converges almost surely to} \nabla^2\varphi(\boldsymbol{\theta}_0). \tag{11.10}$$

By Slutsky's theorem along with asymptotic normality of $n^{1/2}\varphi_n(\boldsymbol{\theta}_0)$ and (11.10), we conclude that

$$n^{1/2}(\hat{\boldsymbol{\theta}}-\boldsymbol{\theta}_0) \text{ is asymptotically distributed with } N\left(\boldsymbol{\theta}, \mathbf{V}(\boldsymbol{\theta}_0)\boldsymbol{\Sigma}(\boldsymbol{\theta}_0)\mathbf{V}(\boldsymbol{\theta}_0)^T\right),$$

where $\mathbf{V}(\boldsymbol{\theta}_0) = (\nabla^2\varphi(\boldsymbol{\theta}_0))^{-1}$. □

Remark 1. Note that, for $a_1 \le \lambda \le b_1$, the Box–Cox transformation and the Yeo–Johnson transformation satisfy the following inequalities;

$$\begin{aligned} \left|\psi(x,\lambda)\right| &\le \left|\psi(x,a_1)\right| + \left|\psi(x,b_1)\right| = h(x) \\ \psi_1(x,\lambda) &\le \psi_1(x,a_1) + \psi_1(x,b_1) = h_1(x) \\ \left|\psi_2(x,\lambda)\right| &\le \left|\psi_2(x,a_1)\right| + \left|\psi_2(x,b_1)\right| = h_2(x). \end{aligned}$$

Here $\psi_1(x,\lambda) \ge 0$ for all (x,λ). This was established in Hernandez and Johnson (1980) and Yeo and Johnson (2000), respectively, where it is also shown that $\psi(x,\lambda)$, $\psi_1(x,\lambda)$, and $\psi_2(x,\lambda)$ are continuous in (x,λ).

11.4 Some Exact Calculations with Weight Functions

We have to decide upon a specific weight function to calculate $\varphi_n(\boldsymbol{\theta})$. Our choice here is to obtain some weight functions that yield a closed form for the distance function. Suppose the weight function $w(t)$ is a symmetric probability density function about zero and its characteristic function is $v(\cdot)$. Then, $v(a) = \int \cos(at)w(t)\,dt$ and the first term in (11.1) is written as $n^{-1}\sum_{j<k} v\left(Z_j(\boldsymbol{\theta}) - Z_k(\boldsymbol{\theta})\right)$. Some examples for weight functions and their characteristic functions are as follows;

$$
\begin{aligned}
w(t) &= \frac{1}{\sqrt{2\pi}\delta} e^{-t^2/2\delta^2}, \quad -\infty < t < \infty, \qquad v(s) = e^{-\delta^2 s^2/2} \\
w(t) &= \frac{|t|^{\alpha-1}}{2\Gamma(\alpha)\delta^{\alpha}} e^{-|t|/\delta}, \quad -\infty < t < \infty, \quad v(s) = \frac{1}{2}\left(\frac{1}{1+\delta^2 s^2}\right)^{\alpha} \\
w(t) &= \frac{1}{2\delta}, \quad -\delta < t < \delta, \qquad\qquad v(s) = \frac{\sin(\delta s)}{\delta s}.
\end{aligned}
\tag{11.11}
$$

Note that weight distributions are indexed by a scale parameter, $\delta > 0$. As mentioned in Epps and Pulley (1983), $w(t)$ should assign high weight in some interval around the origin. This implies that the scale parameter must be a small value. A simulation study shows that the shape of weight distribution may not exert a strong influence on the estimation if the scale parameter is sufficiently small.

When the target distribution is normal, the normal density function $w(t)$ gives the closed form for the second term in (11.1) as follows;

$$
\begin{aligned}
\int \phi(t)w(t)\cos(tz)\,dt &= \frac{1}{\sqrt{1+\delta^2}} \int \frac{\sqrt{1+\delta^2}}{\sqrt{2\pi}\delta} \exp\left(-\frac{1+\delta^2}{2\delta^2}t^2\right)\cos(tz)\,dt \\
&= \frac{1}{\sqrt{1+\delta^2}} \exp\left\{-\frac{\delta^2 z^2}{2(1+\delta^2)}\right\}.
\end{aligned}
$$

Since the integration gives the same family as the weight function, we call this type of weight a conjugate weight. Hence, the second term in (11.1) is written as

$$
\sum_{j=1}^{n} \int w(t)\phi(t)\cos\left(t Z_j(\boldsymbol{\theta})\right) dt = \frac{1}{\sqrt{1+\delta^2}} \sum_{j=1}^{n} \exp\left\{-\frac{\delta^2 Z_j(\boldsymbol{\theta})^2}{2(1+\delta^2)}\right\}.
$$

Consequently, when the normal density function with the standard deviation (SD) δ is employed as the weight function $w(t)$, estimates are obtained by minimizing

$$
\begin{aligned}
\varphi_n^*(\boldsymbol{\theta}) \quad \propto \quad & \frac{1}{n}\sum_{j<k} \exp\left\{-\frac{\delta^2}{2}(Z_j(\boldsymbol{\theta}) - Z_k(\boldsymbol{\theta}))^2\right\} \\
& - \frac{1}{\sqrt{1+\delta^2}} \sum_{j=1}^{n} \exp\left\{-\frac{\delta^2}{2(1+\delta^2)} Z_j(\boldsymbol{\theta})^2\right\}.
\end{aligned}
$$

Note that if the degrees of freedom m are odd, the characteristic function of t_m distribution is

$$\phi(t) = \exp(-\sqrt{m}|t|) \sum_{k=0}^{p-1} c_{k,p-1} \sqrt{m}|t|^k$$

where $p = (m+1)/2$ and the $c_{k,p}$s are some constants given in Johnson et al. (1995, p. 367). If $w(t)$ is the double gamma density function such as (11.11), for some $k \geq 0$ and $a > 0$,

$$\int |t|^k \exp(-a|t|) w(t) \cos(tz) dt = \frac{\delta^k \Gamma(\alpha + k)}{2\Gamma(\alpha)} \left\{ (1 + a\delta) \left(1 + \frac{\delta^2 z^2}{(1 + a\delta)^2} \right) \right\}^{-(\alpha+k)}$$

and we can also have a closed form for $\varphi_n^*(\boldsymbol{\theta})$ when the target distribution is t-distribution with m degree of freedom and m is odd. Suppose the goal of transformation is to achieve the Cauchy distribution. Then $\phi(t) = \exp(-|t|)$ and, for some $\alpha > 0$,

$$\begin{aligned} \int \phi(t) w(t) \cos(tz) dt &= \frac{1}{(1+\delta)^\alpha} \int \frac{|t|^{\alpha-1}}{2\Gamma(\alpha)} \left(\frac{1+\delta}{\delta} \right)^\alpha \exp\left(-\frac{1+\delta}{\delta} |t| \right) \cos(tz) dt \\ &= \frac{1}{2(1+\delta)^\alpha} \left\{ 1 + \frac{\delta^2 z^2}{(1+\delta)^2} \right\}^{-\alpha}. \end{aligned}$$

The estimates are obtained by minimizing

$$\varphi_n^*(\boldsymbol{\theta}) \propto \frac{1}{n} \sum_{j<k} \left\{ 1 + \delta^2 (Z_j(\boldsymbol{\theta}) - Z_k(\boldsymbol{\theta}))^2 \right\}^{-\alpha} - \sum_{j=1}^{n} \left\{ (1+\delta) \left(1 + \frac{\delta^2 Z_j(\boldsymbol{\theta})^2}{(1+\delta)^2} \right) \right\}^{-\alpha}.$$

11.5 Simulation Study

In this section, we present a small simulation to compare the proposed method (MECF) with maximum likelihood estimation (MLE) of λ. A series of 1,000 replications, of samples of size $n = 30$, 50, and 100, were generated for $\lambda_0 = 0.0$ and 0.5 according to $\psi(X, \lambda_0) \sim f$ where ψ is Yeo–Johnson transformation and f is one of following distributions: t_m with degrees of freedom $m = 3$, 5, 7 and the standard normal distribution. The double exponential weight function for t-distribution and the normal weight function for standard normal distribution were employed and $\delta = 0.1$ was applied. The R program 'nlminb' is used to obtain optimizers of the likelihood and $\varphi_n(\boldsymbol{\theta})$.

Since our goal of transformation is to approximate symmetry, we also calculate the Pearson's skewness of transformed data as follows;

$$\sqrt{b_1} = \frac{1}{n-1} \sum_{j=1}^{n} \left(\frac{\psi(x_j, \hat{\lambda}) - \bar{\psi}}{s_\psi} \right)^3$$

Table 11.1 The Monte Carlo Bias, standard deviation (SD) and mean squared error (MSE) of $\hat{\lambda}$ and $\sqrt{b_1}$ for MLE and MECF

			$\lambda_0 = 0.0$				$\lambda_0 = 0.5$			
			$\hat{\lambda}$		$\sqrt{b_1}$		$\hat{\lambda}$		$\sqrt{b_1}$	
Target	n		MLE	MECF	MLE	MECF	MLE	MECF	MLE	MECF
t_3^a	30	Bias	0.069	0.055	0.086	0.101	0.041	0.038	0.023	0.037
		SD	0.228	0.262	0.754	0.328	0.300	0.315	0.674	0.303
		MSE	0.057	0.071	0.576	0.117	0.092	0.100	0.455	0.093
	50	Bias	0.056	0.055	0.064	0.079	0.018	0.021	0.015	0.004
		SD	0.173	0.198	0.983	0.447	0.216	0.234	1.006	0.484
		MSE	0.033	0.042	0.969	0.206	0.047	0.055	1.011	0.234
	100	Bias	0.032	0.034	0.000	−0.052	0.012	0.019	−0.095	−0.096
		SD	0.114	0.125	1.117	0.779	0.152	0.168	1.367	0.777
		MSE	0.014	0.017	1.247	0.608	0.023	0.029	1.876	0.613
t_5	30	Bias	0.068	0.063	0.051	0.089	0.024	0.022	0.037	0.051
		SD	0.234	0.255	0.373	0.263	0.273	0.285	0.326	0.245
		MSE	0.059	0.069	0.142	0.077	0.075	0.082	0.108	0.062
	50	Bias	0.036	0.048	0.028	0.102	0.019	0.028	−0.007	0.042
		SD	0.166	0.177	0.387	0.273	0.212	0.214	0.374	0.239
		MSE	0.029	0.034	0.150	0.085	0.045	0.047	0.139	0.059
	100	Bias	0.015	0.035	0.007	0.103	0.007	0.022	−0.022	0.036
		SD	0.112	0.115	0.436	0.284	0.140	0.144	0.462	0.245
		MSE	0.013	0.014	0.190	0.091	0.020	0.021	0.214	0.061
t_7	30	Bias	0.060	0.058	0.063	0.099	0.024	0.025	0.016	0.032
		SD	0.218	0.240	0.223	0.226	0.270	0.279	0.225	0.224
		MSE	0.051	0.061	0.054	0.061	0.073	0.078	0.051	0.051
	50	Bias	0.041	0.058	0.035	0.111	0.011	0.018	0.016	0.043
		SD	0.181	0.190	0.249	0.230	0.205	0.206	0.254	0.211
		MSE	0.034	0.039	0.063	0.065	0.042	0.043	0.065	0.046
	100	Bias	0.018	0.043	−0.006	0.096	0.006	0.020	−0.020	0.028
		SD	0.118	0.119	0.286	0.229	0.143	0.142	0.259	0.195
		MSE	0.014	0.016	0.082	0.061	0.020	0.020	0.068	0.039
$N(0,1)$	30	Bias	0.030	0.041	0.017	0.032	0.005	0.013	−0.004	0.008
		SD	0.205	0.211	0.299	0.336	0.232	0.220	0.264	0.285
		MSE	0.043	0.046	0.090	0.114	0.054	0.049	0.070	0.081
	50	Bias	0.033	0.031	0.060	0.071	0.019	0.006	0.038	0.024
		SD	0.194	0.183	0.506	0.511	0.193	0.197	0.374	0.397
		MSE	0.039	0.034	0.260	0.266	0.038	0.039	0.141	0.158
	100	Bias	0.039	0.038	0.130	0.125	0.013	0.024	0.028	0.057
		SD	0.176	0.165	0.663	0.617	0.158	0.170	0.362	0.412
		MSE	0.032	0.029	0.456	0.396	0.025	0.029	0.132	0.172

[a] Usual proof of asymptotic normality of $\hat{\lambda}$ does not hold because the necessary moments do not exist

where $\bar{\psi}$ and s_ψ are the sample mean and the sample SD of $\psi(x_j, \hat{\lambda})$s, respectively. For each estimation method, we summarize performance by calculating the means, the SD and the mean squared errors (MSE) of $\hat{\lambda}$ and $\sqrt{b_1}$.

Table 11.1 gives bias, standard deviation and mean squared error of estimates for λ_0 and $\sqrt{\beta_0} = 0$. One unexpected finding is that MLE provides better estimates $\hat{\lambda}$

when the underlying distribution has heavier tails, especially for small n, and MECF perform well for the normal distribution with large n and $\lambda_0 = 0$. For $n = 100$, both methods provide similar performances. However, based on the inspection of the skewness $\sqrt{b_1}$ of transformed data, in all cases where the underlying distribution is t-distribution, MECF is definitely better than MLE. From the Pearson skewness point of view, this suggests that transforming data by our method leads to be more symmetric when the population has heavy tails.

Acknowledgements We first met in 1968 when you spoke at a session I chaired at an IMS Regional Meeting in Madison. From our regular contacts since that time, I have become very impressed with your major contributions in the areas of nonparametric and semi-parametric inference. May you continue to make important contributions for a long time to come. (R.J.)

References

Box GEP, Cox DR (1964) An analysis of transformations. J R Stat Soc B 26:211–252

Burbidge JB, Magee L, Robb AL (1988) Alternative transformations to handle extreme values of the dependent variable. J Amer Stat Assoc 83:123–127

Epps TW, Pulley LB (1983) A test for normality based on the empirical characteristic function. Biometrika 70:723–726

Fan Y (1997) Goodness-of-fit tests for a multivariate distribution by the empirical characteristic function. J Multivar Anal 62:36–63

Hernandez F, Johnson RA (1980) The large-sample behavior of transformations to normality. J Amer Stat Assoc 75:855–861

Jimenez-Gamero MD, Alba-Fernandez V, Munoz-Garcia J, Chalco-Cano Y (2009) Goodness-of-fit tests based on empirical characteristic functions. Comput Stat Data Anal 53:3957–3971

John JA, Draper NR (1980) An alternative family of transformations. Appl Stat 29:190–197

Johnson NL (1949) Systems of frequency curves generated by methods of translation. Biometrika 36:149–176

Johnson NR, Kotz S, Balakrishnan N (1995) Continuous univariate distributions Vol 2, 2nd Edn. Wiley, New York

Klar B, Meintanis SG. (2005) Tests for normal mixtures based on the empirical characteristic function. Comput Stat Data Anal 49:227–242

Koutrouvelis IA. (1980) A goodness-of-fit test of simple hypothesis based on the empirical characteristic function. Biometrika, 67, 238–240

Koutrouvelis IA, Kellermeier J (1981) A goodness-of-fit based on the empirical characteristic function when parameters must be estimated. J R Stat Soc B 43:173–176

Lee AJ (1990) U-statistics : theory and practice. Marcel Dekker, New York

Rubin H (1956) Uniform convergence of random functions with applications to statistics. Ann Math Stat 27:200–203

Taylor JMG (1985) Power transformations to symmetry. Biometrika 72:145–152

Szekely GJ, Rizzo ML, Bakirov NK (2007) Measuring and testing dependence by correlation of distances. Ann Stat 35:2769–2794

van Zwet WR (1964) Convex transformations of random variables. Mathematisch Centrum, Amsterdam

Yeo IK. (2001) Selecting a transformation to reduce skewness. J Korean Stat Soc 30:563–571

Yeo IK, Johnson RA (2000) A new family of power transformations to improve normality or symmetry. Biometrika 87:954–959

Yeo IK, Johnson RA. (2001) A uniform strong law of large numbers for U-statistics with application to transforming to near symmetry. Stat Probab Lett 51:63–69

Yeo IK, Johnson RA, Deng X (2013) An empirical characteristic function approach for robust transformation to normality. Manuscript

Chapter 12
Averaged Regression Quantiles

Jana Jurečková and Jan Picek

12.1 Introduction

Consider the linear regression model

$$\mathbf{Y}_n = \mathbf{X}_n\beta + \mathbf{U}_n \tag{12.1}$$

with observations $\mathbf{Y}_n = (Y_1, \ldots, Y_n)^\top$, i.i.d. errors $\mathbf{U}_n = (U_1, \ldots, U_n)^\top$ with an unknown distribution function F, and unknown parameter $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^\top$. The $n \times (p+1)$ matrix $\mathbf{X} = \mathbf{X}_n$ is known and $x_{i0} = 1$ for $i = 1, \ldots, n$ (i.e., β_0 is an intercept). The α-regression quantile $\widehat{\beta}_n(\alpha)$ of model (12.1) is a solution of the minimization

$$\sum_{i=1}^{n} \rho_\alpha(Y_i - \mathbf{x}_i^\top \mathbf{b}) := \min \tag{12.2}$$

with respect to $\mathbf{b} = (b_0, \ldots, b_p)^\top \in \mathbb{R}^{p+1}$, where $\mathbf{x}_i^\top$ is the i-th row of $\mathbf{X}_n$, $i = 1, \ldots, n$ and $\rho_\alpha(z) = |z|\{\alpha I[z > 0] + (1-\alpha)I[z < 0]\}$, $z \in \mathbb{R}^1$. The population counterpart of $\widehat{\beta}_n(\alpha)$ is the vector $\beta(\alpha) = (\beta_0 + F^{-1}(\alpha), \beta_1, \ldots, \beta_p)^\top$. For the brevity, we shall occasionally use the notation

$$\mathbf{x}_i^* = (x_{i1}, \ldots, x_{ip})^\top, \quad i = 1, \ldots, n \quad \text{and } \mathbf{X}_n^* = \left[\mathbf{x}_1^*, \ldots, \mathbf{x}_n^*\right]^\top$$

and $\widehat{\beta}_n^*(\alpha) = (\hat{\beta}_1(\alpha), \ldots, \hat{\beta}_p(\alpha))^\top$. Assume that the distribution function $F(x)$ of the errors U_i is increasing on the set $\{x : 0 < F(x) < 1\}$. For any fixed $\alpha \in (0,1)$, denote $U_{i\alpha} = U_i - F^{-1}(\alpha)$, $i = 1, \ldots, n$. Then $U_{1\alpha}, \ldots, U_{n\alpha}$ are i.i.d. random

J. Jurečková (✉)
Faculty of Mathematics and Physics, Department of Statistics,
Charles University in Prague, Sokolovská 83, 186 75 Prague, Czech Republic
e-mail: jurecko@karlin.mff.cuni.cz,

J. Picek
Department of Applied Mathematics, Technical University in Liberec,
Voroněžská 1329/13, 460 01 Liberec, Czech Republic
e-mail: jan.picek@tul.cz

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_12,

variables with distribution function $F_\alpha(x) = F(x + F^{-1}(\alpha))$, $x \in \mathbb{R}$, and $F_\alpha^{-1}(u) = F^{-1}(u) - F^{-1}(\alpha)$, $0 < u < 1$, so that $F_\alpha^{-1}(\alpha) = 0$. It is sometimes convenient to rewrite the model (12.1) in the following way:

$$Y_{ni} = \beta_0(\alpha) + \mathbf{x}_{ni}^{*\top}\beta^* + U_{i\alpha},\ i = 1, \ldots, n \tag{12.3}$$

with $\beta_0(\alpha) = \beta_0 + F^{-1}(\alpha)$. We shall omit the subscript n whenever it does not cause a confusion. The α-regression quantile for the reparametrized model (12.3) is then a solution of the minimization

$$\sum_{i=1}^{n} \left\{\alpha[Y_i - b_0(\alpha) - \mathbf{x}_i^{*\top}\mathbf{b}]^+ + (1-\alpha)[Y_i - b_0(\alpha) - \mathbf{x}_i^{*\top}\mathbf{b}]^-\right\} = \min$$
$$\text{with respect to } b_0(\alpha) \in \mathbb{R}^1, \mathbf{b} \in \mathbb{R}^p, \tag{12.4}$$

where $z^+ = \max\{z, 0\}$, $z^- = \max\{-z, 0\}$ for $z \in \mathbb{R}^1$.

The α-regression quantile was introduced by Koenker and Bassett (1978), who used a linear programming algorithm for its calculation. They also used the following dual algorithm as a computational device:

$$\begin{aligned} & \sum_{i=1}^{n} Y_i\ \hat{a}_i := \max \\ \text{under the constraint} \quad & \sum_{i=1}^{n} \hat{a}_i = n(1-\alpha), \\ & \sum_{i=1}^{n} x_{ij}\hat{a}_i = (1-\alpha)\sum_{i=1}^{n} x_{ij},\ j = 1, \ldots, p, \\ & 0 \le \hat{a}_i \le 1, \quad i = 1, \ldots, n, \quad 0 < \alpha < 1. \end{aligned} \tag{12.5}$$

The components of the optimal solution of (12.5),

$$\widehat{\mathbf{a}}_n(\alpha) = (\hat{a}_{n1}(\alpha), \ldots, \hat{a}_{nn}(\alpha))^\top, \quad 0 \le \alpha \le 1$$

were named the *regression rank scores* by Gutenbrunner and Jurečková (1992), who used them for construction of the rank tests in the linear model. The matrix form of program (12.5) is more compact:

$$\begin{aligned} & \mathbf{Y}_n^\top \widehat{\mathbf{a}} := \max \\ \text{under the constraint} \quad & (\mathbf{X}_n)^\top \widehat{\mathbf{a}} = (1-\alpha)(\mathbf{X}_n)^\top \mathbf{1}_n, \\ & \widehat{\mathbf{a}} \in [0,1]^n,\ 0 \le \alpha \le 1. \end{aligned} \tag{12.6}$$

This implies that the regression rank scores are *invariant with respect to the shift in location and scale and to the changes of* β, i.e.,

$$\widehat{\mathbf{a}}_n(\alpha, \mathbf{Y} + \mathbf{X}_n\mathbf{b}) = \widehat{\mathbf{a}}_n(\alpha, \mathbf{Y}) \quad \forall \mathbf{b} \in \mathbb{R}^{p+1}. \tag{12.7}$$

As $\widehat{\boldsymbol{\beta}}_n(\alpha)$ and $\widehat{\mathbf{a}}_n(\alpha)$ are dual to each other, we get from the linear programming theory that

$$\hat{a}_{ni}(\alpha) = \begin{cases} 1 & \dots \quad Y_i > \mathbf{x}_i^\top \widehat{\beta}_n(\alpha), \\ 0 & \dots \quad Y_i < \mathbf{x}_i^\top \widehat{\beta}_n(\alpha), \quad i = 1, \dots, n \end{cases} \tag{12.8}$$

and if $Y_i = \mathbf{x}_i^\top \widehat{\beta}_n(\alpha)$ for some i (the exact fit), then $0 < \hat{a}_{ni}(\alpha) < 1$; there are exactly $p+1$ such components for each α, corresponding to the optimal base among $\mathbf{x}_1, \dots, \mathbf{x}_n$. The pertinent values of $\hat{a}_{ni}(\alpha)$ are determined by the constraints in (12.6).

Assume the following regularity conditions on the matrix $\mathbf{X}_n$:

A1 $\lim_{n\to\infty} \mathbf{Q}_n = \mathbf{Q}$, where $\mathbf{Q}_n = n^{-1}\mathbf{X}_n^\top \mathbf{X}_n$ and $\mathbf{Q}$ is a positive definite matrix.

A2 $n^{-1}\sum_{i=1}^n x_{ij}^4 = \mathcal{O}(1)$, as $n \to \infty$, for $j = 1, \dots, p$.

Then the α-regression quantile admits the following Bahadur-type representation (for the proof see e.g., Jurečková et al. 2012):

Theorem 1. *Suppose that the distribution function F is continuous and twice differentiable in a neighborhood of $F^{-1}(\alpha)$ and that $F'(F^{-1}(\alpha)) = f(F^{-1}(\alpha)) > 0$, $0 < \alpha < 1$. Then, under the conditions* **A1**–**A2**,

$$\widehat{\beta}_n(\alpha) - \widetilde{\beta}(\alpha) = \frac{1}{nf(F^{-1}(\alpha))}\mathbf{Q}_n^{-1}\sum_{i=1}^n \mathbf{x}_i\,\psi_\alpha(U_i - F^{-1}(\alpha)) + \mathbf{R}_n(\alpha), \tag{12.9}$$

where $\|\mathbf{R}_n(\alpha)\| = \mathcal{O}_p(n^{-3/4})$ as $n \to \infty$ and

$$\widetilde{\beta}(\alpha) = (\beta_0 + F^{-1}(\alpha), \beta_1, \dots, \beta_p)^\top, \quad \psi_\alpha(z) = \alpha - I[z < 0], \quad z \in \mathbb{R}^1.$$

The convergence is uniform on interval $[\varepsilon, 1-\varepsilon]$ for every fixed $\varepsilon \in (0, 1/2)$. The process on the right-hand side of (12.9) is the weighted empirical process. Such processes and their asymptotic properties were systematically studied by H. L. Koul; we refer to his excellent monograph Koul (2002) with a rich bibliography.

The regression quantiles were intensively applied in the statistical and econometric inference; here we refer to Koenker's (2005) monograph and to the references cited in, among others. Their extension to the autoregression processes was studied by Koul and Saleh (1995).

Parallelly, the *two-step α-regression quantile* was proposed by the authors in Jurečková and Picek (2005): It first estimates the slope components $\boldsymbol{\beta}^*$ by means of an R-estimate $\widetilde{\boldsymbol{\beta}}_R^*(\alpha) \in \mathbb{R}^p$ as a minimizer of the Jaeckel's measure of rank dispersion (Jaeckel 1972)

$$\sum_{i=1}^n (Y_i - \mathbf{x}_i^{*\top}\mathbf{b}^*)[a_i(\alpha, \mathbf{b}^*) - (1-\alpha)] = \min \text{ with respect to } \mathbf{b}^* \in \mathbb{R}^p \tag{12.10}$$

where

$$a_i(\alpha, \mathbf{b}^*) = \begin{cases} 0 & \dots \quad R_{ni}(Y_i - \mathbf{x}_i^{*\top}\mathbf{b}^*) < n\alpha \\ R_i - n\alpha & \dots \quad n\alpha \le R_{ni}(Y_i - \mathbf{x}_i^{*\top}\mathbf{b}^*) < n\alpha + 1 \\ 1 & \dots \quad n\alpha + 1 \le R_{ni}(Y_i - \mathbf{x}_i^{*\top}\mathbf{b}^*), \end{cases} \tag{12.11}$$

$R_{ni}(Y_i - \mathbf{x}_i^{*\top}\mathbf{b}^*)$ are the ranks of the residuals and $a_i(\alpha, \mathbf{b}^*)$ are known as *Hájek's rank scores* (Hájek 1965), $i = 1, \ldots, n$. The second step of the procedure determines the $[n\alpha]$-quantile $\tilde{\beta}_{0R}(\alpha)$ of the residuals $\{Y_i - \mathbf{x}_i^{*\top}\widetilde{\boldsymbol{\beta}}_R^*(\alpha)\}$, $i = 1, \ldots, n$. Then the two-step regression quantile is $\widetilde{\beta}_R(\alpha) = (\tilde{\beta}_{0R}(\alpha), \tilde{\boldsymbol{\beta}}_R^{*\top}(\alpha))^\top$. It is asymptotically equivalent to the standard regression quantile $\widehat{\beta}_n(\alpha)$, i.e.,

$$\left\|\widehat{\beta}_n(\alpha) - \widetilde{\beta}_{nR}(\alpha)\right\| = o_p(n^{-1/2}) \tag{12.12}$$

as $n \to \infty$. The common population counterpart of $\widehat{\beta}_n(\alpha)$ and $\widetilde{\beta}_n(\alpha)$ is $(F^{-1}(\alpha) + \beta_0, \beta_1, \ldots, \beta_p)^\top$. The finite-sample relations of both versions of regression quantiles are studied in Jurečková and Picek (2005); for special α's their values exactly coincide.

If the inference concerns mainly the functionals of $F^{-1}(\alpha)$ rather than the regressors, we try to reduce the influence of the matrix $\mathbf{X}_n$. It turns out that a suitable projection of $\widehat{\beta}_n(\alpha)$ (a special weighted empirical process) depends asymptotically only on the quantile of the model errors $U_1, \ldots, U_n$. This considerably simplifies the inference, and we shall deal with this phenomenon further.

12.2 Averaged Regression Quantiles

We shall call the scalar statistic

$$\bar{B}_n(\alpha) = \bar{\mathbf{x}}_n^\top \widehat{\beta}_n(\alpha), \qquad \bar{\mathbf{x}}_n = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_{ni} \tag{12.13}$$

the *averaged regression quantile*, and will study its properties and relations to other statistics. Notice that $\bar{B}_n(\alpha)$ is scale equivariant and it is regression equivariant in the sense that

$$\bar{B}_n(\alpha; \mathbf{Y} + \mathbf{X}\mathbf{b}) = \bar{B}_n(\alpha, \mathbf{Y}) + \bar{\mathbf{x}}^\top \mathbf{b} \quad \forall \mathbf{b} \in \mathbb{R}_{p+1}.$$

Some properties of $\bar{B}_n(\alpha)$ are surprising; indeed, $\bar{B}_n(\alpha)$ is asymptotically equivalent to the $[n\alpha]$-quantile of the location model. The following useful identity for $\bar{B}_n(\alpha)$ was first proven in Hallin and Jurečková (1999) for the linear autoregression model:

Lemma 1 *(i) If $\alpha \in (0, 1)$ is a continuity point of $\widehat{\beta}_n(\alpha)$, then*

$$\bar{B}_n(\alpha) = -\frac{1}{n}\sum_{i=1}^n Y_i \, \frac{d}{d\alpha}\hat{a}_i(\alpha).$$

(ii) $\bar{B}_n(\alpha)$ and hence also $-\frac{1}{n}\sum_{i=1}^n Y_i \, \frac{d}{d\alpha}\hat{a}_i(\alpha)$ are nondecreasing step-functions of $\alpha \in (0, 1)$.

Proof. The duality between $\widehat{\beta}_n(\alpha)$ and $\widehat{\mathbf{a}}_n(\alpha)$ implies that

$$\sum_{i=1}^{n} \rho_\alpha(Y_i - \mathbf{x}_i^\top \widehat{\beta}_n(\alpha)) = \sum_{i=1}^{n} Y_i(\hat{a}_i(\alpha) - (1-\alpha)).$$

Hence, for $0 < \alpha_1 < \alpha_2 < 1$,

$$\begin{aligned} &\sum_{i=1}^{n} \left[\rho_{\alpha_2}(Y_i - \mathbf{x}_i^\top \widehat{\beta}(\alpha_1)) - \rho_{\alpha_1}(Y_i - \mathbf{x}_i^\top \widehat{\beta}(\alpha_1))\right] \\ = \quad &(\alpha_2 - \alpha_1) \sum_{i=1}^{n} (Y_i - \mathbf{x}_i^\top \widehat{\beta}_1(\alpha_1)) \geq \sum_{i=1}^{n} Y_i \left[\hat{a}_i(\alpha_2) - \hat{a}_i(\alpha_1) + (\alpha_2 - \alpha_1)\right], \end{aligned}$$

thus

$$(\alpha_2 - \alpha_1) \sum_{i=1}^{n} \mathbf{x}_i^\top \widehat{\beta}(\alpha_1)) \leq -\sum_{i=1}^{n} Y_i(\hat{a}_i(\alpha_2) - \hat{a}_i(\alpha_1)). \tag{12.14}$$

Analogously, we obtain

$$(\alpha_2 - \alpha_1) \sum_{i=1}^{n} \mathbf{x}_i^\top \widehat{\beta}_1(\alpha_2)) \geq -\sum_{i=1}^{n} Y_i(\hat{a}_i(\alpha_2) - \hat{a}_i(\alpha_1)). \tag{12.15}$$

(12.14) and (12.15) imply

$$\bar{\mathbf{x}}_n^\top \widehat{\beta}_n(\alpha_1) \leq -\frac{1}{n} \sum_{i=1}^{n} Y_i \, \frac{\hat{a}_i(\alpha_2) - \hat{a}_i(\alpha_1)}{\alpha_2 - \alpha_1} \leq \bar{\mathbf{x}}_n^\top \widehat{\beta}_n(\alpha_2).$$

This entails the monotonicity of $\bar{\mathbf{x}}_n^\top \widehat{\beta}_n(\alpha)$. On the other hand, $\widehat{\beta}_n(\alpha)$ is a step-function, and $\widehat{\mathbf{a}}_n(\alpha)$ is a piecewice linear function of α, and the points of discontinuity of $\widehat{\beta}_n(\alpha)$ and of $\frac{d}{d\alpha}\widehat{\mathbf{a}}_n(\alpha)$ coincide. Hence, letting $\alpha_2 \to \alpha_1$, we obtain the Lemma. □

The following theorem shows that the averaged regression α-quantile is asymptotically equivalent to the location α-quantile:

Theorem 2. *Under the conditions of Theorem 1,*

$$n^{1/2} \left[\bar{\mathbf{x}}_n^\top (\widehat{\beta}_n(\alpha) - \beta) - U_{n:[n\alpha]}\right] = \mathcal{O}_p(n^{-1/4}) \tag{12.16}$$

as $n \to \infty$, *where* $U_{n:1} \leq \ldots \leq U_{n:n}$ *are the order statistics corresponding to* $U_1, \ldots, U_n$.

Proof. By **A1**, **A2** and Theorem 1,

$$\begin{aligned} &\sqrt{n}\bar{\mathbf{x}}_n^\top (\widehat{\beta}_n(\alpha) - \widetilde{\beta}(\alpha)) \qquad (12.17) \\ &= \frac{1}{f(F^{-1}(\alpha))} \sqrt{n}\bar{\mathbf{x}}_n^\top (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \sum_{i=1}^{n} \mathbf{x}_i(\alpha - I[U_i < F^{-1}(\alpha)]) + \mathcal{O}_p(n^{-1/4}) \end{aligned}$$

$$
\begin{aligned}
&= [\sqrt{n}f(F^{-1}(\alpha))]^{-1} \sum_{k,i=1}^{n} \mathbf{x}_k^\top (\mathbf{X}_n^\top \mathbf{X}_n)^{-1} \mathbf{x}_i (\alpha - I[U_i < F^{-1}(\alpha)]) + \mathcal{O}_p(n^{-1/4}) \\
&= [\sqrt{n}f(F^{-1}(\alpha))]^{-1} \mathbf{1}_n^\top \widehat{\mathbf{H}}_n \mathbf{c}_n(\alpha) + \mathcal{O}_p(n^{-1/4}) \\
&= [\sqrt{n}f(F^{-1}(\alpha))]^{-1} \mathbf{1}_n^\top \mathbf{c}_n(\alpha) + \mathcal{O}_p(n^{-1/4}) \\
&= [\sqrt{n}f(F^{-1}(\alpha))]^{-1} \sum_{i=1}^{n} \big(\alpha - I[U_i < F^{-1}(\alpha)]\big) + \mathcal{O}_p(n^{-1/4}) \\
&= \sqrt{n}\left(U_{n:[n\alpha]} - F^{-1}(\alpha)\right) + \mathcal{O}_p(n^{-1/4})
\end{aligned}
$$

where $\mathbf{c}_n(\alpha) = (\alpha - I[U_1 < F^{-1}(\alpha)], \ldots, \alpha - I[U_n < F^{-1}(\alpha)])^\top$ and $\widehat{\mathbf{H}}_n = \mathbf{X}_n(\mathbf{X}_n^\top \mathbf{X}_n)^{-1}\mathbf{X}_n^\top$ is the projection matrix.

Remark 1 *It follows from (12.12) that the approximation (12.16) is true also for the two-step regression quantile* $\tilde{\beta}_R(\alpha)$. *Moreover,*

$$
n^{1/2}[\tilde{\beta}_{0R}(\alpha) - \beta_0 - U_{n:[n\alpha]}] = o_p(1) \quad \text{as } n \to \infty. \tag{12.18}
$$

Theorem 2 has an easy corollary:

Corollary 1 *Under the conditions of Theorem 1,*

$$
n^{1/2}\left[\bar{\mathbf{x}}_n^\top(\widehat{\beta}_n(\alpha_2) - \widehat{\beta}_n(\alpha_1)) - (U_{n:[n\alpha_2]} - U_{n:[n\alpha_1]})\right] = \mathcal{O}_p(n^{-1/4}) \tag{12.19}
$$

for any $0 < \alpha_1 \le \alpha_2 < 1$.

The statistics of type $\bar{\mathbf{x}}_n^\top(\widehat{\beta}_n(\alpha_2) - \widehat{\beta}_n(\alpha_1))$ are invariant to the regression with design $\mathbf{X}$ and equivariant with respect to the scale. As such, they provide a tool for studentization of M-estimators in linear regression model and always when one needs to make a statistic scale-equivariant. The properties of studentized M-estimators are thoroughly studied in Jurečková et al. (2012). In Jurečková et al. (2003), the authors use the regression interquartile range with $\alpha_1 = \frac{1}{4}$, $\alpha_2 = \frac{3}{4}$ in goodness-of-fit testing with nuisance regression and scale.

12.3 Local Heteroscedasticity

The approximation (12.16) remains true under a sequence of local alternative distributions, contiguous with respect to the sequence $\left\{\prod_{i=1}^n F(u_{ni})\right\}$. Among them, the local heteroscedasticity deserves a special study. The frequent heteroscedastic model has the form

$$
Y_i = \beta_0 + \mathbf{x}_i^\top \beta + \sigma_i U_i, \quad i = 1, \ldots, n \tag{12.20}
$$

where $\mathbf{U}_n = (U_1, \ldots, U_n)^\top$ are the i.i.d. errors with the joint distribution function F and

$$\sigma_i = \exp\{\mathbf{d}_i^\top \gamma\}, \quad i = 1, \ldots, n \tag{12.21}$$

with known or observable $\mathbf{d}_i \in \mathbb{R}^q$, $1 \le i \le n$ and unknown parameter $\gamma \in \mathbb{R}^q$. We assume that

$$\begin{aligned} &\sum_{i=1}^n d_{ij} = 0, \quad j = 1, \ldots, q, \\ &\max_{1 \le i \le n} \|\mathbf{d}_i\| = o(n^{\frac{1}{2}}) \text{ as } \quad n \to \infty, \\ &\lim_{n\to\infty} \mathbf{D}_n = \lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n \mathbf{d}_i \mathbf{d}_i^\top = \mathbf{D}, \\ &\max_{1 \le i \le n} \left\{ \mathbf{d}_i^\top \left(\sum_{k=1}^n \mathbf{d}_k \mathbf{d}_k^\top \right)^{-1} \mathbf{d}_i \right\} \to 0 \text{ as } n \to \infty \end{aligned} \tag{12.22}$$

where $\mathbf{D}$ is positive definite $(q \times q)$ matrix. The homoscedasticity means that $\gamma = \mathbf{0}$; then (12.16) applies. The *local heteroscedasticity* means that

$$\gamma = \gamma_n = n^{-\frac{1}{2}} \delta, \quad \delta \in \mathbb{R}^q, \ \delta \neq \mathbf{0}, \ \|\delta\| \le C < \infty. \tag{12.23}$$

The following theorem shows that (12.16) remains true under the local heteroscedasticity:

Theorem 3. *Consider the model* (12.20) *under the local heteroscedasticity satisfying* (12.21), (12.22) *and* (12.23). *Then* (12.16) *remains true for any fixed* $\alpha \in (0, 1)$. *Moreover*,

$$\begin{aligned} &\sqrt{n} \bar{\mathbf{x}}_n^\top (\widehat{\beta}_n(\alpha) - \beta - \mathbf{e}_0 F^{-1}(\alpha)) \\ &\qquad = \frac{1}{\sqrt{n} f(F^{-1}(\alpha))} \sum_{i=1}^n \left(\alpha - I[U_i < F^{-1}(\alpha)] \right) + \mathcal{O}_p(n^{-1/4}), \\ &\sqrt{n}(U_{n:[n\alpha]} - F^{-1}(\alpha)) \\ &\qquad = \frac{1}{\sqrt{n} f(F^{-1}(\alpha))} \sum_{i=1}^n \left(\alpha - I[U_i < F^{-1}(\alpha)] \right) + \mathcal{O}_p(n^{-1/4}) \end{aligned} \tag{12.24}$$

and both $\{\sqrt{n} \bar{\mathbf{x}}_n^\top (\widehat{\beta}_n(\alpha) - \beta - \mathbf{e}_0 F^{-1}(\alpha))\}$ and $\{\sqrt{n}(U_{n:[n\alpha]} - F^{-1}(\alpha))\}$ *are asymptotically normally distributed* $\mathcal{N}\big(0, \frac{\alpha(1-\alpha)}{f^2(F^{-1}(\alpha))}\big)$ *also under local heteroscedasticity*; $\mathbf{e}_0 = (1, 0, \ldots, 0)^\top \in \mathbb{R}^{p+1}$.

Proof. Under (12.20), (12.21) and (12.23), the random vector $\mathbf{Y}$ has the density

$$q_{n\gamma}(y_1, \ldots, y_n) = \prod_{i=1}^n \exp\{\mathbf{d}_i^\top \gamma\} f\left(y_i \exp\{\mathbf{d}_i^\top \gamma\} \right). \tag{12.25}$$

Under local heteroscedasticity (12.23), the sequence of densities $\{q_{n\gamma}\}$ is contiguous to $\{q_{n0}\}$ corresponding to $\gamma = \mathbf{0}$ (see Hájek 1965, Chap. VI). Hence, then propositions (12.16) and (12.17) remain true. The asymptotic distributions follow from (12.24), using expansions of the moments. □

12.4 Quantile Density Function

The quantile density function $q(u) = \frac{1}{f(F^{-1}(u))}$ is used in nonparametric statistical inference, as in the studentization, adaptive procedures, in the sequential confidence sets, in tests on β based on L_1-regression and elsewhere. It is a scale statistic, being location invariant and scale equivariant. The sum of quantile densities is again a quantile density of some random variable. A typical term in the asymptotic variance of empirical α-quantile is $q^2(\alpha)$. Siddiqui (1960), Bloch and Gastwirth (1968), Bofinger (1975), Lai et al. (1983), and others considered the histogram estimate of $q(\alpha)$ in the location model. Parzen (1979), Yang (1985), Falk (1986), Zelterman (1990), Soni et al. (2012), among others, considered kernel-type estimators of $q(\alpha)$. Xiang (1995) studied the kernel estimator of the conditional quantile density function.

Based on observations $Y_{n1}, \ldots, Y_{nn}$ in model (12.1), we want to estimate $q(\alpha)$ at the point α. Such estimator should be regression invariant and scale equivariant. The first estimates of $q(\alpha)$ in the linear regression model were proposed by Koenker and Bassett (1978) and Welsh (1987). Welsh (1987) constructed a class of estimators of $q(\alpha)$ based on a kernel smoothing the empirical quantile function of the residuals from an estimator of β. Dodge and Jurečková (1995) extended Falk's (1986) estimator to the linear model, using the first component of $\widehat{\beta}(\alpha)$ under the assumption that $\bar{x}_j = \sum_{i=1}^n x_{ij} = 0$ for $j = 1, \ldots, p$.

Applying Theorem 1, we can construct analogues of estimators of Dodge and Jurečková (1995) based on $\bar{B}(\alpha)$. These estimators, not demanding $\bar{x}_j = 0$ for $j = 1, \ldots, p$, can be used also in autoregression and sequential models, where this condition does not hold.

Let us first consider the histogram type estimate

$$H_n(\alpha) = \frac{1}{2\nu_n}[\bar{B}_n(\alpha + \nu_n) - \bar{B}_n(\alpha - \nu_n)] \tag{12.26}$$

where

$$\nu_n = o(n^{-1/3}), \quad n\nu_n \to \infty \text{ as } n \to \infty.$$

Then $H_n(\alpha)$ is consistent and asymptotically normal.

Theorem 4. *Under* (12.26) *and under the conditions of Theorem 2,*

$$H_n(\alpha) - q(\alpha) = \mathcal{O}_p(n\nu_n)^{-1/2} \quad \text{as } n \to \infty, \tag{12.27}$$

uniformly in $\alpha \in (\varepsilon, 1-\varepsilon)$, $\forall\, \varepsilon \in (0, 1/2)$.

Moreover, $H_n(\alpha)$ is asymptotically normal for every fixed $\alpha \in (\varepsilon, 1-\varepsilon)$,

$$(n\nu_n)^{1/2}(H_n(\alpha) - q(\alpha)) \xrightarrow{\mathcal{D}} \mathcal{N}\left(0, \frac{1}{2}q^2(\alpha)\right) \quad \text{as } n \to \infty. \tag{12.28}$$

Proof. Let F_n denote the empirical distribution function of $U_{n1}, \ldots, U_{nn}$. By (12.17) and (12.26),

$$\begin{aligned} H_n(\alpha) &= (2\nu_n)^{-1}\left(F^{-1}(\alpha+\nu_n) - F^{-1}(\alpha-\nu_n)\right) \\ &+(2\nu_n)^{-1}\left\{q(\alpha+\nu_n)[\alpha+\nu_n - F_n(F^{-1}(\alpha+\nu_n))]\right. \\ &\left.-q(\alpha-\nu_n)[\alpha-\nu_n - F_n(F^{-1}(\alpha-\nu_n))]\right\} + \mathcal{O}_p(n^{-3/4}\nu_n^{-1}) \\ &= q(\alpha) + \mathcal{O}_p((n\nu_n)^{-1/2}), \end{aligned}$$

what demonstrates (12.27).

To prove (12.28), notice that by Csörgö and Révész (1978), there exists a sequence of Brownian Bridges $\mathcal{B}_n(\cdot)$, dependent on $U_{n1}, \ldots, U_{nn}$, respectively, such that

$$(2n\nu_n)^{1/2}(H_n(\alpha) - q(\alpha)) = (2\nu_n)^{-1/2}q(\alpha)[\mathcal{B}_n(\alpha+\nu_n) - \mathcal{B}_n(\alpha-\nu_n)] + o_p(1)$$

as $n \to \infty$. This implies (12.28).

Following Falk (1986) and Dodge and Jurečková (1995), define the kernel estimate of $q(\alpha)$ as follows:

$$\widehat{\kappa}_n(\alpha) = \frac{1}{\nu_n^2}\int_0^1 \bar{B}_n(u)k\Big(\frac{\alpha-u}{\nu_n}\Big)du, \tag{12.29}$$

assuming that

$$\nu_n \downarrow 0, \quad n\nu_n^3 \downarrow 0 \text{ and } n\nu_n^2 \to \infty \text{ as } n \to \infty. \tag{12.30}$$

The kernel function $k : \mathbb{R}^1 \mapsto \mathbb{R}^1$ is assumed to satisfy the following condition:

K1: $k(\cdot)$ is continuous on its compact support and

$$\int k(x)dx = 0, \qquad \int xk(x)dx = -1.$$

The estimator $\widehat{\kappa}_n(\alpha)$ is consistent and asymptotically normal:

Theorem 5. *In the model (12.1), let distribution function F of U_1 have continuous density f which is positive and finite in $\{x : 0 < F(x) < 1\}$. Let F^{-1} be twice differentiable with bounded second derivative in a neighborhood of α. Then, under the conditions of Theorem 2,*

$$\widehat{\kappa}_n(\alpha) - q(\alpha) = \mathcal{O}_p((n\nu_n))^{-1/2} \quad \text{as } n \to \infty. \tag{12.31}$$

Moreover, $\widehat{\kappa}_n(\alpha)$ is asymptotically normally distributed,

$$(n\nu_n)^{1/2}(\widehat{\kappa}_n(\alpha) - q(\alpha)) \xrightarrow{\mathcal{D}} \mathcal{N}\big(0, q^2(\alpha)\int K^2(x)dx\big), \tag{12.32}$$

where $K(x) = \int_{-\infty}^{x} k(y)dy$.

Proof. First notice that

$$\widehat{\kappa}_n(\alpha) = \frac{1}{\nu_n^2}\int_0^1 [\bar{B}_n(u) - \bar{\mathbf{x}}_n^\top \beta]k\big(\frac{\alpha - u}{\nu_n}\big)du.$$

Starting with $n \geq n_0$, the interval $\big(\frac{\alpha-1}{\nu_n}, \frac{\alpha}{\nu_n}\big)$ contains the support of $k(\cdot)$. Hence, by (12.16) and (12.17),

$$\begin{aligned}
&\widehat{\kappa}_n(\alpha) = \nu_n^{-2}\int_0^1 F^{-1}(u)k\big(\frac{\alpha - u}{\nu_n}\big)du \\
&+\nu_n^{-2}\int_0^1 q(u)[u - F_n(F^{-1}(u))]k\big(\frac{\alpha - u}{\nu_n}\big)du + \mathcal{O}_p(n^{-3/4}\nu_n^{-1}) \\
&= q(\alpha) + n^{-1}\nu_n^{-2}\sum_{i=1}^{n}\int_0^1 q(u)\{u - I[F(U_i) \leq u]\}k\big(\frac{\alpha - u}{\nu_n}\big)du \\
&+\mathcal{O}_p(n^{-3/4}\nu_n^{-1}) \qquad (12.33) \\
&= q(\alpha) + (n\nu_n)^{-1}\sum_{i=1}^{n}\int q(\alpha - \nu_n z)\{\alpha - \nu_n z - I[F(U_i) \leq (\alpha - \nu_n z)]\}dK(z) \\
&+\mathcal{O}_p(n^{-3/4}\nu_n^{-1}) = q(\alpha) + \mathcal{O}_p((n\nu_n)^{-1/2}),
\end{aligned}$$

what proves (12.31). Applying the central limit theorem in the fourth line of (12.33), we arrive at (12.32).

Remark 2 *As an example of kernel satisfying **K1**, consider the Epanechnikov* (1969) *kernel with*

$$k(x) = \begin{cases} -\frac{3}{2b^3}\cdot x & \text{if} \quad -b \leq x \leq b \\ 0 & \text{elsewhere.} \end{cases}$$

The kernel estimate gets ahead of the histogram for $b > \frac{6}{5}$, when $\int K^2(x)dx = \frac{3}{5b} < \frac{1}{2}$.

12.5 Numerical Illustrations

In order to illustrate the differences of the averaged regression α-quantile and the location α-quantile for moderate samples we have conducted a simulation study. We considered the following linear regression model

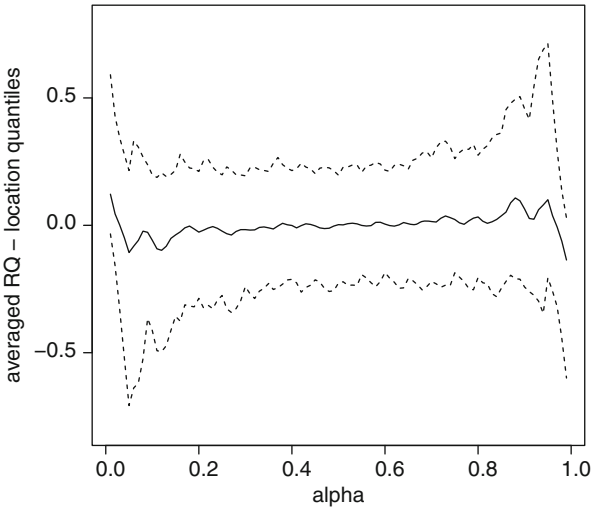


Fig. 12.1 The median, 5 %-, 95 %-quantiles in the sample of 10,000 differences between averaged regression and location α-quantiles in model (12.34); normal distributions of errors; sample sizes $n = 20$

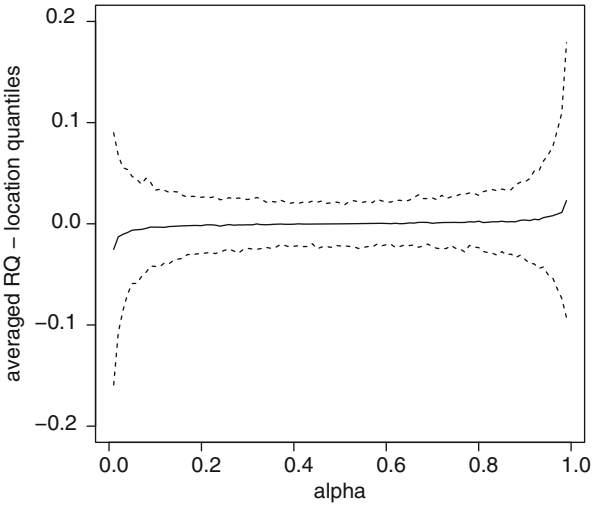


Fig. 12.2 The median, 5 %-, 95 %-quantiles in the sample of 10,000 differences between averaged regression and location α-quantiles in model (12.34); normal distributions of errors; sample sizes $n = 500$

$$Y_i = \beta_0 + x_i\beta_1 + U_i, \quad i = 1, \ldots, n, \tag{12.34}$$

The errors U_i, $i = 1, \ldots, n$, were simulated from the normal, exponential and Cauchy distributions. The design points $x_{1,1}, \ldots, x_{1,n}$ were generated from the uniform distribution on the interval $(-5, 50)$. They remain fixed for all simulations under given

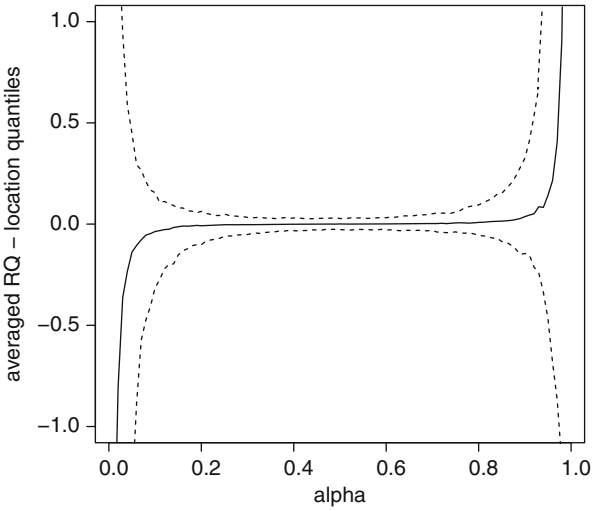


Fig. 12.3 The median, 5 %-, 95 %-quantiles in the sample of 10,000 differences between averaged regression and location α-quantiles in model (12.34); Cauchy distribution of errors; sample sizes n =500

Table 12.1 Mean, standard deviation and quantiles of difference between averaged regression and location 0.55-quantiles in model (12.34)

n, law	Mean	Stand. dev.	Quantiles						
			0	0.05	0.25	0.5	0.75	0.95	1
20, N	−0.155	0.280	−1.856	−0.639	−0.270	−0.098	−0.010	0.209	1.017
20, E	−0.006	0.049	−0.236	−0.081	−0.025	−0.006	0.008	0.071	0.303
20, C	−20.032	100.811	−1699.680	−76.138	−7.176	−1.875	−0.298	0.467	16.006
100, N	−0.033	0.102	−0.484	−0.202	−0.081	−0.025	0.018	0.132	0.368
100, E	0.000	0.013	−0.042	−0.018	−0.007	−0.001	0.005	0.025	0.055
100, C	−1.486	2.988	−38.802	−6.527	−1.803	−0.635	−0.074	0.656	4.648
500, N	−0.006	0.034	−0.143	−0.062	−0.023	−0.005	0.011	0.050	0.132
500, E	0.000	0.004	−0.013	−0.006	−0.002	0.000	0.002	0.007	0.021
500, C	−0.234	0.535	−3.737	−1.200	−0.440	−0.141	0.044	0.475	1.503

Sample sizes n =20, 100, 500, and 10,000 replications
N normal, E exponential, C Cauchy distributions of errors

n. The following parameter values of models were used: $n = 20,\ 100,\ 500$; $\beta_0 = 1$ and $\beta_1 = -2$.

Our interest is comparing the averaged regression α-quantiles and the location α-quantiles. We chose $\alpha = 0.05,\ 0.15,\ 0.55,\ 0.95$ and 10,000 replications of the models were simulated for each combination of the parameters and each α, and the averaged regression α-quantiles and the location α-quantiles were then computed. Figures 12.1–12.3 and Tables 12.1–12.3 compare some characteristics of differences of the averaged regression α-quantile and the location α-quantile for different combination of the parameters.

Table 12.2 Mean, standard deviation and quantiles of difference between averaged regression and location 0.05-quantiles in model (12.34)

n, law	Mean	Stand. dev.	Quantiles						
			0	0.05	0.25	0.5	0.75	0.95	1
20, N	−0.055	0.179	−0.773	−0.368	−0.128	−0.042	0.018	0.224	0.666
20, E	0.003	0.056	−0.182	−0.079	−0.024	−0.004	0.026	0.100	0.312
20, C	−0.788	2.107	−37.602	−3.385	−0.853	−0.247	−0.018	0.413	5.238
100, N	−0.016	0.061	−0.233	−0.117	−0.052	−0.014	0.016	0.087	0.210
100, E	0.001	0.019	−0.076	−0.025	−0.008	−0.001	0.009	0.036	0.078
100, C	−0.131	0.289	−1.947	−0.691	−0.220	−0.066	0.012	0.220	1.017
500, N	−0.001	0.019	−0.071	−0.032	−0.011	−0.001	0.008	0.029	0.079
500, E	0.000	0.006	−0.020	−0.008	−0.003	0.000	0.003	0.010	0.027
500, C	−0.021	0.070	−0.333	−0.140	−0.056	−0.013	0.015	0.083	0.260

Sample sizes n =20, 100, 500, and 10,000 replications
N normal, *E* exponential, *C* Cauchy distributions of errors

Table 12.3 Mean, standard deviation and quantiles of difference between averaged regression and location 0.15-quantiles in model (12.34)

n, law	Mean	Stand. dev.	Quantiles						
			0	0.05	0.25	0.5	0.75	0.95	1
20, N	0.065	0.179	−0.592	−0.215	−0.014	0.047	0.149	0.387	1.065
20, E	0.123	0.269	−0.871	−0.214	−0.008	0.073	0.230	0.669	1.453
20, C	0.880	2.314	−2.577	−0.378	0.024	0.260	0.875	3.899	32.680
100, N	0.012	0.061	−0.281	−0.089	−0.018	0.008	0.042	0.118	0.281
100, E	0.032	0.098	−0.275	−0.113	−0.019	0.021	0.078	0.196	0.579
100, C	0.139	0.325	−0.629	−0.217	−0.017	0.067	0.228	0.692	3.420
500, N	0.002	0.019	−0.060	−0.030	−0.008	0.003	0.013	0.034	0.085
500, E	0.007	0.031	−0.123	−0.042	−0.009	0.005	0.021	0.059	0.159
500, C	0.023	0.074	−0.206	−0.085	−0.015	0.013	0.057	0.152	0.401

Sample sizes n =20, 100, 500, and 10,000 replications
N normal, *E* exponential, *C* Cauchy distributions of errors

Acknowledgements The authors thank the Editors for their assistance and the Referee for the careful reading the text. The research of J. Jurečková was supported by the Czech Republic Grant P201/12/0083. The research of Jan Picek was supported by the Czech Republic Grant P209/10/2045.

References

Bloch DA, Gastwirth JL (1968) On a simple estimate of the reciprocal of the density function. Ann Math Statist 36:457–462

Bofinger E (1975) Estimation of a density function using the order statistics. Austral J Statist 17:1–7

Csörgö M, Révész P (1978) Strong approximation of the quantile process. Ann Statist 6:882–894

Dodge Y, Jurečková J (1995) Estimation of quantile density function based on regression quantiles. Stat Probab Lett 23:73–78

Epanechnikov VA (1969) Nonparametric estimation of a multivariate probability density. Theor Probab Appl 14:153–158

Falk M (1986) On the estimation of the quantile density function. Statist Probab Letters 4:69–73
Gutenbrunner C, Jurečková J (1992) Regression rank scores and regression quantiles. Ann Stat 20:305–330
Hájek J (1965). Extensions of the Kolmogorov-Smirnov tests to regression alternatives. Bernoulli-Bayes-Laplace Seminar, (ed. L. LeCam), University California Press, California, pp 45–60
Hájek J, Šidák Z (1967) Theory of rank tests. Academia, Prague
Hallin M, Jurečková J (1999). Optimal tests for autoregressive models based on autoregression rank scores. Ann Stat 27:1385–1414
Jaeckel LA (1972) Estimating regression coefficients by minimizing the dispersion of the residuals. Ann Math Stat 43:1449–1459
Jones MC (1992) Estimating densities, quantiles, quantile densities and density quantiles. Ann Inst Stat Math 44:721–727
Jurečková J, Picek J, Sen PK (2003) Goodness-of-fit tests with nuisance regression and scale. Metrika 58:235–258
Jurečková J, Picek J (2005) Two-step regression quantiles. Sankhya 67/2:227–252
Jurečková J, Picek J (2012) Regression quantiles and their two-step modifications. Stat Probab Lett 83:1111–1115
Jurečková J, Sen PK, Picek J (2012) Methodological tools in robust and nonparametric statistics. Chapman & Hall/CRC, Boca Raton
Koenker R (2005) Quantile regression. Cambridge University Press, Cambridge
Koenker R, Bassett G (1978) Regression quantiles. Econometrica 46:33–50
Koul HL (2002) Weighted empirical processes in dynamic nonlinear models. Lecture Notes in Statistics, vol 166, Springer, New York
Koul HL, Saleh AKMdE (1995) Autoregression quantiles and related rank-scores processes. Ann Statist 23:670–689
Lai TL, Robbins H, Yu KF (1983) Adaptive choice of mean or median in estimating the center of a symmetric distribution. Proc Nat Acad Sci USA 80:5803–5806
Parzen E (1979) Nonparametric statistical data modelling. J Am Stat Assoc 74:105–122
Parzen E (2004) Quantile probability and statistical data modeling. Stat Sci 19:652–662
Siddiqui MM (1960) Distribution of quantiles in samples from a bivariate population. J Res Nat Bur Standards 6411:145–150
Soni P, Dewan I, Jain K (2012) Nonparametric estimation of quantile density function. Comput Stat Data Anal 56:3876–3886
Welsh AH (1987) One-step L-estimators for the linear model. Ann Statist 15:626–641. Correction: Ann Stat (1988) 16:481
Welsh AH (1987) Kernel estimates of the sparsity function. In: Dodge Y (ed) Statistical data analysis based on the L1-norm and related methods. Elsevier, Amsterdam, pp 369–378
Xiang X (1995) Estimation of conditional quantile density function. J Nonparametr Stat 4:309–316
Yang SS (1985) A smooth nonparametric estimator of quantile function. J Am Stat Assoc 80: 1004–1011
Zeltermann D (1990) Smooth nonparametric estimation of the quantile function. J Stat Plan Infer 26:339–352

Chapter 13
A Study of One Null Array of Random Variables

Estate Khmaladze (with contribution from Thuong Nguyen)

13.1 The Sum

Suppose $U_1, \ldots, U_m$ are independent and uniformly distributed on $[0, 1]$, and let m be a large integer. Consider the sum

$$S_m = \sum_{i=1}^{m} U_i^m.$$

It certainly is sum of asymptotically negligible random variables with distribution function

$$\mathbb{P}(U^m \le s) = \mathbb{P}(U \le s^{1/m}) = s^{1/m}.$$

We want to say as much as we can about the limit distribution of S_m.

This limiting distribution is, certainly, infinitely divisible. In order to obtain the characteristic function of it, consider the characteristic function of the sum S_m:

$$\begin{aligned}[\phi_m(t)]^m &= \exp\,[m \ln \phi_m(t)] \sim \exp\,[m(\phi_m(t) - 1)] \\ &= \exp\,[m \int_0^1 (e^{its} - 1)\frac{1}{m} s^{1/m-1} ds]\end{aligned}$$

where

$$\phi_m(t) = \frac{1}{m} \int_0^1 e^{its} s^{1/m-1} ds$$

is characteristic function of one summand. This immediately implies that

$$\lim_{m\to\infty} [\phi_m(t)]^m = \exp\,[\lim_{m\to\infty} \int_0^1 \frac{e^{its} - 1}{s} s^{1/m} ds] = \exp \int_0^1 \frac{e^{its} - 1}{s} ds. \qquad (13.1)$$

E. Khmaladze (✉)
Victoria University of Wellington, Wellington, New Zealand
e-mail: Estate.Khmaladze@vuw.ac.nz

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_13,

Later, we use notation

$$\psi(t) = \int_0^1 \frac{e^{its} - 1}{s} ds,$$

so that the characteristic function of the limit distribution of S_m is

$$\exp[\psi(t)].$$

This, basically, means that the L 'evy - Khinchine measure of the limiting distribution is $ds/s, 0 < s < 1$.

Since the distribution function of U^m is a B- distribution function with the density

$$\frac{1}{B(\alpha, \beta)} s^{\alpha-1}(1-s)^{\beta-1}, \quad \text{where} \quad \alpha = 1/m, \beta = 1,$$

we have the representation of its characteristic function as an infinite series, see Johnson et al. (1995),

$$\phi_m(t) = 1 + \sum_{k=1}^{\infty} \frac{1/m}{1/m + k} \frac{(it)^k}{k!}.$$

Therefore, for our earlier limit we obtain

$$m(\phi_m(t) - 1) = \sum_{k=1}^{\infty} \frac{1}{1/m + k} \frac{(it)^k}{k!} \to \sum_{k=1}^{\infty} \frac{1}{k} \frac{(it)^k}{k!},$$

which is another expression for $\psi(t)$.

It is good to verify that the two expressions agree. To do this, it seems easiest to differentiate both expressions. Differentiating the infinite series, one obtains

$$\frac{d}{dt} \sum_{k=1}^{\infty} \frac{(it)^k}{k!} = \sum_{k=1}^{\infty} \frac{d}{dt} \frac{1}{k} \frac{(it)^k}{k!} = \sum_{k=1}^{\infty} \frac{1}{t} \frac{(it)^k}{k!} = \frac{1}{t}[e^{it} - 1].$$

Differentiating the integral form of $\psi(t)$, one obtains

$$\frac{d}{dt} \int_0^1 \frac{e^{its} - 1}{s} ds = i \int_0^1 e^{its} ds = \frac{1}{t}[e^{it} - 1],$$

and the derivatives coincide. To sum up: ψ can be written in two different forms,

$$\psi(t) = \int_0^1 \frac{e^{its} - 1}{s} ds = \sum_{k=1}^{\infty} \frac{1}{k} \frac{(it)^k}{k!} = \int_0^t \frac{1}{\tau}[e^{i\tau} - 1] d\tau.$$

13.2 Slightly Different View

Let us split the sum S_m as follows:

$$S_m = \sum_{i=1}^{m} U_i^m I\!I(U_i \geq 1-\epsilon) + \sum_{i=1}^{m} U_i^m I\!I(U_i < 1-\epsilon) = \sum_{i=1}^{m} U_i^m I\!I(U_i \geq 1-\epsilon) + o_P(1),$$

which is correct because the second sum is asymptotically negligible:

$$\sum_{i=1}^{m} U_i^m I\!I(U_i < 1-\epsilon) \leq m(1-\epsilon)^m \to 0, \quad \text{as} \quad m \to \infty.$$

This sum stays asymptotically negligible with slowly decreasing ϵ, say, $\epsilon \sim m^{-2/3}$. Rewrite the first sum as

$$\sum_{i=1}^{m} U_i^m I\!I(U_i \geq 1-\epsilon) = \sum_{i=1}^{m} e^{m \ln U_i} I\!I(U_i \geq 1-\epsilon)$$

$$= \sum_{i=1}^{m} e^{-m(1-U_i)} e^{mO(1-U_i)^2} I\!I(U_i \geq 1-\epsilon) \sim \sum_{i=1}^{m} e^{-m(1-U_i)} I\!I(U_i \geq 1-\epsilon),$$

which is true because $1 \leq e^{m(1-U_i)^2} \leq e^{m\epsilon^2} \to 1$ with our choice of ϵ.

Finally, we can drop $I\!I(U_i \geq 1-\epsilon)$, which will increase the last sum by asymptotically negligible random variable. Altogether

$$\sum_{i=1}^{m} U_i^m = \sum_{i=1}^{m} e^{-m(1-U_i)} + o_p(1). \tag{13.2}$$

This relationship may not be obvious initially, but becomes apparent when we turn to characteristic functions. The distribution function of each summand $e^{-m(1-U_i)}$ is

$$\mathbb{P}(e^{-m(1-U_i)} \leq x) = 1 + \frac{1}{m} \ln x, \quad \text{for} \quad x \in [e^{-m}, 1].$$

Therefore its characteristic function is

$$\varphi_m(t) = \frac{1}{m} \int_{e^{-m}}^{1} e^{its} \frac{1}{s} ds$$

and what 13.2 actually says is simply the probabilistic equivalent of the fact that

$$\lim_{m\to\infty} m(\phi_m(t) - 1) = \lim_{m\to\infty} m(\varphi_m(t) - 1)$$

or

$$\lim_{m\to\infty} \int_0^1 (e^{its} - 1) s^{1/m-1} ds = \lim_{m\to\infty} \int_{e^{-m}}^1 (e^{its} - 1) s^{-1} ds.$$

13.3 Two More Modifications of the Same Facts

Suppose the limit distribution of sum

$$S_m = \sum_{i=1}^{m} U_i^m$$

is some F. It certainly is a member of semigroup of distributions $\mathbb{F} = \{F_s(x), s \geq 0\}$ with convolution, see, e.g., Feller. We can always assume $F = F_1$. Infinitesimal operator of this semigroup of distributions is the operator defined as

$$\mathcal{U}a(x) = \lim_{m\to\infty} mE[a(x - U^m) - a(x)] = \lim_{m\to\infty} m \int_0^1 [a(x - y) - a(x)]dy^{1/m}.$$

Therefore

$$\mathcal{U}a(x) = \int_0^1 [a(x - y) - a(x)]\frac{dy}{y}.$$

The operator defined by the distribution F_s can be represented then as

$$\int_0^\infty a(x - y)dF_s(y) = e^{s\mathcal{U}}a = \sum_{k=0}^{\infty} \frac{s^k}{k!}\mathcal{U}^k a(x).$$

The power $\mathcal{U}^k$, and the operator $\mathcal{U}$ itself, is intuitively very appealing. Namely,

$$\mathcal{U}^k a(x) = \int_0^1 \cdots \int_0^1 \frac{\Delta_{y_1,y_2,\ldots,y_k} a(x)}{y_1 y_2 \ldots y_k} dy_1 dy_2 \ldots dy_k$$

where $\Delta_{y_1,y_2,\ldots,y_k} a(x)$ is the k-th increment of a, i.e.,

$$\Delta_{y_1} a(x) = a(x - y_1) - a(x)$$

is the first increment,

$$\Delta_{y_1,y_2} a(x) = a(x - y_1 - y_2) - a(x - y_1) - a(x - y_2) + a(x)$$

is the second increment, and so on. Since $y_1 y_2 \ldots y_k$ is the area of k-th increment, $\mathcal{U}^k a(x)$ is the "average size of k-th increment".

As we discovered in the previous sections, the $dK(y) = \mathbb{I}(0 < y < 1)\frac{dy}{y}$ is Lévy - Khinchine measure, corresponding to the semigroup $\mathbb{F}$, and its relationship to $\mathcal{U}$ is

$$\mathcal{U}a(x) = \int [a(x - y) - a(x)]dK(y),$$

which is true in general.

The second remark here is that

$$\sum_{i=1}^{m} U_i^m = m \int_0^1 x^m d\widehat{F}_m(x),$$

where $\widehat{F}_m$ is the empirical distribution function of m uniformly distributed random variables.

If we choose now $\varepsilon = \varepsilon_m = C/m$, with C – large but fixed, which is radically smaller than in the pervious section, we can proceed as

$$m \int_0^1 x^m d\widehat{F}_m(x) = m \int_0^{1-C/m} x^m d\widehat{F}_m(x) + m \int_{1-C/m}^1 x^m d\widehat{F}_m(x)$$

The first integral

$$m \int_0^{1-C/m} x^m d\widehat{F}_m(x) = \Sigma_{i=1}^m U_i^m \mathbb{I}(0 < U_i < \frac{C}{m})$$

has expected value

$$m \int_0^{1-C/m} x^m dx = \frac{m}{m+1}(1 - \frac{C}{m})^{m+1} \to e^{-C},$$

and the variance

$$m \left[\int_0^{1-C/m} x^{2\,m} dx - \left(\frac{1}{m+1}(1 - \frac{C}{m})^{m+1} \right)^2 \right] \to e^{-2\,C},$$

and both can be made arbitrarily small for sufficiently large C.

In the second integral, we can change the variable, $z = m(1-u)$ and consider the empirical distribution function $\widehat{F}_{m,Z}(z)$ of random variables $m(1 - U_i) = Z_i$, which are uniformly distributed on $[0, m]$:

$$m \int_{1-C/m}^1 x^m d\widehat{F}_m(x) = m \int_0^C (1 - \frac{z}{m})^m d\widehat{F}_{m,Z}(z)$$

However, on interval $[0, C]$, for any fixed C, the process

$$m\widehat{F}_{m,Z} \xrightarrow{d} \xi,$$

where ξ is standard Poisson process, see, e.g., Reiss and Thomas (2007) or Karr (2002). From this, it can be derived that

$$m \int_0^C (1 - \frac{z}{m})^m d\widehat{F}_{m,Z}(z) \xrightarrow{d} \int_0^C e^{-z} d\xi(z). \tag{13.3}$$

At the same time, the random variable

$$\int_C^\infty e^{-z} d\xi(z)$$

also has small expected value and variance; they are

$$\int_C^\infty e^{-z} dz = e^{-C} \text{ and } \int_C^\infty e^{-2z} dz = e^{-2\,C}.$$

Therefore,

$$m \int_{1-C/m}^1 x^m d\widehat{F}_m(x) \xrightarrow{d} \int_0^\infty e^{-z} d\xi(z),$$

and, altogether, we the following stochastic representation for the limiting random variable:

$$\sum_{i=1}^m U_i^m \xrightarrow{d} \int_0^\infty e^{-z} d\xi(z). \tag{13.4}$$

Limit theorem 13.3 is very suitable for simulations.

13.4 Distribution Function

We obtained the limit of characteristic function of our sum

$$S_m = \sum_{i=1}^m U_i^m$$

and the stochastic representation 13.4 of this limit. Let us use it now to derive what we can for distribution function of this limit.

We first obtain two equations for the distribution function, although we will not use them. We have

$$X = \int_0^\infty e^{-z} d\xi(z) = \sum_{i=1}^\infty e^{-T_i},$$

where T_i is the moment of i-th jump of ξ (or i-th arrival time). Then

$$\sum_{i=1}^\infty e^{-T_i} = e^{-T_1}[1 + \sum_{i=2}^\infty e^{-T_i+T_1}].$$

Since T_1 is standard exponential random variable, e^{-T_1} is uniform random variable on $[0, 1]$, independent from the infinite sum on the right hand side, while two infinite

sums have the same distribution. In notation,

$$X = U[1 + X'], \quad X \stackrel{d}{=} X', \; U \perp X'.$$

This implies

$$\mathbb{P}(X \le x) = \mathbb{P}(X' \le \frac{x}{U} - 1)$$

or

$$F(x) = \int_0^1 F(\frac{x}{u} - 1)du.$$

This also implies

$$\mathbb{P}(\frac{X}{U} \le x) = \mathbb{P}(1 + X' < x),$$

or

$$\int_0^1 F(x\,u)du = F(x - 1).$$

Both equations define F uniquely, but we find it easier to use another equation later. This equation is recursive, and leads to explicit expression for the limiting distribution function.

Start with

$$\int_0^\infty e^{-z}d\xi(z) = \int_0^\varepsilon e^{-z}d\xi(z) + e^{-\varepsilon}\int_\varepsilon^\infty e^{-z+\varepsilon}d\xi(z),$$

or, in different but similar notations,

$$X = \int_0^\varepsilon e^{-z}d\xi(z) + e^{-\varepsilon}X'$$

For the integral, we have

$$\mathbb{P}(\int_0^\varepsilon e^{-z}d\xi(z) - e^{-T_1}I\!I(T_1 \le \varepsilon) \ne 0) = \mathbb{P}(T_2 \le \varepsilon) = O(\varepsilon^2)$$

and therefore

$$\mathbb{P}(X \le x) = \mathbb{P}(X' \le e^{\varepsilon}(x - e^{-T_1}I\!I(T_1 \le \varepsilon))) = F(e^{\varepsilon}x)e^{-\varepsilon} + \int_{e^{-\varepsilon}}^1 F(e^{\varepsilon}(x - u))du,$$

or

$$F(x) = (F(x) + f(x)x\varepsilon)(1 - \varepsilon) + F(x - 1)\varepsilon + o(\varepsilon),$$

and, finally,

$$xf(x) = F(x) - F(x - 1).$$

This can be solved recurrently: $F(x) = 0$ for $x \le 0$ and then

$$F(x) = x[c - \int_0^x \frac{1}{y^2}F(y - 1)dy]. \tag{13.5}$$

Namely, for $0 \le x \le 1$, $F(x)$ is just linear,

$$F(x) = cx, \quad f(x) = c;$$

for $1 \le x \le 2$

$$F(x) = c[2x - x\ln x - 1], \quad f(x) = c[1 - \ln x].$$

For $2 \le x \le 3$, Thuong Nguyen derived

$$F(x) = c[3x - x\ln x - 3 + \ln(x-1) - x\ln(x-1) + x\int_2^x \frac{\ln(y-1)}{y}dy].$$

To derive an analytic expression beyond $x = 3$ is possible, but looks like somewhat unnecessary hassle, because $F(3) = 0.988$, quite high value already. The value of the constant c is about 0.563.

13.5 The Stretched Sum

Now, as $m \to \infty$, let $a_m \to \infty$ but $a_m^{1/m} \to 1$. There are many such sequences. For example, $a_m = m^\alpha$ for any constant $\alpha > 0$, has this property. Consider random variable

$$a_m U^m$$

It is asymptotically negligible: for any $x > 0$

$$\mathbb{P}(a_m U^m < x) = \min(\frac{x^{1/m}}{a_m^{1/m}}, 1) \to 1.$$

Still, its moments can very well diverge to ∞, or remain bounded away from zero, because

$$Ea_m U^m = a_m \frac{1}{m+1}, \quad Var\, a_m U^m = a_m^2 \frac{1}{2m+1}\left(\frac{m}{m+1}\right)^2.$$

How small is $a_m U^m$? To answer this, note that

$$\mathbb{P}(a_m U^m < z^m) = \min(\frac{z}{a_m^{1/m}}, 1) \to \begin{cases} 1, \text{ if } z > 1, \\ z, \text{ if } z \le 1. \end{cases}$$

and that

$$\mathbb{P}(U^m < z^m) = \min(z, 1) = \begin{cases} 1, \text{ if } z > 1, \\ z, \text{ if } z \le 1. \end{cases}$$

Therefore, on the scale z^m random variables $a_m U^m$ and U^m look equally small.

Can we say something about the limit behavior of the sum

$$R_m = \sum_{i=1}^{c_m} a_m U_i^m = a_m S_{c_m},$$

and what should be the choice of c_m? In the previous case it was $c_m = m$, but now—not clear.

To find the limit for the characteristic function of R_m is the same as to find the limit for $c_m[\phi_m(ta_m) - 1]$. We have:

$$c_m[\phi_m(ta_m) - 1] = c_m \int_0^1 [e^{it\, a_m s} - 1] ds^{1/m} = \frac{c_m}{m} \frac{1}{(a_m)^{1/m}} \int_0^{a_m} \frac{e^{ity} - 1}{y} y^{1/m} dy$$

Here, as $m \to \infty$, the integral tends to ∞, while $(a_m)^{1/m} \to 1$. In order to obtain a limit of the form

$$\int_0^\infty \frac{e^{ity} - 1}{y} dM(y)$$

the measure

$$dM_m(y) = \frac{c_m}{m} y^{1/m} dy, \quad y \le a_m,$$

should have a weak limit $dM(y)$ and $1/y$ should be integrable with respect to this limit. But this can not happen. The sequence of normalized measures

$$\frac{m}{c_m a_m} dM_m(y),$$

which are probability distributions on $[0, a_m]$, "runs away" from the space.

13.6 Notes and Acknowledgment

This work was motivated by authors interest to null-arrays of positive random variables and their limiting Lévy measures. In the author's view, they should play an essential role in the statistical analysis of large number of rare events (statistical theory of diversity), see Khmaladze (1989). Hopefully, this will be demonstrated in subsequent publication(s).

Representation 13.1 describes the limiting infinitely divisible distribution as self-decomposable. This also follows from stochastic representation 13.4—see for these results Sato (2002), pp. 109–112. The fact that the density of F is constant on some interval $[0, a]$ is known in much more general situation, described in Sato (2002), Lemma 53.2. At the same time, instead of looking for an aggregate general statement it may be easier to derive the fact directly.

Limiting characteristic function 13.1 appears in the literature in several contexts. For example, it was used as early as 1957 by I.A. Ibragimov (1957) in an attempt to construct example of nonunimodal density of self-decomposable distribution.

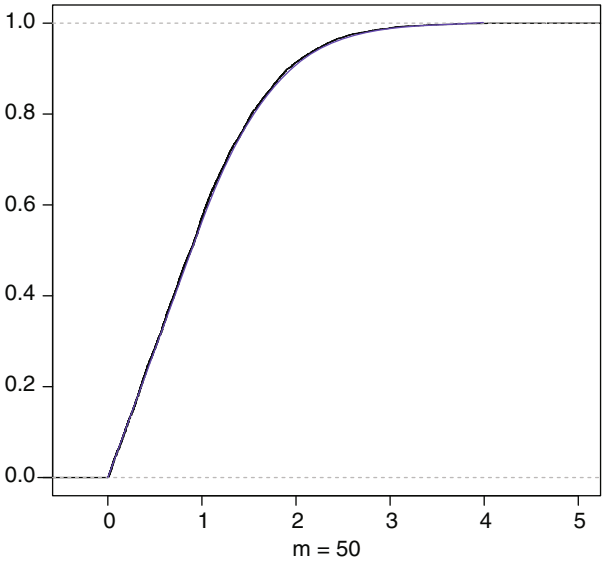


Fig. 13.1 The graph of the distribution function F along with computer simulated distribution function of S_m with $m = 50$. The difference is barely visible. For m as small as 30 the difference between the two is noticeable. In the interval $[3, 4]$ the recurrence 13.5 was used for calculation of F.

I am obliged for these remarks to Ken-iti Sato and Jean Bertoin, who kindly agreed to read the text and provide a feedback.

Acknowledgements Hira L. Koul is not only an old friend but also coauthor. It was a great experience working with him. The note I submitted here is just a beginning of research in the theory of statistical diversity. I hope to work on this with him in the future.

References

Feller W (1966) An introduction to probability theory and its applications, vol. 2, Wiley.
Ibragimov IA (1957) A remark on probabilty distribution of class L, Th Probab Appl, 1957, 117–119
Johnson NL, Kotz S, Balakrishnan N (1995) Continuous univariate distributions Vol. 2 (2nd ed.). Wiley,
Karr A (2002) Point processes and their statistical inference, Marcel Dekker, Inc.
Khmaladze EV (1989) Statistical analysis of large number of rare events, Center for Mathematics and Informatics (CWI), Report
Khmaladze EV(2011) Convergence properties in certain occupancy problems including the Karlin-Rouault Law, J Appl Prob (2011), 48:1095–1113
Reiss RD, Thomas M (2007) Statistical analysis of extreme values with applications to insurance, finance, hydrology and other fields, Birkhaüser-Verlag
Sato K (2002) Lévy processes and infinitely divisible distributions, Cambridge Univesity Press

Chapter 14
Frailty, Profile Likelihood, and Medfly Mortality

Roger Koenker and Jiaying Gu

14.1 Introduction

The notion of frailty to describe unobserved heterogeneity of population risks has become a familiar feature of demographic analysis since Vaupel et al. (1979), and has gradually spread to other statistical domains. A valuable early exposition of the impact of frailty in models of treatment evaluation is provided by Shepard and Zeckhauser (1980). Often, as in the aforementioned sources, parametric models are posited for the frailty effects, but it is usually difficult to justify such assumptions given the unobserved nature of the frailty components. Recent progress in estimation and inference for general, nonparametric mixture models has opened the way to a more flexible approach. We will illustrate some features of such an approach with a reanalysis of the influential Carey et al. (1992) study of medfly mortality.

14.2 Data

In the largest of the three experiments reported in Carey et al. (1992), 1.2 million Mediterranean fruit flies (*Ceratitis capitata*) were raised in a large facility in Mexico,

> ... Pupae were sorted into one of five size classes using a pupal sorter. This enabled size dimorphism to be eliminated as a potential source of sex-specific mortality differences. Approximately, 7,200 medflies (both sexes) of a given size class were maintained in each of 167 mesh covered, 15 cm by 60 cm by 90 cm aluminum cages. Adults were given a diet of sugar and water, ad libitum, and each day dead flies were removed, counted and their sex determined ...

The primary objective of the experiment was to study the upper tail of the mortality distribution, an endeavor that revealed several surprising features.

R. Koenker (✉) · J. Gu
University of Illinois, IL 61801 Urbana, USA
e-mail: rkoenker@uiuc.edu

J. Gu
e-mail: gu17@uiuc.edu

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_14,

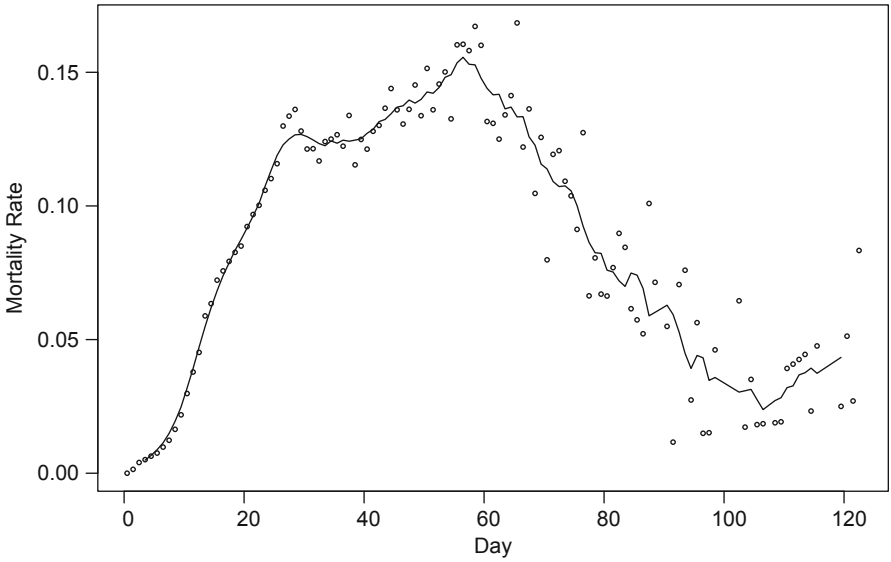


Fig. 14.1 Raw daily medfly mortality rates and moving average smooth

14.3 Declining Mortality Rates

Prior to this experiment it was an article of faith throughout biology that within species mortality (hazard) rates were monotonically increasing with age. Indeed it was commonly suggested that each species had a species specific upper bound for age rendering the whole notion of investigating the "tail behavior" of the mortality distribution pointless. In Fig. 14.1 we plot raw daily mortality rates from the experiment and superimpose a smoothed, geometric moving average curve. More explicitly, let y_t denote the number of flies alive (at risk) at the beginning of day t, then the raw mortality rates plotted in Fig. 14.1 are, $r_t = 1 - y_{t+1}/y_t$, and the smoothed (geometric) weekly moving average. Contrary to the received wisdom, mortality rates actually declined after about age 60. This finding provoked an extensive reappraisal of the biology of aging. The observed decline in mortality offered no consolation to the 99.8 % of the flies that were already dead by age 60, but to the remaining, more than 2000 less frail ones, it offered some hope of a prolonged retirement. The oldest flies in the experiment expired on day 172.

How should we interpret this remarkably long tail? One explanation, suggested by Vaupel and Carey (1998), was that the population under study was really a mixture of several subpopulations of varying frailties. Rather than assume a particular parametric form for the mixing distribution, Vaupel and Carey adopted a nonparametric mixture model. While their two-page note in *Science* precluded a detailed description of their computational methods, we have been able to "reverse engineer" an approach that closely mimics the results reported in their Figure 1.

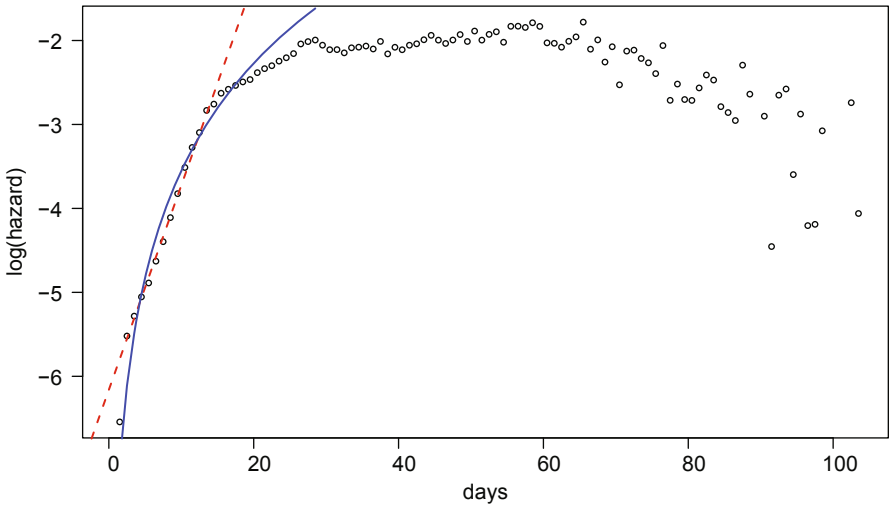


Fig. 14.2 Estimated baseline Gompertz and Weibull hazard models: linear (Gompertz) and log linear (Weibull) fits to the initial k observations of raw daily log mortality rates

The first question is: What are we mixing? Here we follow Vaupel and Carey and consider both Gompertz and Weibull mixtures. The Gompertz model assumes that log hazard is linear in age, while the Weibull model assumes that log hazard is linear in log age. Figure 14.2 illustrates raw log-hazard rates plotted against age, and superimposed are two estimates of the baseline model. The dashed line represents the estimated baseline Gompertz model fit to the data for the first 15 days of the experiment by weighted least squares, with weights given by the relative frequencies of the daily counts. It appears that the first day is an outlier in this plot, however since few flies died on the first day, it exerts little influence on the fitted line. The solid curve represents the baseline Weibull fit based on the first 20 days of the experiment. How many observations to use to estimate the parameters of the baseline model is obviously somewhat debatable, in this respect the problem is somewhat similar to the notorious controversies over how to choose k in the Hill estimator of the Pareto exponent. We will not indulge in further speculation about these choices, but simply remark that our k selection yields baseline Gompertz hazard of $h(t) = 0.002 \exp(0.24t)$, while Vaupel and Carey use $h(t) = 0.003 \exp(0.3t)$, and for the Weibull model we obtain $h(t) = 0.0004t^{1.85}$, against Vaupel and Carey's $h(t) = 0.001t^2$. The intercept in these models is not crucial since the estimated mixture distribution is scale-equivariant. It simply fixes a normalization. The shape parameter is more important, but in both cases our approach of fitting the left tail of the distribution yields rather similar estimates to those employed by Vaupel and Carey. An intriguing, open theoretical and practical question remains: can likelihood methods be brought to bear to estimate these shape parameters. We will return to this question when we consider profile likelihoods.

Given our estimated baseline models it is now time to address the problem of estimating the mixing or frailty distribution. There is a long history and extensive literature on this subject. Lindsay (1995) provides a thorough overview. Kiefer and Wolfowitz (1956) demonstrated that such mixture models were consistently estimable under weak conditions by maximum likelihood. If we write the baseline density as $\varphi(x, \theta)$ and the mixture density as,

$$g(x) = \int \varphi(x, \theta) dF(\theta),$$

then given independent and identically distributed (iid) observations, $x_1, \cdots, x_n$ from g, we wish to solve

$$\max_{F \in \mathcal{F}} \sum_{i=1}^{n} \log(g(x_i)).$$

Following Laird (1978), the expectation-maximization (EM) algorithm, or a variant of it, has been employed to solve such problems. However, EM is notoriously slow to converge. Koenker and Mizera (2013) proposed an alternative computational strategy based on convex optimization. Let, $t_0 < t_1 < \cdots < t_m$ denote a grid of values for the potential mass points of the distribution F, and let f_i denote the mass associated with the ith grid interval. Then, we can rewrite the MLE problem as,

$$\max_{f \in \mathbb{R}^m} \left\{ \sum \log(g(x_i)) \mid g = Af, \ \sum f_i \Delta t_i = 1, \ f \geq 0 \right\},$$

where A denotes the n by m matrix with typical element, $\varphi(x_i, t_j)$, and g denote the n vector with typical element $g(x_i)$. This is a garden variety convex optimization problem that can be efficiently solved by modern interior point methods. We employ MOSEK (Andersen 2010) for this purpose. The R package, REBayes (Koenker 2012), implements a variety of related problems, all of the computational results reported here were carried out in this environment.

In Fig. 14.3 we plot the two mixing distributions estimated by the Kiefer–Wolfowitz maximum likelihood procedure. Note that the vertical axis in these plots is the cube root of the density to exaggerate the smaller mass points that are nearly invisible on the original $f(\theta)$ scale. The Kiefer–Wolfowitz estimator is known to deliver a discrete distribution, here represented by a "density" with a small number of "almost" point masses. The Weibull model is considerably more parsimonious in this respect with only six distinct points of support. The implied hazard functions for the two estimated mixture densities are shown in Fig. 14.4, superimposed over the raw mortality rates. Fewer mass points in the Weibull model translates to much smoother behavior of the hazard function, but this is ultimately traceable back to the forms of the base density, the Gompertz being more sharply peaked and consequently generating a rougher mixture. In both cases the mixing parameter θ functions as a scale parameter, but the mixing distribution is estimated on the $\log \theta$ scale, so we can interpret the mixing as convolution as with the familiar kernel density estimator.

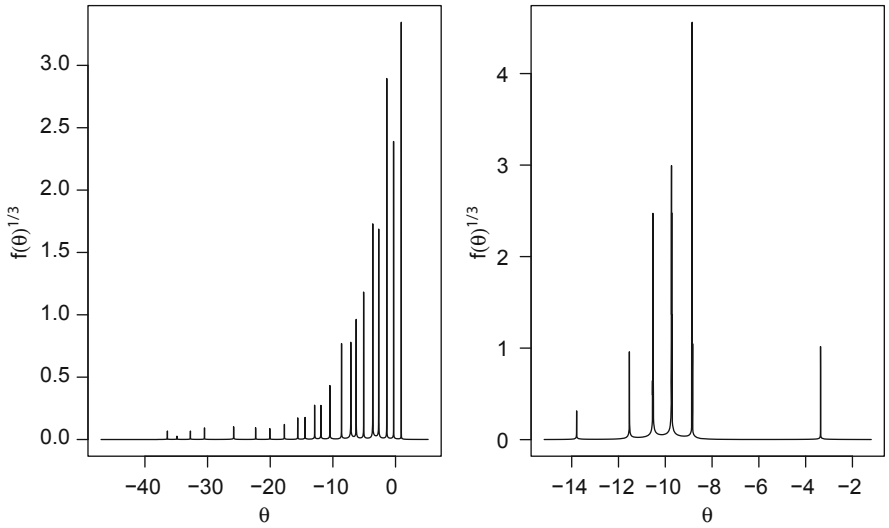


Fig. 14.3 Estimated mixing distributions for the Gompertz (*left*) and Weibull (*right*) models

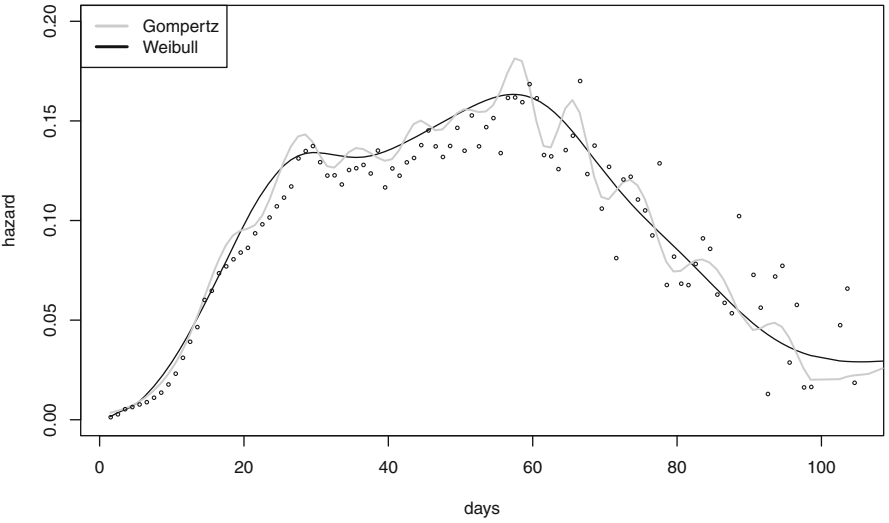


Fig. 14.4 Hazard functions for the estimated Gompertz and Weibull models

14.4 Gender Crossover

An obvious source of *observed* heterogeneity is gender differences. Again, the Carey et al. experiment revealed some surprising new facts. When we repeat our prior exercise fitting separate baseline Weibull models for males and females, we obtain the results appearing in Fig. 14.5. The Weibull model fits considerably better in both

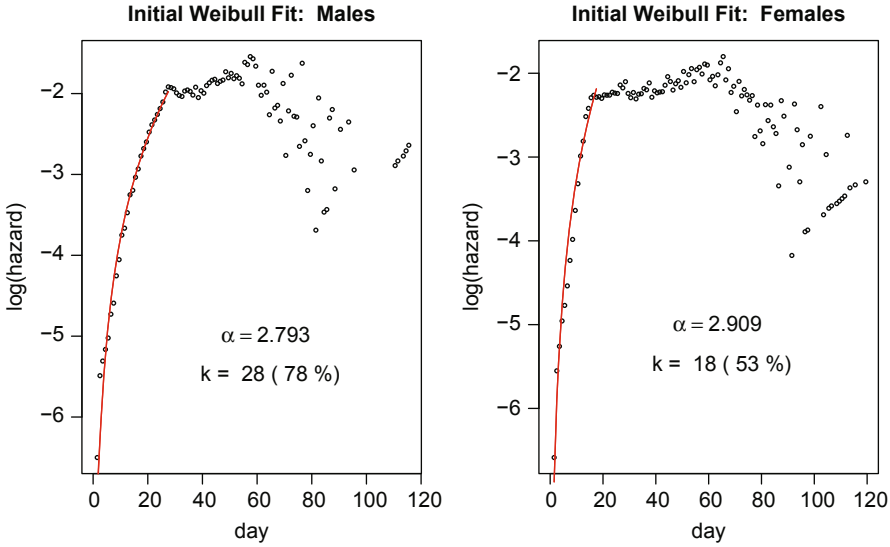


Fig. 14.5 Gender specific of baseline Weibull models: weighted least squares fitting of the initial k observations on daily mortality rates. The percentage of the sample population dead by day k is given in parentheses. The estimated shape parameter of the baseline Weibull model is α

of these plots than in the previous aggregated plot, and considerably better than the corresponding Gompertz plots, so we will restrict attention henceforth to the Weibull model. Given the baseline models the Kiefer–Wolfowitz estimates of the mixture model yields the gender-specific hazard functions of Fig. 14.6. Several features of this plot are worth noting. Until about age 20, female mortality is higher than that of males, but after age 20, female mortality is substantially below male mortality. This crossover of the hazard functions clearly contradicts the proportional hazard assumption that is frequently made in survival analysis. The second crossover of the estimated hazard curves at about age 75 probably shouldn't be taken too seriously, but the initial crossing is quite precisely estimated and induces a crossing of the estimated gender-specific survival functions at about age 36. It is impossible to resist noting that this pattern reverses the typical finding for human populations for which males are more frail than females with a possible crossover only at very advanced ages.

14.5 Profile Likelihood and Covariate Effects

If nonparametric maximum likelihood estimation of frailty effects were restricted to univariate survival models, it would still be a very valuable addition to the statistical repertoire, but it would be much more useful if it could be extended to semiparametric applications including covariate effects. Of course we already have the proportional hazard model for this purpose, however frailty offers another valuable perspective.

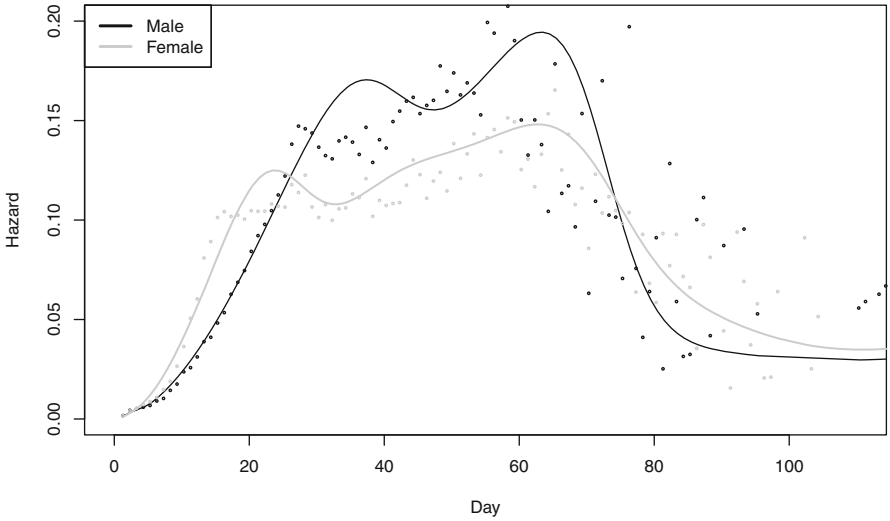


Fig. 14.6 Gender Specific of hazard functions for the Weibull Mixture model: raw daily mortality rates are plotted in black for males and red for females, superimposed are the estimated hazard functions for the Weibull mixture models using the baseline models shown in Fig. 14.6

Factorization of the likelihood makes the proportional hazard assumption especially convenient from a computational viewpoint. The Weibull mixture model has no comparable factorization; nevertheless, it is possible to employ a profile likelihood formulation to elaborate the model to include covariate effects.

From the beginning a controversial aspect of the Carey experiment was the effect of cage density. Critics claimed that flies raised in more crowded cages would be more likely to die earlier. Carey et al. (1993) responded that the cage density was quite low after 60 days, only 16 flies per cage, on average, survived beyond this age, so it seemed difficult to attribute differences in mortality rates in elderly medflies to differences in crowding. To investigate whether differences in initial cage density had a significant impact on mortality we considered a model in which it entered as a linear multiplicative scale shift in the Weibull model, that is, the baseline Weibull scale becomes $\theta_0 \exp(d_i \beta)$ where d_i denotes initial cage density. To estimate the density effect parameter, β, we simply evaluated the profiled likelihood on a grid of values on the interval $[-1, 1]$, yielding Fig. 14.7. This exercise yields a point estimate of about $\hat{\beta} = -0.5$ that is quite precise, at least if we are to believe the confidence bounds implied by the classical Wilks, $2 \log \lambda \rightsquigarrow \chi^2_1$, theory. Leaving the reliability of such intervals to future investigation, we conclude simply that the negative estimated coefficient implies that higher density shifts the survival distribution to the right, thus prolonging lifetimes, and directly contradicting the conjecture of the Carey critics. This finding is confirmed by other methods, see for example Koenker and Geling (2001), where similar results are reported for both the Cox model and several quantile regression models.

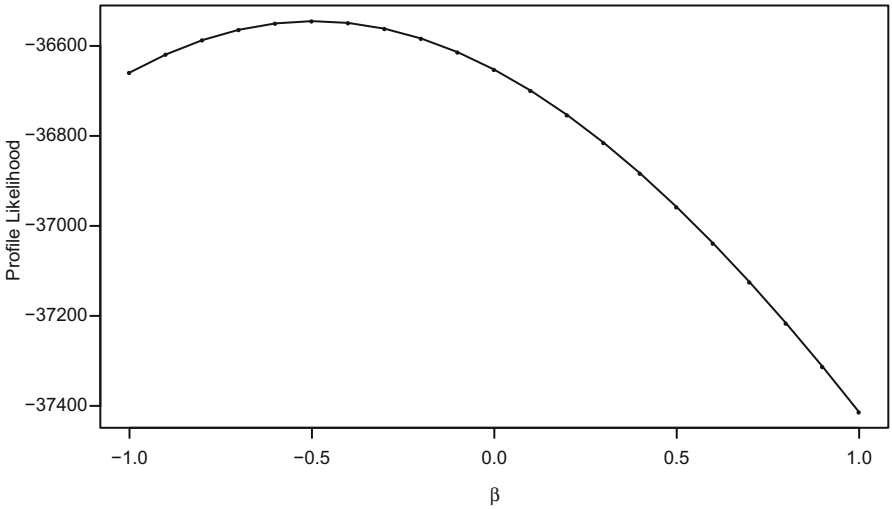


Fig. 14.7 Profile likelihood for the initial cage density effect in the Weibull mixture model

The success of profile likelihood in a few cases prompts one to wonder how far similar methods can be extended to other semiparametric mixture settings. There is a considerable literature on this topic, pioneered by Lindsay. When profiling leads to fully adaptive estimation of structural parameters, not only do we get efficient estimates of those parameters, as a by-product we also get valid inference from the profiled likelihood ratio statistic, see Murphy and Van der Vaart (2000). The latter bonus is sometimes referred to as the Wilks phenomenon, e.g., Fan et al. (2001).

But profiling is not always so perceptive; sometimes it can lead the unwary toward disaster. To illustrate this less benign side of profile likelihood for mixture models, we would like to briefly reconsider estimation of the Weibull shape parameter, α, based on the medfly data. In Fig. 14.8 we show the profile likelihood for α based on the full medfly data. Based on our earlier results we know that $\alpha \approx 2.8$ fits the initial portion of the log hazard plot quite well. What does the profile likelihood have to say about it? The message is a bit confusing: the profile likelihood increases sharply up to about $\alpha = 2.8$, and then dramatically flattens out. In fact, closer examination reveals that the profile likelihood continues to increase beyond this value, but very, very gradually. Indeed, as $\alpha \to \infty$, the profile likelihood also tends to infinity. To understand this better it is helpful to consider how the estimated mixture distribution responds to changes in α. For small α, the estimated mixture distribution has only a single mass point, and this single mass point persists for a while, by the time we get to α between 2.5 and 3.0 though we have 5 or 6 mass points as in Fig. 14.3. As α becomes larger we get more and more mass points, eventually yielding positive mass corresponding to virtually all the distinct observed values. This is reminiscent of the familiar Dirac catastrophe produced by kernel density bandwidths chosen by maximum likelihood. Indeed, the situation is quite similar, as α becomes large the effective bandwidth of the baseline Weibull model becomes narrower and more mass points are needed in the mixture distribution to mimic the density of the observed data.

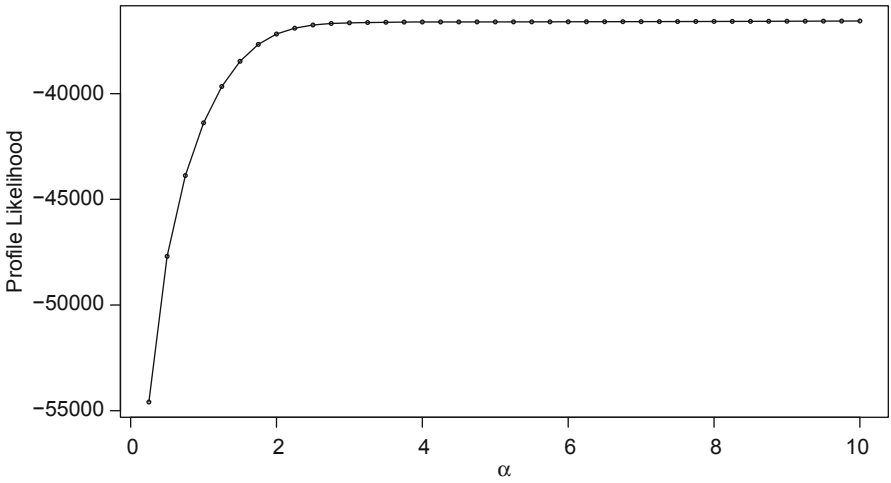


Fig. 14.8 Profile likelihood for the shape parameter in the Weibull mixture model

So profile likelihood has failed us. Now what? There is a familiar litany of circumstances in which naive adherence to the principle of maximum likelihood leads to absurd results: various Gaussian examples in which driving variance parameters to zero yields unbounded likelihood at unlikely places in parameter space, estimation of the threshold parameter of the three-parameter lognormal distribution, and many others. One approach that has proven successful in such situations is the maximum product-spacing methods introduced by Cheng and Amin (1983) and Ranneby (1984). Roeder (1990) describes an application of this approach in astronomy that although based on Gaussian assumptions is qualitatively quite similar to our Weibull problem.

Log product spacings optimization can be viewed as a discretization of classical maximum likelihood. Let $G(x, \theta)$ denote the distribution function of a parametric model for a scalar random variable, X. Given a sample, $X_1, \cdots, X_n$ of identical copies of X, let

$$\Delta G_i(\theta) = G(X_{(i)}, \theta) - G(X_{(i-1)}, \theta),$$

for $i = 1, \ldots, n+1$ with $X_{(0)} = -\infty$ and $X_{(n+1)} = +\infty$ and $X_{(i)} : \ i = 1, \cdots, n$ denoting the order statistics of the original sample. Since $G(X, \theta_0)$ is uniform when evaluated at the true parameter, θ_0 of the model, the $\Delta G_i(\theta_0)$ constitute a sample of uniform spacings for which there is an extensive theory. Considering

$$R_n(\theta) = \frac{1}{\sqrt{n+1}} \sum_{i=1}^{n+1} (\log(\Delta G_i(\theta)(n+1)) + \gamma)/(\pi^2/6 - 1)^{1/2}$$

with $\gamma \approx 0.577216$, the Euler constant, we have a normalized sum that satisfies a central limit theorem with a standard normal limiting distribution. Maximizing $R_n(\theta)$

with respect to θ requires computing the Kiefer–Wolfowitz mixture distribution, $\hat{G}(x, \theta)$, at each θ to obtain the profile log product spacing objective function. The function $R_n(\theta)$ behaves like the usual log-likelihood; this is to be expected since the summands can be viewed as difference quotient approximations of $g(\tilde{x}_i, \theta)$ for $\tilde{x}_i \in (X_{(i-1)}, X_{(i)})$. However, by avoiding the direct evaluation of densities we circumvent the pathological behavior of the log likelihood.

An important feature of the maximum product spacing method noted by Roeder (1990), is that for given θ, it selects an $\hat{G}(x, \theta)$ that is asymptotically equivalent to the mixture distribution estimated by nonparametric maximum likelihood, that we have focused on thus far. For θ taking various values, we get a profiled objective function similar to the profiled nonparametric likelihood. Yet unlike the problematic profiled likelihood, the limiting form of $R_n(\theta)$ yields an estimating function centered at zero for the true parameter and a simple confidence interval construction for the structural parameter. Further details regarding the maximum product spacing method can be found in Roeder (1990), Roeder (1992), and Ekström (2008).

We have seen already that an α parameter that fits the left tail of the survival distribution can be estimated well by a simple regression of log hazards on log event times using data from the first few days of the experiment. This assumes that flies that only survive for the first k days are all from a homogeneous parametric survival model. When we move on to the semiparametric mixture model using all the observations, a natural question becomes how reasonable is it to assume a global value for α while allowing scale heterogeneity with frailty. We employ a first-order form of the log-product-spacing method and find that the test strongly rejects the mixture models with $\alpha = 2.85$. However, when we use only the observations surviving up to 50 days, a subsample that actually contains 99.5 % of the full sample, we obtain a test statistic of only 0.33 and the model is not rejected. Similar conclusions are drawn when we estimate gender-specific models. The message seems to be that the Weibull semiparametric mixture model fits the majority of the data quite well, but fails to perform adequately for the extreme right tail.

This conclusion may simply reassert that estimating a fixed shape parameter in the Weibull mixture model is an extremely difficult task; this is indeed the impression one gets from the prior literature. Hahn (1994) shows that the information matrix is singular for mixed Weibull proportional hazard model. When there are no covariates, the score function for α is identically zero, hence also the Fisher information. This means that the Weibull shape parameter can not be estimated at a root-n rate. Various estimation strategies for α are nevertheless available, for example, Honoré (1990), Honoré (1997), and Ishwaran (1996). We would like to highlight what seems to be a somewhat neglected paper by Ishwaran (1999) discussing the information loss phenomenon for a class of semiparametric mixture models. Ishwaran shows that for the Weibull mixture model, there is information loss for αs bigger than the true value α_0, so that with $\alpha > \alpha_0$, one can find a mixing distribution that produces a model that is arbitrarily close to the true model in the sense of Hellinger distance. This corresponds to the flat region in our profile likelihood. On the other hand, as he notes, it is curious that the same information loss phenomenon does not occur for α's that are smaller than α_0. Whether one could take advantage of this asymmetric behavior for estimation of α is left for future investigations.

Acknowledgements It is a pleasure to acknowledge the inspiration that Hira Koul's work has provided over the years; his enthusiasm for statistics, indeed for life more generally, is evident in his writings, his talks, and in his dancing. The authors wish to thank Olga Geling for stimulating their initial interest in the medfly experiment, and James Carey for sharing the data. This research was partially supported by NSF Grant 11-53548.

References

Andersen ED (2010) The MOSEK optimization tools manual, version 6.0, available from http://www.mosek.com

Carey J, Liedo P, Orozco D, Vaupel J (1992) Slowing of mortality rates at older ages in large medfly cohorts. Science 258:457–61

Carey J, Curtsinger J, Vaupel J (1993) Fruit fly aging and mortality. response to letters to the editor. Science 260:1567–1569

Cheng R, Amin N (1983) Estimating parameters in continuous univariate distributions with a shifted origin. Journal of the Royal Statistical Society Series B (Methodological) pp 394–403

Ekström M (2008) Alternatives to maximum likelihood estimation based on spacings and the kullback–leibler divergence. Journal of Statistical Planning and Inference 138:1778–1791

Fan J, Zhang C, Zhang J (2001) Generalized likelihood ratio statistics and wilks phenomenon. The Annals of Statistics 29:153–193

Honoré B (1990) Simple estimation of a duration model with unobserved heterogeneity. Econometrica 58:453–473

Honoré B (1997) A note on the rate of convergence of estimators of mixtures of weibulls, preprint, Princeton

Ishwaran H (1996) Uniform rates of estimation in the semiparametric weibull mixture models. Annals of Statistics 24:1560–1571

Ishwaran H (1999) Information in semiparametric mixtures of exponential families. Annals of Statistics 27:159–177

Kiefer J, Wolfowitz J (1956) Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. The Annals of Mathematical Statistics 27:887–906

Koenker R (2012) REBayes: An R package for empirical Bayes methods, available from http://www.econ.uiuc.edu/roger/research/ebayes/ebayes.html

Koenker R, Geling O (2001) Reappraising medfly longevity: a quantile regression survival analysis. J of Am Stat Assoc 96:458–468

Koenker R, Mizera I (2013) Convex optimization, shape constraints, compound decisions and empirical bayes rules, http://www.econ.uiuc.edu/roger/research/ebayes/ebayes.html

Laird N (1978) Nonparametric maximum likelihood estimation of a mixing distribution. J Amer Statistical Assoc 73:805–811

Lindsay B (1995) Mixture models: theory, geometry and applications. In: NSF-CBMS regional conference series in probability and statistics

Murphy SA, Van der Vaart AW (2000) On profile likelihood. J Amer Statistical Assoc 95:449–465

Ranneby B (1984) The maximum spacing method. an estimation method related to the maximum likelihood method. Scandinavian Journal of Statistics pp 93–112

Roeder K (1990) Density estimation with confidence sets exemplified by superclusters and voids in the galaxies. J Amer Statistical Assoc 85:617–624

Roeder K (1992) Semiparametric estimation of normal mixture densities. Annals of Statistics 20:929–943

Shepard D, Zeckhauser R (1980) Long-term effects of interventions to improve survival in mixed populations. J Chronic Disease 33:413–433

Vaupel J, Carey J (1998) Compositional interpretations of medfly mortality. Science 260:1666–1667

Vaupel J, Manton K, Stollard E (1979) The impact of heterogeneity in individual frailty on the dynamics of mortality. Demography 16:439–454

Chapter 15
Comparison of Autoregressive Curves Through Partial Sums of Quasi-Residuals

Fang Li

15.1 Introduction

This chapter is concerned with testing the equality of two autoregressive functions against two sided alternatives when observing two independent strictly stationary and ergodic autoregressive times series of order one. More precisely, let $Y_{1,i}$, $Y_{2,i}$, $i \in \mathbb{Z} := \{0, \pm 1, \cdots\}$, be two observable autoregressive time series such that for some real valued functions μ_1 and μ_2, and for some positive functions σ_1, σ_2,

$$Y_{1,i} = \mu_1(Y_{1,i-1}) + \sigma_1(Y_{1,i-1})\varepsilon_{1,i}, \qquad Y_{2,i} = \mu_2(Y_{2,i-1}) + \sigma_2(Y_{2,i-1})\varepsilon_{2,i}. \quad (15.1)$$

The errors $\{\varepsilon_{1,i},\ i \in \mathbb{Z}\}$ and $\{\varepsilon_{2,i},\ i \in \mathbb{Z}\}$ are assumed to be two independent sequences of independent and identically distributed (i.i.d.) r.v.'s with mean zero and unit variance. Moreover, $\varepsilon_{1,i}$, $i \geq 1$ are independent of $Y_{1,0}$, and $\varepsilon_{2,i}$, $i \geq 1$ are independent of $Y_{2,0}$. And the time series are assumed to be stationary and ergodic.

Consider a bounded interval $[a, b]$ of $\mathbb{R}$. The problem of interest is to test the null hypothesis:

$$H_0 : \mu_1(x) = \mu_2(x), \qquad \forall\, x \in [a, b],$$

against the two sided alternative hypothesis:

$$H_1 : \mu_1(x) \neq \mu_2(x), \quad \text{for some } x \in [a, b], \quad (15.2)$$

based on the data set $Y_{1,0}, Y_{1,1}, \cdots, Y_{1,n_1}, Y_{2,0}, Y_{2,1}, \cdots, Y_{2,n_2}$.

In hydrology, autoregressive time series are often used to model water reservoirs, see, e.g., Bloomfield (1992). The above testing problem could be applied in comparing the water levels of two rivers.

Few related studies had been conducted under the two sample autoregressive setting. Koul and Li (2005) adapts the covariate matching idea used in regression setting to a one-sided tests for the superiority among two time series. Li (2009)

1990 IMS Subject Classification: Primary 62M10, Secondary 62F03.

F. Li (✉)
Department of Mathematical Sciences, Indiana University Purdue University at Indianapolis, IUPUI, 402 N. Blackford Street, 46202 Indianapolis, IN, USA
e-mail: fli@math.iupui.edu

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_15,

studied the same testing problem, but the test is based on the difference of two sums of quasi-residuals. This method is also an extension of T_2 in Koul and Schick (1997) from regression setting to autoregressive setting.

The papers that address the above two sided testing problem in regression setting include Hall and Hart (1990); Delgado (1993); Kulasekera (1995) and Scheike (2000). In particular, Delgado (1993) used the absolute difference of the cumulative regression functions for the same problem, assuming common design in the two regression models. Kulasekera (1995) used quasi-residuals to test the difference between two regression curves, under the conditions that do not require common design points or equal sample sizes. The current article adapts Delgado's idea of using partial sum process and Kulasekera's idea of using quasi residuals to construct the tests for testing the difference between two autoregressive functions.

Similarly, as in Delgado (1993), let

$$\Delta(t) := \int_a^t \big(\mu_1(x) - \mu_2(x)\big)\,(f_1(x) + f_2(x))\,dx, \quad \forall\, a \le t \le b, \tag{15.3}$$

where μ_1, μ_2 are assumed to be continuous on $[a, b]$ and f_1, f_2 are the stationary densities of the two time series $Y_{1,i}$ and $Y_{2,i}$, respectively. We also assume that f_1, f_2 are continuous and positive on $[a, b]$. It is easy to show that $\Delta(t) \equiv 0$ when the null hypothesis holds and $\Delta(t) \neq 0$ for some t under H_a. This suggests to construct tests of H_0 vs. H_a based on some consistent estimators of $\Delta(t)$. One such estimator is obtained as follows.

First, as in Kulasekera (1995), we define quasi-residuals

$$e_{1,i} = Y_{1,i} - \hat{\mu}_2(Y_{1,i-1}), \qquad i = 1, \cdots, n_1, \tag{15.4}$$

and

$$e_{2,j} = Y_{2,j} - \hat{\mu}_1(Y_{2,j-1}), \qquad j = 1, \cdots, n_2. \tag{15.5}$$

Here, $\hat{\mu}_1$ and $\hat{\mu}_2$ are appropriate estimators, such as Nadaraya–Watson estimators used in this article, of μ_1 and μ_2. See Nadaraya (1964) and Watson (1994).

Now, let

$$U_n(t) = \frac{1}{n_1}\sum_{i=1}^{n_1} e_{1,i}\,1_{[a \le Y_{1,i-1} \le t]} - \frac{1}{n_2}\sum_{j=1}^{n_2} e_{2,j}\,1_{[a \le Y_{2,j-1} \le t]}, \tag{15.6}$$

where the subscript n, here and through out the chapter, represents the dependence on n_1 and n_2. With uniformly consistent estimators $\hat{\mu}_1$ and $\hat{\mu}_2$ of μ_1 and μ_2 such as kernel estimates and under some mixing condition on the time series $Y_{1,i}$ and $Y_{2,j}$ such as strongly $\alpha-$mixing, $U_n(t)$ can be shown to be $U_{1n}(t) + U_{2n}(t) + U_{3n}(t)$ with

$$\begin{aligned} U_{1n}(t) = {} & \frac{1}{n_1}\sum_{i=1}^{n_1} \sigma_1(Y_{1,i-1})\varepsilon_{1,i}\,1_{[a \le Y_{1,i-1} \le t]} \\ & - \frac{1}{n_2}\sum_{j=1}^{n_2} \sigma_2(Y_{2,j-1})\varepsilon_{2,j}\,1_{[a \le Y_{2,j-1} \le t]} = o_P(1), \end{aligned}$$

$$
\begin{aligned}
U_{2n}(t) &= \frac{1}{n_1}\sum_{i=1}^{n_1}(\mu_1(Y_{1,i-1})-\mu_2(Y_{1,i-1}))1_{[a\le Y_{1,i-1}\le t]} \\
&\quad - \frac{1}{n_2}\sum_{j=1}^{n_2}(\mu_2(Y_{2,j-1})-\mu_1(Y_{2,j-1}))1_{[a\le Y_{2,j-1}\le t]} \\
&= \int_a^t \big(\mu_1(x)-\mu_2(x)\big)\,(f_1(x)+f_2(x))\,dx + o_P(1), \\
U_{3n}(t) &= \frac{1}{n_1}\sum_{i=1}^{n_1}(\mu_2(Y_{1,i-1})-\hat{\mu}_2(Y_{1,i-1}))1_{[a\le Y_{1,i-1}\le t]} \\
&\quad - \frac{1}{n_2}\sum_{j=1}^{n_2}(\mu_1(Y_{2,j-1})-\hat{\mu}_1(Y_{2,j-1}))1_{[a\le Y_{2,j-1}\le t]} = o_P(1),
\end{aligned}
$$

uniformly for all $t \in [a,b]$. Thus, $U_n(t)$ provides a uniformly consistent estimator of $\Delta(t)$. This suggests to base tests of H_0 on some suitable functions of this process. In this chapter, we shall focus on the Kolmogorov–Smirnov type test based on $\sup_{a\le t\le b}|U_n(t)|$.

To determine the large sample distribution of the process $U_n(t)$, one needs to normalize this process suitably. Let

$$
\begin{aligned}
\tau_n^2(t) &= q_1 E\left\{\sigma_1^2(Y_{1,0})\left(1+\frac{f_2(Y_{1,0})}{f_1(Y_{1,0})}\right)^2 1_{[a\le Y_{1,0}\le t]}\right\} \\
&\quad + q_2 E\left\{\sigma_2^2(Y_{2,0})\left(1+\frac{f_1(Y_{2,0})}{f_2(Y_{2,0})}\right)^2 1_{[a\le Y_{2,0}\le t]}\right\},
\end{aligned}
\tag{15.7}
$$

where, $q_1 = \frac{N}{n_1} = \frac{n_2}{n_1+n_2}$, $q_2 = \frac{N}{n_2} = \frac{n_1}{n_1+n_2}$ and $N = \frac{n_1 n_2}{n_1+n_2}$.

We consider the following normalized test statistics:

$$
T := \sup_{a\le t\le b}\Big|\frac{N^{1/2}U_n(t)}{\sqrt{\tau_n^2(b)}}\Big|. \tag{15.8}
$$

In the case σ_i's and f_i's are known, the tests of H_0 could be based on T, being significant for its large value. But, usually those functions are unknown which renders T of little use. This suggests to replace τ_n with its estimate $\hat{\tau}_n^2$ which satisfies

$$
\frac{\hat{\tau}_n^2(b)}{\tau_n^2(b)} \to_P 1. \tag{15.9}
$$

An example of such estimator $\hat{\tau}_n(t)$ of $\tau_n(t)$ is

$$
\begin{aligned}
\hat{\tau}_n^2(t) &= q_1\frac{1}{n_1}\sum_{i=1}^{n_1}\left\{\big(Y_{1,i}-\tilde{\mu}_1(Y_{1,i-1})\big)^2\left(1+\frac{\hat{f}_2(Y_{1,i-1})}{\hat{f}_1(Y_{1,i-1})}\right)^2 1_{[a\le Y_{1,i-1}\le t]}\right\} \\
&\quad + q_2\frac{1}{n_2}\sum_{j=1}^{n_2}\left\{\big(Y_{2,j}-\tilde{\mu}_2(Y_{2,j-1})\big)^2\left(1+\frac{\hat{f}_1(Y_{2,j-1})}{\hat{f}_2(Y_{2,j-1})}\right)^2 1_{[a\le Y_{2,j-1}\le t]}\right\},
\end{aligned}
\tag{15.10}
$$

where, $\tilde{\mu}_i$'s and $\hat{f}_i$'s are appropriate estimators, such as kernel estimators used in this paper, of μ_i's and f_i's. Therefore, the proposed tests will be based on the adaptive version of T, namely

$$\hat{T} := \sup_{a \le t \le b} \left| \frac{N^{1/2} U_n(t)}{\sqrt{\hat{\tau}_n^2(b)}} \right| \tag{15.11}$$

We shall study the asymptotic behavior of $\hat{T}$ as the sample sizes n_1 and n_2 tend to infinity. Theorem 2.1 of Sect. 15.2 shows that under H_0, T weakly converge to supremum of Brownian motion over $[0, 1]$, under some general assumptions and with $\hat{\mu}_1$ and $\hat{\mu}_2$ being Nadaraya–Watson estimators of μ_1 and μ_2. Then, in Corollary 2.1, under some general assumptions on the estimates $\tilde{\mu}_1$, $\tilde{\mu}_2$ and $\hat{f}_1$, $\hat{f}_2$, we derive the same asymptotic distributions of $\hat{T}$ under H_0. Remark 2.2 proves that the power of the test basted on $\hat{T}$ converges to 1, at the fixed alternative (15.2) or even at the alternatives that converge to H_0 at a rate lower than $\sqrt{\tau_n^2(b)}$. In Sect. 15.3, we conduct a Monte Carlo simulation study of the finite sample level and power behavior of the proposed test $\hat{T}$. The simulation results are shown to be consistent with the asymptotic theory at the moderate sample sizes considered. In Sect. 15.4, we study some properties of kernel smoothers and weak convergence of both empirical processes and marked empirical processes. Those studies facilitate the proof of our main results in Sect. 15.2. But, they may also be of interest on their own, hence are formulated and proved in Sect. 15.4. The other proofs are deferred to Sect. 15.5.

15.2 Asymptotic Behavior of T and $\hat{T}$

This section investigates the asymptotic behavior of T given in (15.8) and the adaptive statistic $\hat{T}$ given in (15.11) under the null hypothesis and the alternatives (15.2). We write P for the underline probability measures and E for the corresponding expectations. In this chapter we consider Nadaraya–Watson estimators $\hat{\mu}_1$, $\hat{\mu}_2$ of μ_1 and μ_2, i.e.,

$$\hat{\mu}_i(x) = \frac{\sum_{j=1}^{n_i} Y_{i,j} K_{h_i}(Y_{i,j-1} - x)}{\sum_{j=1}^{n_i} K_{h_i}(Y_{i,j-1} - x)}, \quad i = 1, 2, \tag{15.12}$$

where $K_{h_i}(x) = \frac{1}{h_i} K(\frac{x}{h_i})$, with K being a kernel density function on the real line with compact support $[-1, 1]$, h_1, $h_2 > 0$ are the bandwidths. First, we recall the following definition from Bosq (1998):

Definition 2.1 *For any real discrete time process $(X_i, i \in \mathbb{Z})$ define the strongly mixing coefficients*

$$\alpha(k) := \sup_{t \in \mathbb{Z}} \alpha(\sigma\text{-field}(X_i, i \le t),\ \sigma\text{-field}(X_i, i \ge t + k)); \quad k = 1, 2, \ldots$$

where, for any two sub σ-fields $\mathcal{B}$ and $\mathcal{C}$,

$$\alpha(\mathcal{B}, \mathcal{C}) = \sup_{B \in \mathcal{B}, C \in \mathcal{C}} |P(B \cap C) - P(B)P(C)|.$$

Definition 2.2. *The process* $(X_i, i \in \mathbb{Z})$ *is said to be geometrically strong mixing (GSM) if there exists* $c_0 > 0$ *and* $\rho \in [0,1)$ *such that* $\alpha(k) \leq c_0 \rho^k$, *for all* $k \geq 1$.

The following assumptions are needed in this paper.

(A.1) The autoregressive functions μ_1, μ_2 are continuous on an open interval containing $[a,b]$ and they have continuous derivatives on $[a,b]$.

(A.2) The kernel function $K(x)$ is a symmetric Lipschitz-continuous density on $\mathbb{R}$ with compact support $[-1,1]$.

(A.3) The bandwidths h_1, h_2 are chosen such that $h_i^2\, N^{1-c} \to \infty$ for some $c > 0$ and $h_i^4\, N \to 0$.

(A.4) The densities f_1 and f_2 are bounded and their restrictions to $[a,b]$ are positive. Moreover, they have continuous second derivatives over an open interval containing $[a,\, b]$.

(A.5) The conditional variance functions σ_1^2 and σ_2^2 are positive on $[a,b]$ and continuous on an open interval containing $[a,b]$.

(A.6) $Y_{1,i}$, $Y_{2,i}$, $i \in \mathbb{Z}$ are GSM processes.

(A.7) For some $M < \infty$, we have

$$E(\varepsilon_{i,1}^4) \leq M, \qquad i = 1,2.$$

(A.8) For $i = 1,2$, the joint densities $g_{i,l}$ between $Y_{i,0}$ and $Y_{i,l}$ for all $l \geq 1$ are uniformly bounded over an open interval $\mathcal{I}_0$ containing $\mathcal{I}$, i.e., $\sup_{l\geq 1} \sup_{x,y\in\mathcal{I}_0} g_{i,l}(x,y) < \infty$.

(A.9) The densities g_1 and g_2 of the innovations $\varepsilon_{1,1}$ and $\varepsilon_{2,1}$ are bounded.

Let $\mathcal{K}(y) = \int_{-1}^{y} K(t)\,dt$ be the distribution function corresponding to the kernel density $K(y)$ on $[-1,\, 1]$ and let

$$\begin{aligned} V_n(t) = {} & \frac{1}{n_1}\sum_{i=1}^{n_1} \varepsilon_{1,i}\, \sigma_1(Y_{1,i-1}) \left(1_{[a\leq Y_{1,i-1}\leq t]} + \frac{f_2(Y_{1,i-1})}{f_1(Y_{1,i-1})} \left(\mathcal{K}\left(\frac{t - Y_{1,i-1}}{h_2} \right) \right.\right. \\ & \left.\left. - \mathcal{K}\left(\frac{a - Y_{1,i-1}}{h_2} \right) \right) \right) \\ & - \frac{1}{n_2}\sum_{j=1}^{n_2} \varepsilon_{2,j}\, \sigma_2(Y_{2,j-1}) \left(1_{[a\leq Y_{2,j-1}\leq t]} + \frac{f_1(Y_{2,j-1})}{f_2(Y_{2,j-1})} \left(\mathcal{K}\left(\frac{t - Y_{2,j-1}}{h_1} \right) \right.\right. \\ & \left.\left. - \mathcal{K}\left(\frac{a - Y_{2,j-1}}{h_1} \right) \right) \right) \end{aligned} \tag{15.13}$$

and

$$\begin{aligned} W_n(t) = {} & \frac{1}{n_1}\sum_{i=1}^{n_1} (\mu_1(Y_{1,i-1}) - \mu_2(Y_{1,i-1})) 1_{[a\leq Y_{1,i-1}\leq t]} \\ & + \frac{1}{n_2}\sum_{j=1}^{n_2} (\mu_1(Y_{2,j-1}) - \mu_2(Y_{2,j-1})) 1_{[a\leq Y_{2,j-1}\leq t]} \end{aligned} \tag{15.14}$$

We are now ready to state the main result.

Theorem 2.1 *Suppose, the conditions* (A.1)–(A.9) *hold true. Then, under both null and alternative hypotheses, as* $n_1 \wedge n_2 \to \infty$,

$$\sup_{a\le t\le b}\left|\frac{N^{1/2}}{\sqrt{\tau_n^2(b)}}\left(U_n(t)-V_n(t)-W_n(t)\right)\right| = o_P(1). \tag{15.15}$$

Here, U_n *is given in (15.6) with* $\hat{\mu}_1$, $\hat{\mu}_2$ *of (15.12), and* V_n *and* W_n *are given in (15.13) and (15.14). Consequently,*

$$\frac{N^{1/2}}{\sqrt{\tau_n^2(b)}}\left(U_n(t)-W_n(t)\right) \Longrightarrow B\circ\varphi(t), \quad \varphi(t)=\lim_{n_1\wedge n_2\to\infty}\frac{\tau_n^2(t)}{\tau_n^2(b)}, \tag{15.16}$$

in the Skorohod space $D[a,b]$, *where* $B\circ\varphi$ *is a continuous Brownian motion on* $[a,b]$ *with respect to time* φ. *Therefore, under* H_0, T *of (15.8) satisfies*

$$T \Longrightarrow \sup_{0\le t\le 1}|B(t)|,$$

where $B(t)$ *is a continuous Brownian motion on* $\mathbb{R}$.

Proof: The proof is given in Sect. 15.5.

Next, we need the following additional assumption to obtain the asymptotic distribution of $\hat{T}$ given in (15.11)

Assumption 2.1 *Let* $\tilde{\mu}_i$, $\hat{f}_i$ *be estimators of* μ_i *and* f_i, *respectively, satisfying*

$$\sup_{a\le x\le b}|\tilde{\mu}_i(x)-\mu_i(x)| = o_P(1), \quad \sup_{a\le x\le b}|\hat{f}_i(x)-f_i(x)| = o_P(1), \qquad i=1,2,$$

under both null and alternative hypotheses.

Corollary 2.1 *Suppose the conditions of Theorem 2.1 hold true. In addition, suppose that there are estimates* $\tilde{\mu}_i$ *and* $\hat{f}_i$ *in* (15.10) *satisfying Assumption 2.1. Then, as* $n_1\wedge n_2\to\infty$ *and under* H_0, $\hat{T}$ of (15.11) *satisfies*

$$\hat{T} \Longrightarrow \sup_{0\le t\le 1}|B(t)|.$$

Proof: It suffices to prove (15.9). Let

$$O_1 = E\left\{\sigma_1^2(Y_{1,0})\left(1+\frac{f_2(Y_{1,0})}{f_1(Y_{1,0})}\right)^2 1_{[a\le Y_{1,0}\le b]}\right\}$$

and

$$O_{1n} = \frac{1}{n_1}\sum_{i=1}^{n_1}\left\{\left(Y_{1,i}-\tilde{\mu}_1(Y_{1,i-1})\right)^2\left(1+\frac{\hat{f}_2(Y_{1,i-1})}{\hat{f}_1(Y_{1,i-1})}\right)^2 1_{[a\le Y_{1,i-1}\le b]}\right\}.$$

By Assumption 2.1, (A.4), (A.5) and Chebyshev's inequality, it can be derived that

$$O_{1n} = O_1 + o_P(1). \tag{15.17}$$

Similarly, we obtain

$$O_{2n} = O_2 + o_P(1), \tag{15.18}$$

with

$$O_2 = E\left\{\sigma_2^2(Y_{2,0})\left(1 + \frac{f_1(Y_{2,0})}{f_2(Y_{2,0})}\right)^2 1_{[a\le Y_{2,0}\le b]}\right\}$$

and

$$O_{2n} = \frac{1}{n_2}\sum_{j=1}^{n_2}\left\{\left(Y_{2,j} - \tilde{\mu}_2(Y_{2,j-1})\right)^2\left(1 + \frac{\hat{f}_1(Y_{2,j-1})}{\hat{f}_2(Y_{2,j-1})}\right)^2 1_{[a\le Y_{2,j-1}\le b]}\right\}.$$

From (15.17), (15.18) and the fact that O_1, O_2 are some positive constants, we have

$$\frac{\hat{\tau}_n^2(b) - \hat{\tau}_n^2(a)}{\tau_n^2(b) - \tau_n^2(a)} = \frac{q_1(O_1 + o_P(1)) + q_2(O_2 + o_P(1))}{q_1 O_1 + q_2 O_2} \to_P 1,$$

which completes the proof of the corollary. □

Remark 2.1 An example of estimates $\tilde{\mu}_i$ and $\hat{f}_i$ satisfying Assumption 2.1 are: $\tilde{\mu}_i = \hat{\mu}_i$ of (15.12) and

$$\hat{f}_i(x) = \frac{1}{n_i}\sum_{j=1}^{n_i} K_{h_i}(Y_{i,j-1} - x), \quad i = 1, 2, \tag{15.19}$$

with h_1, h_2 being appropriate bandwidths that could be different for constructing $\hat{\mu}_i$ in U_n of (15.6). For example, here we can take $h_i = O(n_i^{-1/5})$, See Bosq (1998). But to construct $\hat{\mu}_i$ in U_n of (15.6), we need to choose bandwidths that satisfy (A.3).

Remark 2.2 Testing property of $\hat{T}$: Under the model (15.1), consider the following alternative that is the same as in (15.2):

$$H_a: \quad \mu_1(x) - \mu_2(x) = \delta(x) \neq 0 \quad \text{for some } x \in [a, b],$$

where δ is continuous on $[a, b]$ since μ_1, μ_2 are continuous.

Theorem 2.1 and its corollary suggest to reject the null hypothesis for large values of $\hat{T}$ given in (15.11) under Assumption 2.1.

Let

$$\hat{T}(t) = \frac{N^{1/2}}{\sqrt{\hat{\tau}_n^2(b)}}\, U_n(t), \qquad \hat{T}_1(t) = \frac{N^{1/2}}{\sqrt{\hat{\tau}_n^2(b)}}(U_n(t) - W_n(t))$$

Then,

$$\hat{T}(t) = \hat{T}_1(t) + h(t), \quad h(t) = \frac{N^{1/2}}{\sqrt{\hat{\tau}_n^2(b)}} W_n(t)$$

By Ergodic theorem,

$$W_n \to_P \int_a^t \delta(x)\, dx,$$

$$h(t) \sim \frac{N^{1/2}}{\sqrt{\hat{\tau}_n^2(b)}} \int_a^t \delta(x)\, dx. \tag{15.20}$$

This, together with $\frac{N^{1/2}}{\sqrt{\tau_n^2(b)}} \to \infty$, (15.9), and the fact that $\int_a^t \delta(x)\, dx$ is not 0 for some $a \le t \le b$ implies,

$$\sup_{a \le t \le b} |h(t)| \to_P \infty. \tag{15.21}$$

Hence, in view of (15.16) and (15.9),

$$\hat{T} = \sup_{a \le t \le b} |\hat{T}(t)| = \sup_{a \le t \le b} |\hat{T}_1(t) + h(t)| \to_P \infty. \tag{15.22}$$

This, together with Corollary 2.1 indicates that the test bases on $\hat{T}$ is consistent for H_a.

Note: By using the same arguments as above, we even can claim that under Assumption 2.1, the test based on $\hat{T}$ is consistent for the alternatives converging to the null hypothesis at any rate α_n that is lower than $N^{-1/2}$, since (15.21) is still satisfied when $\delta(x)$ is replaced by $\delta(x)\alpha_n$. Furthermore, under $H_1 : \mu_1(x) - \mu_2(x) = \frac{\sqrt{\tau_n^2(b)}}{N^{1/2}}\delta(x), \quad x \in [0, 1]$, the limiting powers of the asymptotic level α tests T is computed as

$$\lim_{n \to \infty} P(\hat{T}_1 > b_\alpha) \quad = \quad P\left(\sup_{a \le t \le b} |B \circ \varphi(t) + g(t)| > b_\alpha \right),$$

where b_α is defined such that

$$P\left(\sup_{0 \le t \le 1} |B(t)| > b_\alpha \right) = \alpha.$$

15.3 Simulation

In this section, we investigate the finite sample behavior of the nominal level of the proposed test $\hat{T}$ under H_0 and power of $\hat{T}$ against some nonparametric alternatives. As sample sizes, we choose the moderate sample sizes $n_1 = n_2 = n = 50,\ 100,\ 300,$

Table 15.1 The critical values b_α

α	0.05	0.025	0.01
b_α	2.24241	2.49771	2.80705

600, 1,000, and 2,000 with each simulation being repeated for 1,000 times. The data is simulated from model (15.1), where the two autoregressive functions are chosen to be $\mu_2(x) = 1 + x/2$ and $\mu_1(x) = \mu_2(x) + \delta(x)$, and the innovations $\{\varepsilon_{1,i}\}$ and $\{\varepsilon_{2,i}\}$ are taken to be independent standard normal $\mathcal{N}(0, 1)$. We choose $\delta(x) = 0$ corresponding to H_0 and $\delta(x) = 1,\ 2(x-3)/x$, and $2\left(\frac{n_1 n_2}{n_1+n_2}\right)^{-1/2} = 2\ N^{-1/2}$ corresponding to H_a. Note that the second choice of δ is negative for $x < 3$ and positive for $x > 3$ and converge to 0 for $x \to 3$; the last choice of δ corresponds to the local alternatives with a rate being the same as τ_n of (15.7). For simplicity, the conditional variance functions σ_1 and σ_2 are chosen to be (i) $\sigma_1(x) = \sigma_2(x) = 1$ and (ii) $\sigma_1(x) = \sigma_2(x) = 3/\sqrt{1+x^2}$. Finally, the interval $[a, b]$ in (15.2) is taken to be $[2, 4]$.

To construct the test statistics $\hat{T}$ of (15.11), we consider Nadaraya–Watson estimators $\hat{\mu}_1$, $\hat{\mu}_2$ of (15.12) with kernel $K(x) = \frac{3}{4}(1-x^2)\mathbf{1}(|x| \le 1)$ and we considered three different bandwidths $h_1 = h_2 = 0.15, 0.2$ and 0.25. The estimates $\tilde{\mu}_1$, $\tilde{\mu}_2$ and $\hat{f}_1$, $\hat{f}_2$ in $\hat{\tau}_n$ of (15.10) are from Remark 2.1 with $h_i = n_i^{-1/5}$, $i = 1, 2$. Let b_α satisfy $P(\sup_{0\le t\le 1} |B(t)| > b_\alpha) = \alpha$. Then, the empirical size (power) is computed by the proportion of rejects $\frac{\#\ \text{of}\ [\hat{T} > b_\alpha]}{1000}$.

In Table 15.1, we give the critical values b_α obtained from the formula $P\left(\sup_{0\le t\le 1} |B(t)| < b\right) = P(\,|B(1)| < b\,) + 2\sum_{i=1}^{\infty}(-1)^i P(\,(2i-1)b < B(1) < (2i+1)b\,)$ given on page 553 of the book by Resnick (1992).

The simulation programming was done using R. To generate each of the two samples, we first generated $(500 + n)$ error variables from $\mathcal{N}(0, 1)$. Using these errors and model (15.1) with the initial value $Y_{i,0}$ randomly chosen from $\mathcal{N}(0, 1)$, we generated $(501 + n)$ observations. The last $(n + 1)$ observations from the data thus generated are used in carrying out the simulation study.

The results of the simulation study are shown in Table 15.2 below. Three rows correspond each choice of $\delta(x)$ with the first row corresponding to bandwidth 0.15, the second to 0.2 and the third to 0.25. The finite sample level and power behavior of the tests are shown to be quite stable across the various choices of the bandwidth. One sees that for both choices of σ_1 and σ_2, the empirical sizes of the test are not much different from the nominal levels for most moderate samples sizes, but they are closer to the true levels when the sample size gets larger. The simulated powers under fixed alternative $\delta(x) = 1$ are close to 1 for all moderate sample sizes, even at α-level .025. The simulated powers under fixed alternative $\delta(x) = 2(x-3)/x$ are seen to increase quickly with n and they are quite large for $n \ge 600$. The simulated powers under local alternative $\delta(x) = 2\ N^{-1/2}$ are stable for most moderate sample sizes. In summary, the simulated levels and powers are consistent with the asymptotic theory at most moderate sample sizes considered.

Table 15.2 Proportion of rejections $\{\hat{T} > 2.24241\,(2.49771)\}$ at level $\alpha = .05\,(.025)$ for (i) $\sigma_1 = 1 = \sigma_2$ and (ii) $\sigma_1(x) = 3/\sqrt{1+x^2} = \sigma_2(x)$. Three rows correspond each choice of $\delta(x)$ with the first row corresponding to bandwidth 0.15, the second to 0.2 and the third to 0.25

		(i)						(ii)					
α-level	$\delta(x) \setminus n$	50	100	300	600	1,000	2,000	50	100	300	600	1,000	2,000
0.05	0	0.085	0.056	0.048	0.048	0.036	0.034	0.070	0.062	0.041	0.038	0.044	0.043
		0.090	0.057	0.039	0.040	0.040	0.044	0.084	0.049	0.040	0.042	0.037	0.032
		0.085	0.050	0.042	0.041	0.035	0.045	0.068	0.057	0.042	0.030	0.049	0.044
	1	0.873	0.989	1	1	1	1	0.910	0.993	1	1	1	1
		0.897	0.996	1	1	1	1	0.918	0.995	1	1	1	1
		0.916	0.996	1	1	1	1	0.930	0.995	1	1	1	1
	$2(x-3)/x$	0.191	0.222	0.474	0.827	0.968	1	0.147	0.194	0.349	0.657	0.888	0.998
		0.219	0.247	0.483	0.827	0.970	1	0.165	0.187	0.356	0.655	0.886	0.998
		0.212	0.221	0.449	0.798	0.973	1	0.146	0.181	0.337	0.671	0.878	0.998
	$2\,N^{-1/2}$	0.335	0.317	0.266	0.228	0.243	0.262	0.335	0.295	0.349	0.244	0.221	0.207
		0.333	0.313	0.245	0.258	0.232	0.195	0.320	0.294	0.227	0.209	0.192	0.222
		0.342	0.278	0.256	0.256	0.243	0.232	0.331	0.256	0.238	0.234	0.218	0.214
0.025	0	0.048	0.031	0.028	0.023	0.020	0.020	0.043	0.040	0.025	0.016	0.024	0.025
		0.061	0.029	0.022	0.015	0.020	0.028	0.039	0.036	0.023	0.020	0.024	0.015
		0.055	0.034	0.022	0.020	0.018	0.021	0.041	0.031	0.021	0.017	0.025	0.019
	1	0.824	0.982	1	1	1	1	0.870	0.985	1	1	1	1
		0.841	0.991	1	1	1	1	0.885	0.992	1	1	1	1
		0.874	0.990	1	1	1	1	0.892	0.991	1	1	1	1
	$2(x-3)/x$	0.141	0.154	0.357	0.730	0.937	1	0.103	0.126	0.248	0.533	0.803	0.991
		0.160	0.163	0.356	0.722	0.924	1	0.119	0.134	0.254	0.522	0.807	0.994
		0.148	0.158	0.327	0.704	0.933	1	0.104	0.121	0.224	0.531	0.801	0.994
	$2\,N^{-1/2}$	0.259	0.227	0.193	0.167	0.161	0.177	0.264	0.214	0.167	0.173	0.155	0.150
		0.255	0.225	0.160	0.185	0.171	0.133	0.254	0.226	0.166	0.155	0.132	0.150
		0.248	0.188	0.173	0.170	0.164	0.172	0.246	0.176	0.171	0.155	0.146	0.149

15.4 Properties of Kernel Smoothers and Weak Convergence of Empirical Processes

In this section, we first study the asymptotic behavior of the following kernel smoothers over $[a, b]$ for $i = 1, 2$:

$$\hat{f}_i(x) = \frac{1}{n_i}\sum_{j=1}^{n_i} K_{h_i}(Y_{i,j-1} - x), \quad \Lambda_{i,n}(x) = \frac{1}{n_i}\sum_{j=1}^{n_i} \sigma_i(Y_{i,j-1})\varepsilon_{i,j} K_{h_i}(Y_{i,j-1} - x), \tag{15.23}$$

$$\Psi_i(x, y) = \frac{1}{n_i}\sum_{j=1}^{n_i} \frac{K_{h_i}(Y_{i,j-1} - x)}{\Psi_i(Y_{i,j-1})} 1_{[Y_{i,j-1} \le y]}, \quad \Psi_1 = f_2, \quad \Psi_2 = f_1, \tag{15.24}$$

$$\Gamma_1(y) = \frac{1}{n_1}\sum_{i=1}^{n_1} \varepsilon_{1,i}\Psi_2(Y_{1,i-1}, y), \ \Gamma_2(y) = \frac{1}{n_2}\sum_{i=1}^{n_2} \varepsilon_{2,i}\Psi_1(Y_{2,i-1}, y), \tag{15.25}$$

$$H_{i,j}(y) = \frac{1}{n_i}\sum_{k=1}^{n_i} K_{h_i}(Y_{i,k-1} - y)(Y_{i,k-1} - y)^j, \quad i, j = 1, 2. \tag{15.26}$$

By Lemma 1–4 in Li (2008), we have the following results.

Lemma 4.1 *Suppose conditions (A.2), (A.3) and (A.4)–(A.8) hold. Then,*

$$\sup_{a \le x \le b} \left|\Lambda_{i,n}(x)\right| = O_P\left(\sqrt{\frac{\log n_i}{n_i h_i}}\right), \quad i = 1, 2,$$

where $\Lambda_{i,n}(x)$ *is given in* (15.23).

Lemma 4.2 *Suppose conditions (A.2), (A.3), (A.4), (A.6) and (A.8) hold. Then* f_i of (15.23) *satisfies*

$$\sup_{a \le x \le b} |\hat{f}_i(x) - f_i(x)| = O_P\left(\sqrt{\frac{\log n_i}{n_i h_i}}\right) + O(h_i^2), \quad i = 1, 2.$$

Lemma 4.3 *Suppose condition (A.2), (A.3), (A.4), (A.6) and (A.8) hold. Then,* $\Psi_i(x, y)$ *of* (15.24) *satisfies*

$$\sup_{\substack{a - h_i \le x \le b + h_i \\ a \le y \le b}} Var\{\Psi_i(x, y)\} = O(\frac{1}{n_i h_i}), \qquad i = 1, 2.$$

Lemma 4.4 *Suppose condition (A.2)- (A.4), (A.6) and (A.8) hold. Then* $H_{i,j}$ *of* (15.26) *satisfies*

$$H_{i,j}(y) = h_i^j f_i(y) u_j + O_P\left(h_i^{j+1} + h_j^j \sqrt{\frac{\log n_i}{n_i h_i}}\right),$$

$$u_j = \int_{-1}^{1} K(u) u^j \, du, \qquad i, j = 1, 2$$

uniformly on $a \le x \le b$.

Next, we study the property of some empirical processes. The weak convergence of marked empirical process proved in Theorem 2.2.6 of Koul (2002) implied the following lemma:

Lemma 4.5 *Suppose conditions (A.4), (A.6), (A.7) and (A.9) hold, then for* $i = 1, 2$,

$$\sup_{a \le y \le b} \left| \frac{1}{n_i} \sum_{j=1}^{n_i} (|\varepsilon_{i,j}| - E|\varepsilon_{i,j}|) 1_{[Y_{i,j-1} \le y]} \right| = O_P(n_i^{-1/2}).$$

Next, recall that $\mathcal{K}(y) = \int_{-1}^{y} K(t)\, dt$ is the distribution function corresponding to the kernel density $K(y)$ on $[-1, 1]$. We prove the following lemma:

Lemma 4.6 *Suppose conditions (A.2), (A.3), (A.4), (A.6) and (A.8) hold. Then* Γ_i of (15.25) *satisfies*

$$\sup_{a \le y \le b} \left| N^{1/2} \left(\Gamma_1(y) - \frac{1}{n_1} \sum_{i=1}^{n_1} \varepsilon_{1,i} \frac{f_2(Y_{1,i-1})}{f_1(Y_{1,i-1})} \mathcal{K}\left(\frac{y - Y_{1,i-1}}{h_2} \right) \right) \right| = o_P(1),$$

$$\sup_{a \le y \le b} \left| N^{1/2} \left(\Gamma_2(y) - \frac{1}{n_2} \sum_{i=1}^{n_2} \varepsilon_{2,i} \frac{f_1(Y_{2,i-1})}{f_2(Y_{2,i-1})} \mathcal{K}\left(\frac{y - Y_{2,i-1}}{h_1} \right) \right) \right| = o_P(1).$$

Proof: The proof is similar to the proof of Lemma 1 of Li (2008) which is in turn similar to Lemma 6.1 of Fan and Yao (2003). It is sufficient to prove the first equality. Let C denote a generic constant, which can vary from one place to another. Also let

$$N^{1/2} \left(\Gamma_1(y) - \frac{1}{n_1} \sum_{i=1}^{n_1} \varepsilon_{1,i} \frac{f_2(Y_{1,i-1})}{f_1(Y_{1,i-1})} \mathcal{K}\left(\frac{y - Y_{1,i-1}}{h_2} \right) \right) = A_n(y)$$

Now, decompose $A_n(y)$ into $A_{1,n}(y) + A_{2,n}(y)$ with

$$A_{1,n}(y) = N^{1/2} \frac{1}{n_1} \sum_{i=1}^{n_1} \varepsilon_{1,i} \left(\Psi_2(Y_{1,i-1}, y) - E(\Psi_2(Y_{1,i-1}, y)) \right),$$

$$A_{2,n}(y) = N^{1/2} \frac{1}{n_1} \sum_{i=1}^{n_1} \varepsilon_{1,i} \left(E(\Psi_2(Y_{1,i-1}, y)) - \frac{f_2(Y_{1,i-1})}{f_1(Y_{1,i-1})} \mathcal{K}\left(\frac{y - Y_{1,i-1}}{h_2} \right) \right).$$

First, we show $\sup_{a\le y\le b}|A_{2,n}(y)| = o_P(1)$. For some $(Y^*_{1,i-1} \in [a-2h_2,\ b+2h_2]$, By Tayler expansion, We have

$$
\begin{aligned}
&A_{2,n}(y)\\
&= N^{1/2}\frac{1}{n_1}\sum_{i=1}^{n_1}\varepsilon_{1,i}\left(\int_{-1}^{\frac{y-Y_{1,i-1}}{h_2}} K(u)\left(\frac{f_2(Y_{1,i-1}+h_2u)}{f_1(Y_{1,i-1}+h_2u)} - \frac{f_2(Y_{1,i-1})}{f_1(Y_{1,i-1})}\right)du\right)\\
&= N^{1/2}\frac{1}{n_1}\sum_{i=1}^{n_1}\varepsilon_{1,i}\left(h_2\int_{-1}^{\frac{y-Y_{1,i-1}}{h_2}} uK(u)\frac{f_2'(Y_{1,i-1})f_1(Y_{1,i-1}) - f_1'(Y_{1,i-1})f_2(Y_{1,i-1})}{f_1(Y_{1,i-1}+h_2u)f_1(Y_{1,i-1})}\,du\right.\\
&\quad\left. + \frac{h_2^2}{2}\int_{-1}^{\frac{y-Y_{1,i-1}}{h_2}} u^2\ K(u)\frac{f_2''(Y^*_{1,i-1})f_1(Y_{1,i-1}) - f_1''(Y^*_{1,i-1})f_2(Y_{1,i-1})}{f_1(Y_{1,i-1}+h_2u)f_1(Y_{1,i-1})}\,du\right)\\
&\le N^{1/2}h_2\frac{1}{n_1}\sum_{i=1}^{n_1}\varepsilon_{1,i}\int_{-1}^{\frac{y-Y_{1,i-1}}{h_2}} uK(u)\frac{f_2'(Y_{1,i-1})f_1(Y_{1,i-1}) - f_1'(Y_{1,i-1})f_2(Y_{1,i-1})}{f_1(Y_{1,i-1}+h_2u)f_1(Y_{1,i-1})}\,du\\
&\quad + N^{1/2}h_2^2\frac{1}{n_1}\sum_{i=1}^{n_1}|\varepsilon_{1,i}|\cdot C, \qquad \text{by (A.2) and (A.4)},
\end{aligned}
$$

uniformly over $[a,\ b]$,

By a similar argument in proving Lemma 4.1 or Lemma 1 of Li (2008), it can be shown that

$$
\begin{aligned}
&\sup_{a\le y\le b}\left|\frac{1}{n_1}\sum_{i=1}^{n_1}\varepsilon_{1,i}\int_{-1}^{\frac{y-Y_{1,i-1}}{h_2}} uK(u)\frac{f_2'(Y_{1,i-1})f_1(Y_{1,i-1}) - f_1'(Y_{1,i-1})f_2(Y_{1,i-1})}{f_1(Y_{1,i-1}+h_2u)f_1(Y_{1,i-1})}\,du\right|\\
&\qquad = O_P\left(\sqrt{\frac{\log n_1}{n_1h_2}}\right).
\end{aligned}
$$

Also, $N^{1/2}h_2^2\frac{1}{n_1}\sum_{i=1}^{n_1}|\varepsilon_{1,i}| = O_P(N^{1/2}h_2^2)) = o_P(1)$ by (A.3). Hence, by (A.3) we have

$$
\begin{aligned}
\sup_{a\le y\le b}|A_{2,n}(y)| &= O_P\left(\sqrt{q_1h_2\log n_1}\right) + o_P(1)\\
&= O_P\left(\sqrt{q_1h_2\log\frac{N}{q_1}}\right) + o_P(1) = o_P(1). \qquad (15.27)
\end{aligned}
$$

Now, it is left to prove

$$
\sup_{a\le y\le b}|A_{1,n}(y)| = o_P(1). \qquad (15.28)
$$

Slightly simpler than the proof of Lemma 1 in Li (2008) and Lemma 6.1 in Fan and Yao (2003), the proof consists of the following two steps:

(a) (Discretization). Partition the interval $[a, b]$ with length L into $M = [(N^{1+c})^{1/2}]$ subintervals $\{I_k\}$ of equal length. Let $\{y_k\}$ be the centers of I_k. Then

$$\sup_{a\le y\le b} |A_{1,n}(y)| \le \max_{1\le k\le M} |A_{1,n}(y_k)| + o_P(1). \tag{15.29}$$

(b) (Maximum deviation for discretized series). For any small ϵ,

$$P\left(\max_{1\le j\le M} |A_{1,n}(y_j)| > \epsilon\right) \to 0. \tag{15.30}$$

Let $G_{i,n}(y) = \sqrt{n_i}\left(\frac{1}{n_i}\sum_{j=1}^{n_i} 1_{[Y_{i,j-1}\le y]} - P(Y_{i,j-1} \le y)\right)$. The strong approximation theorem for the empirical process of a stationary sequence of strong mixing random variables established in Berkes and Philipp (1997) and in Theorem 4.3 of the monograph edited by Dehling et al (2002) implied

$$\sup_{1\le k\le M} \sup_{y\in I_k} \left|G_{1,n}(y) - G_{1,n}(y_k)\right| = o_P(1). \tag{15.31}$$

Now, we prove part (a). First, for any $1 \le K \le M$ and all $y \in I_k$, we decompose $A_{1,n}(y) - A_{1,n}(y_k)$ as $D_{1,n}(y) + D_{2,n}(y)$ with

$$D_{1,n}(y) = N^{1/2}\frac{1}{n_1}\sum_{i=1}^{n_1} \varepsilon_{1,i}\left(\Psi_2(Y_{1,i-1}, y) - \Psi_2(Y_{1,i-1}, y_k))\right),$$

$$D_{2,n}(y) = N^{1/2}\frac{1}{n_1}\sum_{i=1}^{n_1} \varepsilon_{1,i}\left(E(\Psi_2(Y_{1,i-1}, y)) - E(\Psi_2(Y_{1,i-1}, y_k))\right),$$

Without losing generality, it is sufficient to consider all $y_k \le y \in I_k$. It is easy to see that

$$\begin{aligned}
&|D_{1,n}(y)| \\
&\le CN^{1/2}\left(\frac{1}{n_1}\sum_{i=1}^{n_1} |\varepsilon_{1,i}| 1_{[y_k - h_2 \le Y_{1,i-1} \le y + h_2]}\right)\left(\frac{1}{n_2 h_2}\sum_{i=1}^{n_2} 1_{[y_k \le Y_{2,i-1} \le y]}\right) \\
&\le CN^{1/2}\left(\frac{1}{n_1}\sum_{i=1}^{n_1} (|\varepsilon_{1,i}| - E|\varepsilon_{1,i}|) 1_{[y_k - h_2 \le Y_{1,i-1} \le y + h_2]} + \frac{1}{n_1}\sum_{i=1}^{n_1} 1_{[y_k - h_2 \le Y_{1,i-1} \le y + h_2]}\right) \\
&\qquad \left(\frac{1}{h_2}\left[\frac{1}{\sqrt{n_2}}(G_{2,n}(y) - G_{2,n}(y_k)) + P(y_k \le Y_{2,i-1} \le y)\right]\right) \\
&= C\frac{N^{1/2}}{h_2}\Big(O_P\left(\frac{1}{\sqrt{n_1}}\right) + \frac{1}{\sqrt{n_1}}(G_{1,n}(y + h_2) - G_{1,n}(y_k - h_2)) \\
&\quad + P(y_k - h_2 \le y_{1,i-1} \le y + h_2)\Big)\left(o_P(\frac{1}{\sqrt{n_2}}) + O_P(\frac{1}{M})\right) \text{ by Lemma 4.5, (15.31)} \\
&= C\frac{N^{1/2}}{h_2}\left(O_P\left(\frac{1}{\sqrt{n_1}}\right) + O_P(h_2)\right)\left(o_P(\frac{1}{\sqrt{n_2}}) + O_P(\frac{1}{M})\right) \text{ again by (15.31)} \\
&= o_P(1),
\end{aligned}$$

and similarly,

$$|D_{2,n}(y)| \le CN^{1/2}\left(\frac{1}{n_1}\sum_{i=1}^{n_1}|\varepsilon_{1,i}|1_{[y_k-h_2\le Y_{1,i-1}\le y+h_2]}\right)\left(\frac{1}{h_2}P(y_k \le Y_{2,i-1} \le y)\right)$$
$$= C\frac{N^{1/2}}{h_2}O_P(h_2)O_P\left(\frac{1}{N^{1/2+c/2}}\right) = o_P(1).$$

Hence, we have

$$\sup_{1\le k\le M}\sup_{y\in I_j}|A_{1,n}(y) - A_{1,n}(y_k)| = o_P(1).$$

This proves part (a). Next,

$$P(\max_{1\le k\le M}|A_{1,n}(y_k)| > \epsilon) \le M\max_{1\le k\le M}E\left(A_{1,n}^2(y_k)\right)/\epsilon^2$$
$$= N^{1/2+c/2}N\frac{1}{n_1}O\left(\frac{1}{n_2h_2}\right) \to 0, \qquad \text{by Lemma 4.3 and (A.3).}$$

This proves part (b) and hence finishes the proof of (15.28) and the Lemma. □

15.5 Proofs

Here, we shall give the proof of our main result, Theorem 2.1. The lemmas proved in Sect. 15.4 will facilitate the proof of this theorem. As usual, let C be a generic constant. It suffices to prove (15.15) and (15.16). Now consider $N^{1/2}U_n(t)$ for all $a \le t \le b$. We decompose $N^{1/2}U_n(t)$ as $B_{1,n}(t) - B_{2,n}(t)$ with

$$B_{1,n}(t) = N^{1/2}\frac{1}{n_1}\sum_{i=1}^{n_1}(Y_{1,i} - \hat{\mu}_2(Y_{1,i-1}))1_{[a\le Y_{1,i-1}\le t]},$$

$$B_{2,n}(t) = N^{1/2}\frac{1}{n_2}\sum_{i=1}^{n_2}(Y_{2,i} - \hat{\mu}_1(Y_{2,i-1}))1_{[a\le Y_{2,i-1}\le t]}.$$

We first consider $B_{1,n}(t)$. Recall definitions (15.12) and (15.23)–(15.25). By decomposition and simple algebra, we rewrite $B_{1,n}(t)$ as $I(t) - II(t) + III(t)$ with

$$I(t) = \frac{N^{1/2}}{n_1}\sum_{i=1}^{n_1}(\varepsilon_{1,i}\sigma_1(Y_{1,i-1}) + (\mu_1(Y_{1,i-1}) - \mu_2(Y_{1,i-1})))\mathbf{1}_{[a\le Y_{1,i-1}\le t]} \tag{15.32}$$

$$II(t) = \frac{N^{1/2}}{n_1}\sum_{i=1}^{n_1}\frac{\sum_{j=1}^{n_2}\varepsilon_{2,j}\sigma_2(Y_{2,j-1})K_{h_2}(Y_{2,j-1} - Y_{1,i-1})}{n_2 f_2(Y_{1,n-1})}\mathbf{1}_{[a\le Y_{1,i-1}\le t]} \tag{15.33}$$

$$III(t) = \frac{N^{1/2}}{n_1}\sum_{i=1}^{n_1}\frac{\sum_{j=1}^{n_2}(\mu_2(Y_{1,i-1}) - \mu_2(Y_{2,j-1}))K_{h_2}(Y_{2,j-1} - Y_{1,i-1})}{n_2\hat{f}_2(Y_{1,i-1})}\mathbf{1}_{[a\le Y_{1,i-1}\le t]} \tag{15.34}$$

Now we consider $II(t)$. By decomposition, we rewrite it as $II_1(t) + II_2(t)$ with

$$II_1(t) = N^{1/2}\frac{1}{n_2}\sum_{j=1}^{n_2}\varepsilon_{2,j}\sigma_2(Y_{2,j-1})(\Psi_1(Y_{2,j-1},t) - \Psi_1(Y_{2,j-1},a))$$

$$= N^{1/2}\frac{1}{n_2}\sum_{j=1}^{n_2}\varepsilon_{2,j}\sigma_2(Y_{2,j-1})\frac{f_1(Y_{2,j-1})}{f_2(Y_{2,j-1})}\left(\mathcal{K}\left(\frac{t-Y_{2,j-1}}{h_1}\right)\right.$$

$$\left. - \mathcal{K}\left(\frac{a-Y_{2,j-1}}{h_1}\right)\right) + o_P(1), \tag{15.35}$$

uniformly on $a \le t \le b$ by Lemma 4.6 and its proof, and uniformly on $a \le t \le b$,

$$II_2(t) = \frac{N^{1/2}}{n_1}\sum_{i=1}^{n_1}\frac{\sum_{j=1}^{n_2}\varepsilon_{2,j}\sigma_2(Y_{2,j-1})K_{h_2}(Y_{2,j-1}-Y_{1,i-1})}{n_2 f_2(Y_{1,i-1})}\mathbf{1}_{[a\le Y_{1,i-1}\le t]}$$

$$\frac{f_2(Y_{1,i-1}) - \hat{f}_2(Y_{1,i-1})}{\hat{f}_2(Y_{1,i-1})}$$

$$\le C\frac{N^{1/2}}{n_1}\sum_{i=1}^{n_1}\left|\frac{1}{n_2}\sum_{j=1}^{n_2}\varepsilon_{2,j}\sigma_2(Y_{2,j-1})K_{h_2}(Y_{2,j-1}-Y_{1,i-1})\right|\mathbf{1}_{[a\le Y_{1,i-1}\le t]}$$

$$\cdot \sup_{a\le Y_{1,i-1}\le t}\left|\frac{f_2(Y_{1,i-1}) - \hat{f}_2(Y_{1,i-1})}{\hat{f}_2(Y_{1,i-1})}\right|$$

$$= CN^{1/2}O_P\left(\sqrt{\frac{\log n_2}{n_2 h_2}}\right)\cdot\left(O_P\left(\sqrt{\frac{\log n_2}{n_2 h_2}}\right) + O_P(h_2^2)\right), \text{by Lemma 4.1 and 4.2}$$

$$= o_P(1), \qquad \text{by (A.3).}$$

Next, we consider $III(t)$. Let $\mu_2^{(1)}$ denote the first derivative of μ_2. Then, by (A.1), uniformly on $a \le t \le b$,

$$III(t) \le \frac{N^{1/2}}{n_1}\sum_{i=1}^{n_1}\frac{\mu_2^{(1)}(Y_{1,i-1})|H_{2,1}(Y_{1,i-1})| + C|H_{2,2}(Y_{1,i-1})|}{\hat{f}_2(Y_{1,i-1})}\mathbf{1}_{[a\le Y_{1,i-1}\le t]}$$

$$= O_P\left(N^{1/2}\left(h_2^2 + h_2\sqrt{\frac{\log n_2}{n_2 h_2}}\right)\right) = 0_P(1), \text{by Lemma 4 and (A.3)} \tag{15.36}$$

Hence, by (15.32) and (15.35)–(15.36), we have uniformly on $a \le t \le b$,

$$B_{1,n}(t) = \frac{N^{1/2}}{n_1} \sum_{i=1}^{n_1} (\varepsilon_{1,i}\sigma_1(Y_{1,i-1}) + (\mu_1(Y_{1,i-1}) - \mu_2(Y_{1,i-1})))\mathbf{1}_{[a \le Y_{1,i-1} \le t]}$$
$$- \frac{N^{1/2}}{n_2} \sum_{j=1}^{n_2} \varepsilon_{2,j}\sigma_2(Y_{2,j-1}) \frac{f_1(Y_{2,j-1})}{f_2(Y_{2,j-1})} \left(\mathcal{K}\left(\frac{t - Y_{2,j-1}}{h_1} \right) \right.$$
$$\left. - \mathcal{K}\left(\frac{a - Y_{2,j-1}}{h_1} \right) \right) + o_P(1) \tag{15.37}$$

Similarly, we have uniformly on $a \le t \le b$,

$$B_{2,n}(t) = \frac{N^{1/2}}{n_2} \sum_{j=1}^{n_2} (\varepsilon_{2,j}\sigma_2(Y_{2,j-1}) + (\mu_2(Y_{2,j-1}) - \mu_1(Y_{2,j-1})))\mathbf{1}_{[a \le Y_{2,j-1} \le t]}$$
$$- \frac{N^{1/2}}{n_1} \sum_{i=1}^{n_1} \varepsilon_{1,i}\sigma_1(Y_{1,i-1}) \frac{f_2(Y_{1,i-1})}{f_1(Y_{1,i-1})} \left(\mathcal{K}\left(\frac{t - Y_{1,i-1}}{h_2} \right) \right.$$
$$\left. - \mathcal{K}\left(\frac{a - Y_{1,i-1}}{h_2} \right) \right) + o_P(1) \tag{15.38}$$

By (15.37) and (15.38), we proved (15.15).

Now, we need to prove (15.16). Applying the CLT for martingales [Hall and Heyde (1980), Corollary 3.1], we first could show that the finite-dimensional distributions of $\frac{N^{1/2}}{\tau_n^2(b)} V_n(t)$ tend to the right limit. Then, apply theorem for weak convergence on functional space [Hall and Heyde (1980), Theorem A.2], we need to prove the tightness of $\frac{N^{1/2}}{\tau_n^2(b)} V_n(t)$. It suffices to prove the tightness of

$$\frac{N^{1/2}}{n_1} \sum_{i=1}^{n_1} \varepsilon_{1,i}\sigma_1(Y_{1,i-1})\mathbf{1}_{[a \le Y_{1,i-1} \le t]}$$

and

$$\frac{N^{1/2}}{n_1} \sum_{i=1}^{n_1} \varepsilon_{1,i}\sigma_1(Y_{1,i-1}) \frac{f_2(Y_{1,i-1})}{f_1(Y_{1,i-1})} \mathcal{K}\left(\frac{t - Y_{1,i-1}}{h_2} \right).$$

The tightness of the first sequence is implied by the weak convergence of a marked empirical process [Koul and Stute (1999), Lemma3.1].

Since $\mathcal{K}\left(\frac{t-Y_{1,i-1}}{h_2}\right) = 1$ for $Y_{1,i-1} \le t - h_2$ and $\mathcal{K}\left(\frac{t-Y_{1,i-1}}{h_2}\right) 1_{[t-h_2 \le Y_{1,i-1} \le t+h_2]}$ just behaves like $h_2 K_{h_2}(t - Y_{1,i-1})$, the second sequence can be rewritten as

$$\frac{N^{1/2}}{n_1} \sum_{i=1}^{n_1} \varepsilon_{1,i}\sigma_1(Y_{1,i-1}) \frac{f_2(Y_{1,i-1})}{f_1(Y_{1,i-1})} 1_{[Y_{1,i-1} \le t-h_2]}$$
$$+ \frac{N^{1/2}}{n_1} \sum_{i=1}^{n_1} \varepsilon_{1,i}\sigma_1(Y_{1,i-1}) \frac{f_2(Y_{1,i-1})}{f_1(Y_{1,i-1})} \mathcal{K}\left(\frac{t - Y_{1,i-1}}{h_2} \right) 1_{[t-h_2 \le Y_{1,i-1} \le t+h_2]},$$

with second term bing $o_P(1)$ uniformly on $a \leq t \leq b$ by a proof similar to that of Lemma 4.1. Again, by the weak convergence of a marked empirical process [Koul and Stute (1999), Lemma3.1], we could prove the tightness of $\frac{N^{1/2}}{n_1}\sum_{i=1}^{n_1}\varepsilon_{1,i}\sigma_1(Y_{1,i-1})\frac{f_2(Y_{1,i-1})}{f_1(Y_{1,i-1})}1_{[Y_{1,i-1}\leq t-h_2]}$. Therefore, we complete the proof of the main theory. □

Acknowledgements The author is Dr. Hira L. Koul's 22nd Ph.D. student. She graduated in 2004 and would like to thank Hira for his patient guidance and generous support for more than ten years now. The author is also very grateful to Hira for his academic advice and knowledge and many insightful discussions and suggestions. This manuscript is specially dedicated to him on the occasion of his 70th Birthday.

References

Berkes I, Philipp W (1997) Approximation theorems for independent and weakly dependent random vectors. Ann Probab 7:29–54

Bloomfield P (1992) Trends in global temperatures. Climatic Change 21:275–287

Bosq D (1998) Nonparametric Statistics for Stochastic Processes: Estimation and Prediction. Springer, New York

Delgado MA (1993) Testing the equality of nonparametric regression curves. Stat Probab Lett 17:199–204

Dehling H, Mikosch T, Sϕrensen M (2002) Empirical Process Techniques for Dependent Data. Birkhauser, Boston. MR1958776

Fan J, Yao Q (2003) Nonlinear Time Series: Nonparametric and Parametric Methods. Springer Verlag, New York

Hall P, Hart JD (1990) Bootstrap test for difference between means in nonparametric regression. J Amer Stat Assoc 85:1039–1049

Hall P, Heyde CC (1980) Martingale limit theory and its Applications. Academic Press, Inc., New York

Koul HL (2002) Weighted Empirical Processes in Dynamic Nonlinear Models. Springler, New York

Koul HL, Li F (2005) Testing for Superiority among two time series. Statistical Inference for Stochastic Processes. 8:109–135

Koul HL, Schick A (1997) Testing the equality of two nonparametric regression curves.J Stat Plann Inference 65:293–314

Kulasekera KB (1995) Comparison of regression curves using quasi-residuals. J Am Stat Assoc 90:1085–1093

Li F (2008) Asymptotic properties of some kernel smoothers. Preprint. pr08-03, Department of Mathematical Sciences. Indiana University Purdue University Indianapolis.

Li F (2009) Testing for the equality of two autoregressive functions using quasi residual. Communicat Stat-Theory Meth 38(9):1404–1421

Nadaraya EA (1964) On estimating regression. Theory Probab Appl 9:141–142

Resnick SI (1992) Adventures in Stochastic Processes. Springer, New York

Scheike TH (2000) Comparison of non-parametric regression functions through their cumulatives. Stat Probab Lett 46:21–32

Watson GS (1964) Smooth regression analysis. Sankhy$\bar{a}$ Ser A 26:359–372

Chapter 16
Testing for Long Memory Using Penalized Splines and Adaptive Neyman Methods

Linyuan Li and Kewei Lu

16.1 Introduction

Autoregressive moving average (ARMA) and autoregressive integrated moving average (ARIMA) models have been used extensively to analyze experimental data that have been observed at different points in time. The landmark work of Box and Jenkins (1970) developed a systematic class of ARIMA models to handle time-correlated modeling and forecasting. However, many economic and financial time series, e.g., inflation and interest rates, exhibit slow decay in correlation which is often referred to as long-range dependence or long memory. These time series are not well captured by ARMA or ARIMA models. Most often, long memory is modeled by fractionally integrated processes of order d ($0 < d < 1$). Non-integer orders of integration provoke the concept of fractional cointegration; see Granger and Joyeux (1980). Hence, there is an increasing number of econometric papers which are concerned with the tests of short memory against long memory processes. The literature on long memory processes is very extensive, see, e.g., the recent monograph by Giraitis et al. (2012) and the references cited therein.

Several test statistics have been proposed in the literature for testing short memory against long memory. Among them, we mention the KPSS statistic by Kwiatkowski et al. (1992), the Lagrange multiplier (LM) test by Robinson (1994) and a variation of LM test by Tanaka (1999). Above tests typically assume a known behavior of the underlying short memory process. Hence, they usually fit a parametric model and whiten the data prior to testing. The difficulty is that one usually does not know the true short memory model, so the fitted model may not be appropriate. In this case, the tests based on the whitened data from the wrong-fitted model may not have the right sizes or levels. Thus, the powers of the corresponding tests may not be reliable. Therefore, one could reach a wrong conclusion. In order to overcome this

2000 Mathematics Subject Classification: Primary: 62F03; Secondary: 62F05, 62M10

L. Li (✉) · K. Lu
Department of Mathematics and Statistics, University of New Hampshire,
Durham, NH 03824, USA
e-mail: linyuan@cisunix.unh.edu

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_16,

issue, Harris et al. (2008) proposed a test which doesn't depend on any short memory parametric model assumption. The rationale for their tests is that, for short memory processes, there exist significant low order sample autocovariances and their high order autocovariances are typically negligible. Thus, they constructed the test from appropriately chosen and weighted high order sample autocovariances to effectively eliminate the effects induced by the short memory. Simulation studies demonstrate that their tests exhibit very good size control across a range of stationary short memory processes and display particularly good power for long memory alternatives.

In this chapter, we propose a new test statistic based on the estimate of spectral density using penalized splines method, which does not assume any short memory parametric models either. For a typical short memory process or a general linear process, its spectral density typically is very smooth. So it can be estimated very well using penalized splines approach with simple rules of thumb for the selection of the number of knots. Ruppert et al. (2003, 2009) and Li and Ruppert (2008) have shown that penalized splines estimate does not depend on the number of knots crucially, as long as the number of knots exceeds the minimum rate with the corresponding sample size. Thus, the corresponding residuals behave like random noises. Therefore, the test based on the corresponding residuals could eliminate the effects induced by short memory. The advantage of our new approach is that our test does not depend on any parameter-selection crucially, because of the adaptivity of the penalized splines estimate. The main contribution of our approach is that our proposed test is completely data-driven or adaptive, avoiding the need to select any smoothing parameters. Under a very general short memory linear process assumption, our test follows a known distribution asymptotically. We perform Monte Carlo simulation experiments that demonstrate that our new statistic can exhibit very good size control across a range of stationary short memory processes and display a very good power property to the well-known tests for a short memory null hypothesis against long memory alternatives.

The remainder of this chapter is organized as follows. We review several current test statistics in Sect. 16.2. In Sect. 16.3, we introduce penalized splines estimate for spectral density, explain Fan's (1996) adaptive Neyman test, and discuss how it can be used in our hypothesis testing. The asymptotic result of our test is presented too. In Sect. 16.4, we provide simulation studies and demonstrate that our proposed new test is very competitive with current test statistics. We conclude in Sect. 16.5 with some remarks. Proofs of the main results are relegated to the Appendix.

16.2 Some Current Test Statistics

Consider the $I(d)$ process z_t

$$(1-L)^d z_t = u_t, \quad t = 1, 2, \ldots, T, \tag{16.1}$$

where u_t is a zero mean stationary short memory process and L is a backward shift operator. Our hypothesis testing problem is

$$H_0 : d = 0 \qquad \text{versus} \qquad H_1 : 0 < d < 0.5. \tag{16.2}$$

In other words, we wish to test the null hypothesis that z_t has $I(0)$ stationary short memory against the alternative that it has $I(d)$ stationary long memory $(0 < d < .5)$. The autocovariance function of z_t is denoted as $\gamma_j = E(z_t z_{t-j})$, which is absolutely summable under H_0 but not under H_1.

Several test statistics have been proposed in the literature for testing such hypotheses. Among them, we mention the KPSS statistic by Kwiatkowski et al. (1992), Lagrange multiplier (LM) test and its variation by Robinson (1994) and Tanaka (1999), and a recent test $\hat{S}_k$ by Harris et al. (2008).

Kwiatkowski et al. (1992) proposed a test for the hypothesis that the deviations of a series from deterministic trend are short memory against $I(d)$ alternatives. For our simpler model (16.1) case (i.e., without a linear trend), their test statistic, known as KPSS statistic, can be simplified as

$$\text{KPSS} = T^{-2} \sum_{t=1}^{T} S_t^2 / s^2(l), \tag{16.3}$$

where S_t is a partial sum process: $S_t = \sum_{i=1}^{t} z_i$, $t = 1, 2, \ldots, T$ and $s^2(l) = T^{-1} \sum_{t=1}^{T} z_t^2 + 2\, T^{-1} \sum_{s=1}^{l} W(s, l) \sum_{t=s+1}^{T} z_t z_{t-s}$ with $W(s, l) = 1 - s/(l+1)$ is the Newey-West (1987) estimator of the long-run variance of the process u_t. Lee and Schmidt (1996) showed that the KPSS test is a consistent test of short memory against long memory. Based on their simulations, they concluded that a rather large sample size, such as $T = 500$ or $1,000$, will be required to distinguish a long memory process from a short memory process. Moreover, the finite sample performance of the KPSS test depends on the selection of l. With a more strongly autocorrelated series, a larger value of l is required to control size distortions under the null. However, choosing a large value of l will reduce the power against the long memory substantially.

Robinson (1994) and Tanaka (1999) proposed the following LM test statistic under the assumption that u_t is a Gaussian white noise

$$\tilde{N} = \sqrt{T} \sum_{j=1}^{T-1} \frac{\tilde{\gamma}_j}{j}, \tag{16.4}$$

where $\tilde{\gamma}_j = T^{-1} \sum_{t=j+1}^{T} z_t z_{t-j}$. They demonstrated that, when u_t is a white noise, $\tilde{N}$ (when suitably studentized) has a standard normal limiting distribution under the null hypothesis. However, in the more general case when u_t is an autocorrelated short memory process, the statistic has an asymptotic size of either zero or one, if standard normal critical values are used. For details, see Harris et al. (2008). In order to overcome this difficulty, Tanaka (1999) suggested that the $\tilde{\gamma}_j$ be calculated not from the z_t but instead using the residuals from an ARMA model estimated for z_t. The resulting statistic has been demonstrated to be asymptotically centered at zero in the case where z_t is generated by a stationary ARMA model and assuming the correct model is fitted.

The effectiveness of the above LM test depends on whether the fitted model is appropriate or not. In practice, one does not know the true underlying short memory model. Hence, the inference based on wrong-fitted model may be misleading. In order to eliminate this effect induced by the dependence in u_t under H_0, Harris et al. (2008) provided a modified statistic $\hat{S}_k$, which is based on $\tilde{N}$. Specifically, they considered the truncated statistic

$$\tilde{N}_k = \sqrt{T-k}\sum_{j=k}^{T-1}\frac{\tilde{\gamma}_j}{j-k+1}, \tag{16.5}$$

which can be viewed as $\tilde{N}$ calculated only from the sample autocovariances at lag k and above. They showed that under H_0, $\hat{S}_k$, which is an appropriately studentized $\tilde{N}_k$ (for details, see Harris et al. 2008, p. 146), has an asymptotic standard normal null distribution. Therefore, provided that the class of linear processes is appropriate, there is no need to postulate and fit a parametric model for any short memory behavior. However, the choice of the starting point of the lag k is very critical. The authors assumed that $k = \sqrt{c\,T}$ for some constant $c > 0$. They demonstrated that the choice of the scaling parameter c, which controls the truncation k, has no asymptotic effect on the properties of $\tilde{N}_k$. Nevertheless, its finite sample performance will inherently depend on the specific value selected by the users. Thus, they considered the values $c = .5,\ 1.0,\ 2.0$ in their Monte Carlo simulation.

16.3 Our New Adaptive Test Statistic

In this chapter, we pursue another nonparametric approach to eliminating the need to postulate and fit a parametric model for u_t. Our test statistic is constructed completely by the data, and does not involve any unknown smoothing parameters seriously. In particular, our test statistic is based on the estimate of the spectral density using penalized splines method. Specifically, the spectral density of z_t is defined as

$$f(\omega) = \frac{1}{2\pi}\sum_{j=-\infty}^{\infty}\gamma_j e^{-i\omega j}, \quad -\pi < \omega \le \pi, \tag{16.6}$$

whenever this exists. Because of the symmetry of $f(\omega)$, we only need to consider $\omega \in [0,\pi)$. It is well known that the basic tool to estimate the spectral density $f(\omega)$ is its periodogram defined at the Fourier frequencies $\omega_j = 2\pi j/T$, $\omega_j \in [0,\pi)$, by

$$I(\omega_j) = \frac{1}{2\pi T}\left|\sum_{t=1}^{T} z_t e^{-it\omega_j}\right|^2. \tag{16.7}$$

For the simplicity and convenience of the exposition of our test, we assume the sample size $T = 2n$. Also, let $x_j = \omega_{j-1}/(2\pi) + 1/T$, $y_j = I(x_j)$, $j = 1, 2, \ldots, n$.

Then, these new notations match those in Li and Ruppert (2008). Therefore, in the proof part of the main result in Appendix below, we can cite the results from Li and Ruppert (2008) directly without any confusion. If z_t is a stationary linear process with i.i.d. Gaussian innovations, then, by Theorem 10.3.2 of Brockwell and Davis (1991), y_j are asymptotically exponentially distributed with mean $f(x_j)$ and that they are approximately independent. That is, with R_j denoting an asymptotically negligible term, we have

$$y_j = f(x_j) + f(x_j)\eta_j + R_j, \tag{16.8}$$

where the random variables η_j for $j = 2, 3, \ldots, n$ are i.i.d. with $\eta_j = \chi^2(2)/2 - 1$ and $\eta_1 = \chi^2(1) - 1$. Above model (16.8) could be written as

$$y_t = f(x_t) + \epsilon_t, \quad t = 1, 2, \ldots, n, \tag{16.9}$$

with $\epsilon_t = f(x_t)\eta_t + R_t$. Li and Ruppert (2008) considered model (16.9) and provided an estimate for f using penalized splines method. Note that, in the estimation step, Li and Ruppert (2008) do not need the assumption that ϵ_t are independent (it is required in the derivation of the limit distribution of their estimator). In this chapter, we simply apply their estimation method to estimate f based on the periodogram y_t as that in Li and Ruppert (2008). Since the remainder term R_j is negligible, our model (16.8) is equivalent to their model (16.9) (with independent errors) asymptotically.

The penalized spline method approximates the regression function f with splines. Typically, the modeling bias due to approximation to the regression function by a spline is negligible compared to the shrinkage bias due to estimation. Specifically, one approximates regression with a spline $f(x) = \sum_{k=1}^{K(n)+p} b_k B_k^{[p]}(x)$, where $\{B_k^{[p]} : k = 1, 2, \ldots, K(n) + p\}$ is the pth degree B-spline basis with knots $0 = \kappa_0 < \kappa_1 < \ldots < \kappa_{K(n)} = 1$. The value of $K(n)$ will depend upon n as discussed below. The penalized least-squares estimator $\hat{b} = (\hat{b}_1, \ldots, \hat{b}_{K(n)+p})$ minimizes

$$\sum_{j=1}^{n} \left\{ y_j - \sum_{k=1}^{K(n)+p} b_k B_k^{[p]}(x_j) \right\}^2 + \lambda_n^* \sum_{k=m+1}^{K(n)+p} \{\Delta^m(b_k)\}^2, \qquad \lambda_n^* \geq 0, \tag{16.10}$$

where Δ is the difference operator, i.e., $\Delta b_k = b_k - b_{k-1}$, m is a positive integer, and $\Delta^m = \Delta(\Delta^{m-1})$. The nonparametric regression estimator $\hat{f}(x) = \sum_{k=1}^{K(n)+p} \hat{b}_k B_k^{[p]}(x)$ is called the P-spline estimator. For details, see Li and Ruppert (2008, p. 415) and Eilers and Marx (1996). In this chapter, we only consider the first-order penalized estimator using zero-degree splines, that is piecewise constant, $m = 1$ and $p = 0$, with equally spaced knots $\kappa_0 = 0, \kappa_1 = 1/K(n), \kappa_2 = 2/K(n), \ldots, \kappa_{K(n)} = 1$. In particular, $B_k^{[0]}(x) = \chi_{\{\kappa_{k-1} < x \leq \kappa_k\}}, 1 \leq k \leq K(n)$, where χ is the indicator function, and $\hat{f}(x) = \hat{b}_k$ for any $x \in (\kappa_{k-1}, \kappa_k], k = 1, 2, \ldots, K(n)$. For this special case, we have explicit expressions for those $\hat{b}_k$ values, which solve $\{I_{K(n)} + \lambda_n (D^m)' D^m\}\hat{b} = \bar{y}$. For more details, again see Li and Ruppert (2008, p. 418–422).

Let $\hat{\eta}_j = (y_j - \hat{f}(x_j))/\hat{f}(x_j)$, where $\hat{f}(x_j)$ is obtained from the above penalized spline estimate as in Li and Ruppert (2008). We propose the following statistic as a test statistic

$$\sum_{j=1}^{n} \hat{\eta}_j^2 = \sum_{j=1}^{n} \frac{(y_j - \hat{f}(x_j))^2}{\hat{f}(x_j)^2}. \tag{16.11}$$

Under the null hypothesis of short memory processes, the spectral densities are more or less very smooth. Hence, splines approximate the underlying regression (or spectral density in our case) very well with certain number of knots, which is equivalent to saying that the modeling bias is asymptotically negligible. Thus, the penalized splines method estimates the spectral densities very well across a broad range of short memory processes. Therefore, $\hat{\eta}_j$ behaves very similar to $\eta_j,\ j = 1, 2, \ldots, n$ under null hypothesis. Furthermore, one can derive a known limit distribution for the above statistic in (16.11) under the null, which can be used to carry out the hypothesis testing.

On the other hand, under long memory alternatives, one observes $f(x) \to \infty$ as $x \to 0$. Hence, splines could not approximate the regression function very well near the zero. Thus, the modeling bias will not be negligible and the the penalized spline estimator $\hat{f}(x)$ will not do a good job in estimating $f(x)$ when x is near 0. Consequently, $\hat{\eta}_j^2$ tends to be relatively large when j's are small. Thus, under H_1, the statistic in (16.11) tends to be very large. Therefore, testing hypothesis could be reached based on the statistic $\sum_{j=1}^{n} \hat{\eta}_j^2$. However, based on the observation of Fan (1996), this test statistic has very low power at certain alternatives. The main reason is that the test involves too many individual terms (n terms in total in this case), which accumulate too much stochastic errors.

Specifically, we will use Fan's (1996) canonical multivariate normal hypothesis testing procedure to construct a new test statistic. Fan (1996) considered the following testing problem: Let $\mathbb{X} \sim N(\theta, \mathbb{I}_n)$ be an n-dimensional normal random vector. One wishes to test

$$H_0 : \theta = 0 \qquad \text{versus} \qquad H_1 : \theta \neq 0. \tag{16.12}$$

He demonstrated that the test based on $||\mathbb{X}||^2$ has very low power at alternative $\theta = \theta_0 \neq 0$, since the test includes too many individual terms and has too many stochastic errors. Neyman (1937) proposed testing the first m-dimensional subspace, leading to the test statistic $\sum_{i=1}^{m} X_i^2$. Based on the power calculation, Fan (1996) proposed adaptive Neyman test for the testing problem (16.12):

$$T_{AN}^* = \max_{1 \le m \le n} \frac{1}{\sqrt{2m}} \sum_{i=1}^{m} (X_i^2 - 1). \tag{16.13}$$

For more details, see Fan (1996). Large values of test T_{AN}^* tend to reject null hypotheses H_0 in (16.12). With theoretical power calculation and empirical simulation studies, Fan (1996) showed that the adaptive Neyman test has a higher power than those of Kolmogorov-Smirnov and Cramér-Von Mises tests.

Although Fan (1996) considered hypothesis testing on n-dimensional normal distribution in (16.12), our testing problem is conceptually equivalent to his test problem. Under H_0, our $\hat{\eta}_j,\ j = 1, 2, \ldots, n$ behave like a sequence of i.i.d. random variables η_j with mean 0. Hence, we can construct a similar test as in Fan (1996). In particular, we propose the following analogous test statistic

$$T_S^* = \max_{1 \le m \le M_n} \frac{1}{\sqrt{8m}} \sum_{i=1}^{m} (\hat{\eta}_i^2 - 1), \tag{16.14}$$

where $M_n = n^{4/5}/(\log\log n)^2$. First, one notices that our proposed test is maximized in m over the range $[1,\ M_n]$ instead of $[1,\ n]$. This modification of the range mainly follows from a technical reason (for details, see the proof of Theorem in Appendix). From the large sample theoretical point of view, this difference between M_n and n is hardly significant. Second, from the practical implementation point of view, one does not have to apply the test with maximization of m over the entire range $[1,\ n]$. The main reason is that the difference between short memory and long memory is mainly carried by those $\hat{\eta}_j$'s with small j's. Thus, with this observation, we know those $\hat{\eta}_i^2$, with smaller j's, are relatively large in alternatives. Thus, the test T_S^* attains the maximum usually when m belongs to range $[1,\ M_n]$. Our simulation studies confirm this observation. In our extensive simulation studies, we find that the test with the maximum over range $[1,\ M_n]$ is not significantly different from that with the maximum over the whole range $[1,\ n]$. We tried both tests (over the range $[1,\ M_n]$ and $[1,\ n]$), and find that their results are very close. Therefore, for the definiteness and convenience, we simply use the test with the maximization of m over $[1,\ n]$ for simplicity. For more discussion on this similar modification over the maximization range, see Fan and Huang (2001, p. 642) and Fan and Yao (2003, p. 300).

Following Fan (1996), we normalize the test statistic to obtain

$$T_S = \sqrt{2\log\log M_n}\, T_S^* - \{2\log\log M_n + .5\log\log\log M_n - .5\log(4\pi)\}. \tag{16.15}$$

Before we provide the asymptotic null distribution of T_S, we require the following technical assumptions on the short memory process u_t in (16.1) to simplify technical arguments in the proof.

A1. The series u_t is a linear Gaussian process; that is $u_t = \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j}$ with $\varepsilon_t \sim i.i.d.\ N(0, \sigma^2)$ and $\sum_{j=-\infty}^{\infty} |\psi_j||j|^2 < \infty$.

A2. The spectral density $f(x)$ of u_t is positive; i.e., it satisfies $\inf_x f(x) > 0$.

The above assumptions are used in the estimation and testing for spectral density, e.g., Fan and Zhang (2004). We mention that, under the above assumption **A1**, one obtains that the second derivative $f''(\cdot)$ of the spectral density $f(\cdot)$ is continuous on $[0,\ \pi]$. Thus, the assumption of Theorem 1 in Li and Ruppert (2008) is satisfied for our penalized splines estimation.

Similar to that in Fan (1996), we have the following asymptotic null distribution of T_S. The proof of the main theorem is postponed to Appendix.

Theorem 3.1. *Assume the assumptions* **A1** *and* **A2** *are satisfied, then, under* H_0 : $d = 0$*, we have*

$$P(T_S < x) \to \exp(-\exp(-x)), \qquad as\ n \to \infty. \tag{16.16}$$

The above limiting distribution could be used to determine the rejection region at a given significant level. Nevertheless, the rate of convergence of T_S to the above limiting distribution is very slow (one needs very large sample size n to approximate the above distribution well). Thus, for any fixed sample size T, the distribution of the test T_S (denoted with J_T) is different from its limit distribution in (16.16). Since the explicit distribution of J_T is very difficult to be derived, we will use Monte Carlo simulation studies to determine the rejection region for the finite sample size under H_0 and the power of the test under alternative H_1. With powerful computers, this computing is no longer an issue. We have MATLAB codes available for computing the distribution of J_T and are happy to provide those codes upon request.

The theoretical power of the test statistic T_S is difficult to evaluate analytically under the alternative hypothesis. However, from our simulation studies in the next section, the asymptotic power tends to 1 as the sample size becomes very large.

16.4 Simulation Studies

16.4.1 Densities Under Null Hypothesis

Because of the technical reason stated before, we propose our test T_S in (16.15) with maximization over range $[1, M_n]$. As we have discussed before, because of the nature of our test problem, this test is very close to the test with maximization over range $[1, n]$. So for the definiteness, we provide the simulation studies in this section using the test with maximization over range $[1, n]$. We still use the same notation T_S, simply replacing M_n with n.

As mentioned earlier, the theoretical limit distributions in (16.16) for T_S in (16.15) are not good approximations for finite sample size n. Thus, one typically uses simulation study to determine the rejection region for finite sample size under H_0 and calculate the power of the test under alternative H_1. Following Fan (1996), we investigate the finite-sample distributions of the test statistic T_S via simulations. As an illustration, we generate a white noise time series for u_t with length $T = 500$ and 1,000 and study $N = 5,000$ simulations. The distribution of 5,000 simulated test statistics under the null hypothesis are presented using a kernel density estimate. More precisely, let $Z_1, Z_2, \ldots, Z_N$ be a sequence of i.i.d. random variables with a common density f. Then its kernel density estimator is

$$\hat{f}_h(x) = N^{-1} \sum_{i=1}^{N} h^{-1} K\{(Z_i - x)/h\},$$

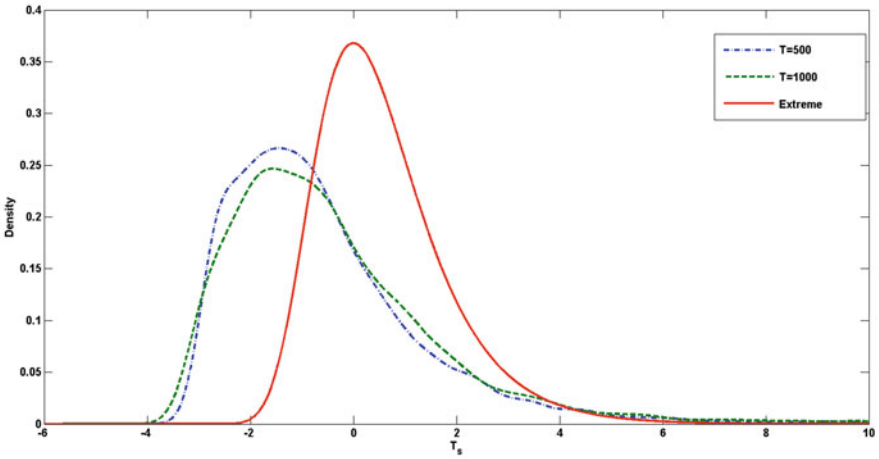


Fig. 16.1 The Estimated Densities for Test Statistic T_S under the Null Hypothesis for $T = 500$ and 1000 based on 5000 simulations, and the limiting extreme distribution in (16.16)

where K is the Gaussian kernel function $K(x) = (2\pi)^{-1/2} \exp(-x^2/2)$, and h is the bandwidth. By the normal reference bandwidth rule, $h = 1.06 * s_N * N^{-1/5}$, where s_N is the standard deviation of $Z_1, Z_2, \ldots, Z_N$. For more discussion on the selections of K and bandwidth h, see Fan and Yao (2003, Chap. 5).

Specifically, we calculated $Z_1 = T_{S,1}$, $Z_2 = T_{S,2}, \cdots, Z_{5000} = T_{S,5000}$ from the 5,000 simulations. The result is presented in Fig. 16.1. From the estimated distribution of our test statistic T_S, one can see that none of the finite sample versions are close to the theoretic limiting distribution in (16.16). This is in line with the previous findings stating that the adaptive Neyman test statistics converge rather slowly toward their asymptotic limit.

16.4.2 Empirical Sizes

This section explores the empirical sizes of the suggested testing procedure for finite sample performances. We consider an ARMA(1,1) data generating process (DGP): $(1 - \rho L)u_t = (1 + \theta L)\varepsilon_t$, $\varepsilon_t \sim$ i.i.d.$N(0, 1)$. The sample sizes are $T = 200, 500, 1,000$. From Li and Ruppert (2008), the number of knots $K(n)$ does not effect the asymptotic distribution of the penalized spline estimators, as long as the number of knots satisfies certain rates with the sample size. In our case with $m = 1$ and $p = 0$, one needs the number of knots $K(n) = Cn^{\gamma}$ with $\gamma > 2/5$. (see Li and Ruppert, 2008, p. 420). In the practical point of view, this does not provide much useful guidelines. Ruppert et al. (2003, p.126) suggest $K(n) = \min(35, n/4)$ from their empirical studies. In the following simulation studies, we select $K(n) = 3n^{0.42}$, where $n = T/2$. Thus for $T = 200, 500$, and 1,000, the number of knots are

$K(n) = 20, 30$, and 40 respectively. We find that it works well in our simulation. The advantage of the penalized spline estimator is that it does not depend on the number of knots crucially, as long as the number of knots exceeds the minimum rates with the corresponding sample size. So the corresponding test does not depend on the parameters (number of knots and penalty term λ) seriously. The penalty term automatically eliminates the autocorrelation induced by the short memory process. Table 16.1 reports the proportion of times the test statistics reject the short memory at level 5 %. The results of empirical sizes of our test are based on the empirical critical values (ECVs). We find that for the sample sizes $T = 200, 500$, and 1,000, the corresponding ECVs are around 4.57, 4.74, and 4.78 based on 5,000 simulations. The standard error of the empirical sizes is $\sqrt{.05 * .95/5,000} = .31$ %.

We also include the simulation results of three other well-known tests of long memory: Harris et al. (2008) test, KPSS test, and Robinson (1994) and Tanaka (1999) test. These numbers are taken from Harris et al. (2008) for comparison purposes. Although being undersized for ARMA(1, 1) with $\theta = -.8$ and $\rho \in \{0, .5, .7\}$, our test statistic generally provides reasonable size control across the ARMA(1, 1) parameter space. The Harris et al. (2008) test ($c = 1.0$) has good size properties in general, although with some oversizing for large values of ρ when $T = 200$ and $T = 500$. However, if one selects a smaller c, say, $c = .5$, the sizes would be deteriorated. On the other hand, the test managed to have perfect size control when $c = 2.0$. For details, see Harris et al. (2008). This suggests that the empirical sizes of their test depend largely on the selection of c. The KPSS test only achieves good size properties for small AR coefficients, and becomes oversized for large values of ρ across all of the sample sizes. The Robinson (1994)/Tanaka (1999) test assuming an ARMA(1, 1) model for u_t has very good size control for all of the ARMA(1, 1) models across all of the sample sizes, although being a little undersized for small values of AR coefficients when $T = 200$ and $T = 500$. However, the size properties of their test are based on the correct model specification of u_t, which, in reality, is usually unknown. Harris et al. (2008) considers another version of the Robinson (1994)/Tanaka (1999) test assuming an AR(1) model for u_t for comparison. The sizes are either badly oversized or undersized in those cases where the AR(1) model is incorrect while ARMA(1, 1) is the true DGP.

In those undersized cases for T_S, the periodograms near 0 are always smaller than the estimated spectral density, which leads to the failure to reject the null (because our test is one-sided, and only large values of T_S reject the null). We also notice that it is modestly oversized for ARMA(1, 1) with $\rho = .9$ and $\theta = -.8$. Increasing the sample size does reduce the size, although the effect is not obvious (the size for $T = 200$ is .09, and for $T = 1,000$ is .08). This is due to the fact that for this particular model, the spectral density near zero is very steep. Thus, the penalized spline estimator requires a fairly large sample size to approximate the true density well. Obviously, the sample size $T = 1,000$ ($n = 500$) is not large enough for our test statistic in this particular model. Other than that, there are no notable size distortions elsewhere.

Table 16.1 ARMA(1, 1) Empirical Type I Errors when d=0.0

	T=200					T=500					T=1,000				
ρ/θ	−0.8	−0.4	0.0	0.4	0.8	−0.8	−0.4	0.0	0.4	0.8	−0.8	−0.4	0.0	0.4	0.8
						T_S									
0.0	0.00	0.01	0.03	0.02	0.02	0.00	0.01	0.03	0.03	0.02	0.00	0.01	0.03	0.03	0.02
0.5	0.00	0.04	0.03	0.03	0.03	0.00	0.05	0.03	0.03	0.03	0.00	0.05	0.03	0.03	0.03
0.7	0.00	0.04	0.03	0.04	0.04	0.00	0.04	0.03	0.04	0.04	0.00	0.04	0.04	0.03	0.04
0.9	0.09	0.06	0.06	0.06	0.05	0.09	0.05	0.05	0.06	0.05	0.08	0.05	0.05	0.05	0.05
						Harris et al. (c=1.0)									
0.0	0.04	0.04	0.04	0.04	0.04	0.05	0.05	0.04	0.04	0.04	0.05	0.05	0.05	0.04	0.05
0.5	0.05	0.04	0.04	0.03	0.03	0.05	0.05	0.04	0.04	0.04	0.05	0.05	0.04	0.05	0.05
0.7	0.05	0.03	0.03	0.03	0.03	0.05	0.04	0.04	0.04	0.04	0.05	0.04	0.04	0.04	0.05
0.9	0.07	0.09	0.09	0.09	0.10	0.06	0.08	0.07	0.06	0.07	0.05	0.05	0.05	0.05	0.05
						KPSS									
0.0	0.16	0.05	0.05	0.05	0.05	0.07	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.05
0.5	0.07	0.05	0.05	0.05	0.05	0.04	0.05	0.06	0.06	0.06	0.03	0.05	0.06	0.06	0.06
0.7	0.04	0.06	0.06	0.06	0.06	0.04	0.07	0.07	0.07	0.07	0.04	0.07	0.07	0.07	0.07
0.9	0.10	0.14	0.14	0.14	0.14	0.11	0.12	0.12	0.12	0.12	0.11	0.11	0.12	0.12	0.12
						Robinson(1994)/Tanaka(1999) ARMA(1,1)									
0.0	0.01	0.02	0.02	0.05	0.05	0.07	0.05	0.05	0.05	0.05	0.05	0.05	0.06	0.06	0.06
0.5	0.01	0.01	0.05	0.05	0.05	0.04	0.05	0.06	0.06	0.06	0.03	0.05	0.06	0.06	0.06
0.7	0.02	0.01	0.06	0.06	0.06	0.04	0.07	0.07	0.07	0.07	0.04	0.07	0.07	0.07	0.07
0.9	0.02	0.01	0.14	0.14	0.14	0.11	0.12	0.12	0.12	0.12	0.11	0.11	0.12	0.12	0.12

16.4.3 Power Comparisons

In this section, we compare the powers of our proposed adaptive Neyman test T_S with those of Harris et al. (2008) test, KPSS test, and Robinson (1994)/Tanaka (1999) test. The empirical powers of the tests are evaluated under the alternatives $d = 0.2$ and $d = 0.4$. The results are summarized in Tables 16.2–16.3. As the spectral density of long memory near 0 is approaching infinity, our penalized-spline-based test T_S is a powerful test procedure. We also notice that the power of T_S increases with both T and d, as expected.

The power of T_S is greater than that of the Harris et al. (2008) test almost everywhere across all of the sample sizes when $\rho \in \{0, 0.5, 0.7\}$. When the AR coefficient takes the largest value of $\rho = 0.9$, the Harris et al. (2008) test with $c = 1.0$ has greater power than T_S when $d = 0.2$. Nevertheless, the two tests are almost equally powerful for $\rho = 0.9$ when $d = 0.4$ (with T_S slightly outperforming the Harris et al. (2008) test when $T = 200$, and the latter slightly outperforming the former when $T = 500$). Moreover, T_S has a significant power improvement over the Harris et al. (2008) test for the other AR coefficients when $d = 0.4$. Considering the fact that the performance of the Harris et al. (2008) test depends on the selection of the truncation parameter c, our adaptive Neyman test T_S, which avoids the need to select any smoothing parameter, would appear to provide a good competitor to the Harris et al. (2008) test as a test of short memory against long memory, particularly for large stationary values of d.

The KPSS test generally has greater power than T_S in most cases when $d = 0.2$. However, when d is increased to 0.4, T_S takes the lead everywhere. With its superior control of size over the KPSS test, T_S would tend to be a competitive testing procedure. The Robinson (1994)/Tanaka (1999) test has great power in some models while it has very small power in the others. Furthermore, in some cases, its power curiously decreases as d increases. As we mentioned earlier, the performance of the Robinson (1994)/Tanaka (1999) test is based on the correct model assumption of the true DGP. If it is misspecified, as illustrated in Harris et al. (2008) by assuming a AR(1) model while the true DGP is ARMA(1, 1), the power of the test would be largely affected, with some extreme situations where the power would be either 1 or 0.

We do find that the power of T_S is 0 for the model ARMA(1, 1) with $\rho = 0$ and $\theta = -0.8$ when $T = 200$ and $d = 0.2$. In this particular case, the periodograms near 0 are smaller than the estimated spectral density. Even though one would expect the periodograms near 0 to be large with $d = 0.2$, in fact they still fall behind the estimated spectral density because of the small sample size, leading to a failure to reject the short memory. Increasing the sample size, however, improves the power significantly; for example, the power is increased to 0.11 when the sample size is 1,000.

For the nonstationary process z_t in (16.1) with $d > 0.5$ case, we find that T_S is still comparable to its competitors in general (for the brevity of expression, the simulation results are not included in this chapter. They are available upon request). When d is approaching 1 ($d = 0.8$ and 1.0), Harris et al. (2008) test performs slightly better than T_S. Both T_S and KPSS tests perform equally well. Therefore, it is hoped that the proposed test statistic will present a useful complement to the the current tests for

Table 16.2 ARMA(1, 1) Empirical Powers when d=0.2

	T=200					T=500					T=1,000				
ρ/θ	−0.8	−0.4	0.0	0.4	0.8	−0.8	−0.4	0.0	0.4	0.8	−0.8	−0.4	0.0	0.4	0.8
						T_S									
0.0	0.00	0.12	0.12	0.11	0.10	0.04	0.25	0.18	0.17	0.17	0.11	0.42	0.27	0.25	0.25
0.5	0.03	0.10	0.09	0.10	0.09	0.08	0.17	0.17	0.16	0.16	0.22	0.25	0.24	0.24	0.24
0.7	0.11	0.10	0.10	0.10	0.10	0.22	0.15	0.16	0.15	0.15	0.36	0.24	0.24	0.23	0.23
0.9	0.15	0.16	0.17	0.17	0.16	0.19	0.19	0.19	0.19	0.18	0.25	0.25	0.25	0.26	0.26
						Harris et al. (c=1.0)									
0.0	0.05	0.07	0.07	0.07	0.08	0.05	0.11	0.14	0.14	0.13	0.06	0.15	0.20	0.19	0.20
0.5	0.06	0.07	0.08	0.08	0.07	0.08	0.14	0.14	0.14	0.15	0.10	0.20	0.20	0.20	0.21
0.7	0.07	0.08	0.08	0.08	0.08	0.12	0.14	0.15	0.15	0.15	0.16	0.22	0.22	0.21	0.22
0.9	0.18	0.24	0.23	0.23	0.24	0.24	0.28	0.29	0.28	0.28	0.28	0.31	0.32	0.32	0.32
						KPSS									
0.0	0.14	0.14	0.14	0.14	0.14	0.15	0.21	0.20	0.20	0.20	0.19	0.25	0.25	0.25	0.25
0.5	0.09	0.14	0.14	0.14	0.14	0.15	0.20	0.20	0.20	0.20	0.21	0.25	0.25	0.25	0.25
0.7	0.10	0.15	0.15	0.15	0.15	0.17	0.20	0.20	0.20	0.20	0.24	0.25	0.25	0.25	0.25
0.9	0.22	0.25	0.25	0.25	0.25	0.26	0.28	0.28	0.28	0.28	0.30	0.31	0.31	0.31	0.31
						Robinson(1994)/Tanaka(1999) ARMA(1,1)									
0.0	0.05	0.22	0.02	0.16	0.28	0.15	0.61	0.14	0.60	0.75	0.35	0.90	0.42	0.91	0.97
0.5	0.03	0.03	0.03	0.03	0.03	0.19	0.18	0.19	0.18	0.18	0.46	0.47	0.49	0.47	0.45
0.7	0.01	0.01	0.01	0.07	0.16	0.07	0.08	0.07	0.16	0.30	0.23	0.28	0.22	0.32	0.54
0.9	0.03	0.13	0.35	0.57	0.72	0.15	0.13	0.51	0.81	0.94	0.42	0.16	0.72	0.96	1.0

Table 16.3 ARMA(1, 1) Empirical Powers when d=0.4

	T=200					T=500					T=1,000				
ρ/θ	−0.8	−0.4	0.0	0.4	0.8	−0.8	−0.4	0.0	0.4	0.8	−0.8	−0.4	0.0	0.4	0.8
						T_S									
0.0	0.33	0.42	0.40	0.40	0.41	0.62	0.54	0.54	0.53	0.53	0.82	0.69	0.66	0.67	0.68
0.5	0.45	0.42	0.40	0.40	0.40	0.59	0.54	0.54	0.53	0.52	0.74	0.68	0.67	0.66	0.67
0.7	0.41	0.41	0.42	0.41	0.40	0.53	0.53	0.53	0.53	0.53	0.68	0.67	0.68	0.68	0.67
0.9	0.46	0.47	0.47	0.47	0.47	0.55	0.57	0.56	0.56	0.56	0.69	0.68	0.69	0.69	0.69
						Harris et al. (c=1.0)									
0.0	0.08	0.18	0.18	0.19	0.19	0.14	0.39	0.39	0.39	0.40	0.24	0.58	0.58	0.58	0.59
0.5	0.15	0.19	0.19	0.19	0.20	0.30	0.40	0.41	0.40	0.40	0.49	0.58	0.59	0.59	0.59
0.7	0.16	0.21	0.22	0.22	0.21	0.37	0.42	0.41	0.42	0.42	0.56	0.59	0.60	0.60	0.60
0.9	0.37	0.43	0.42	0.43	0.43	0.55	0.60	0.60	0.60	0.61	0.69	0.71	0.71	0.71	0.71
						KPSS									
0.0	0.24	0.26	0.26	0.26	0.26	0.38	0.37	0.37	0.37	0.37	0.52	0.49	0.49	0.50	0.50
0.5	0.24	0.26	0.26	0.26	0.26	0.38	0.37	0.37	0.37	0.37	0.50	0.49	0.50	0.50	0.50
0.7	0.24	0.27	0.27	0.27	0.27	0.36	0.37	0.36	0.37	0.36	0.49	0.50	0.50	0.50	0.50
0.9	0.34	0.37	0.37	0.37	0.37	0.42	0.44	0.43	0.43	0.44	0.53	0.54	0.54	0.54	0.54
						Robinson(1994)/Tanaka(1999) ARMA(1,1)									
0.0	0.17	0.27	0.06	0.16	0.37	0.61	0.22	0.39	0.68	0.90	0.92	0.07	0.84	0.96	1.0
0.5	0.04	0.03	0.02	0.04	0.12	0.23	0.31	0.10	0.11	0.21	0.51	0.76	0.28	0.23	0.35
0.7	0.08	0.01	0.16	0.45	0.70	0.52	0.04	0.23	0.67	0.92	0.91	0.12	0.33	0.86	0.99
0.9	0.05	0.48	0.88	0.98	0.99	0.10	0.64	0.99	1.0	1.0	0.23	0.80	1.0	1.0	1.0

Table 16.4 ARMA(1, 1) Empirical Type I Errors when d=0.0 using ECVs assuming i.i.d. normal errors for the i.i.d. non-Gaussian uniform distribution

	T=200					T=500					T=1,000				
ρ/θ	−0.8	−0.4	0.0	0.4	0.8	−0.8	−0.4	0.0	0.4	0.8	−0.8	−0.4	0.0	0.4	0.8
0.0	0.00	0.00	0.02	0.02	0.02	0.00	0.01	0.03	0.03	0.02	0.00	0.01	0.03	0.02	0.03
0.5	0.00	0.03	0.02	0.03	0.03	0.00	0.04	0.04	0.02	0.04	0.00	0.05	0.04	0.02	0.04
0.7	0.00	0.04	0.04	0.03	0.03	0.00	0.04	0.04	0.03	0.04	0.00	0.04	0.04	0.04	0.03
0.9	0.09	0.06	0.06	0.06	0.07	0.09	0.06	0.05	0.06	0.05	0.08	0.06	0.05	0.05	0.05

Table 16.5 ARMA(1, 1) Empirical Type I Errors when d=0.0 using ECVs assuming i.i.d. normal errors for the i.i.d. non-Gaussian lognormal distribution

	T=200					T=500					T=1,000				
ρ/θ	−0.8	−0.4	0.0	0.4	0.8	−0.8	−0.4	0.0	0.4	0.8	−0.8	−0.4	0.0	0.4	0.8
0.0	0.00	0.01	0.02	0.02	0.02	0.00	0.01	0.02	0.02	0.02	0.00	0.01	0.03	0.03	0.02
0.5	0.00	0.04	0.03	0.03	0.02	0.00	0.04	0.03	0.03	0.02	0.00	0.04	0.03	0.03	0.03
0.7	0.00	0.03	0.03	0.03	0.03	0.00	0.03	0.03	0.03	0.03	0.00	0.04	0.03	0.03	0.03
0.9	0.09	0.07	0.06	0.07	0.07	0.07	0.05	0.05	0.05	0.05	0.07	0.05	0.05	0.05	0.05

testing stationary short memory versus long memory alternatives. All the computer codes are written in R and MATLAB, and are available upon request.

16.4.4 Robustness Study Under Non-Gaussian Errors

In the main theorem and previous simulation studies, we assume that the series $u_t = \sum_{j=-\infty}^{\infty} \psi_j \varepsilon_{t-j}$ is a linear Gaussian process (i.e., $\varepsilon_t's$ are $i.i.d.$ normal variables) because of the technical argument reason. In this section, we would like to investigate whether the results remain valid when the errors ε_t are not Gaussian. In particular, we consider uniform and lognormal distributions for ε_t respectively in the following simulation studies. Specifically, we report the Empirical Type I Errors (or empirical levels) using the ECVs assuming i.i.d. normal errors for the i.i.d. non-Gaussian errors (normalized uniform and lognormal distributions respectively). The results are displayed in Tables 16.4 and 16.5. It is seen that under non-normal errors (uniform and lognormal distributions), the empirical levels using the ECVs for normal errors are reasonably close to those for normal errors. This suggests that if the errors are not Gaussian (which is assumed in the main theorem), the test statistics are reasonably robust to the other non-Gaussian errors from this modest simulation study.

16.5 Conclusion

In this chapter, we propose a new test statistic based on penalized splines method and ideas in Fan's (1996) adaptive Neyman tests. Our test does not assume any short memory parametric models and is completely data-driven or adaptive, avoiding the need to select any smoothing parameters. Under a general linear process null hypothesis, our test follows a known distribution asymptotically. Since the convergence of

the proposed test statistics toward their asymptotic distributions is known to be slow, we apply Monte Carlo simulation method to investigate their distributions and powers. Our test is compared to the three well-known tests in literature (KPSS, LM in Robinson (1994) and Tanaka (1999), and the test in Harris et al. (2008)). The empirical powers of our proposed test statistic are always competitive and, for certain alternatives, the most powerful.

16.6 Appendix

Proof of Theorem 3.1 During the following proof, we use C to denote a positive generic constant. The notation $o_p(1)$ is short for a sequence of random variables that converge to zero in probability. The expression $O_p(1)$ denotes a sequence that is bounded in probability. Similarly, $o(1)$ and $O(1)$ are short notations for deterministic sequences. More generally, $U_n = o_p(V_n)$ means $|U_n/V_n| = o_p(1)$, and $U_n = o(V_n)$ indicates $|U_n/V_n| = o(1)$. Similar notations apply to $U_n = O_p(V_n)$ and $U_n = O(V_n)$. There are many rules of calculations with o and O symbols, which we apply without comment. For example, $o_p(1) + o_p(1) = o_p(1)$, $o_p(1) + O_p(1) = O_p(1)$, and $o_p(1)\, O_p(1) = o_p(1)$.

The idea of the proof is similar to that of Theorem 1 in Fan and Huang (2001). They considered the Goodness-of-fit to parametric regression models, whereas we consider the nonparametric (penalized splines) estimates to the spectral densities. So the bias of our estimates (penalized splines estimates) have different rates than theirs (least squares estimate). Because of this difference, we propose our test statistic with a modification on the maximization range over $[1,\ M_n]$. Other than this difference, overall structure of the proof is the same. Denote $f(x_j)$ as f_j and $\hat{f}(x_j)$ as $\hat{f}_j$, in view of (16.8) and (16.11), we have

$$
\begin{aligned}
\sum_{j=1}^{m} \hat{\eta}_j^2 &= \sum_{j=1}^{m} \frac{(y_j - \hat{f}_j)^2}{\hat{f}_j^2} \\
&= \sum_{j=1}^{m} \left(\frac{f_j - \hat{f}_j + f_j \eta_j + R_j}{f_j} \cdot \frac{f_j}{\hat{f}_j} \right)^2 \\
&= \sum_{j=1}^{m} \left[\left(\eta_j + \frac{f_j - \hat{f}_j}{f_j} + \frac{R_j}{f_j} \right) \left(1 + \frac{f_j - \hat{f}_j}{\hat{f}_j} \right) \right]^2 \\
&= \sum_{j=1}^{m} \left[\eta_j + \eta_j \frac{f_j - \hat{f}_j}{\hat{f}_j} + \frac{f_j - \hat{f}_j}{f_j} + \frac{(f_j - \hat{f}_j)^2}{f_j \hat{f}_j} + \frac{R_j}{f_j} + \frac{R_j (f_j - \hat{f}_j)}{f_j \hat{f}_j} \right]^2 \\
&= \sum_{j=1}^{m} \left[\eta_j^2 + \eta_j^2 \frac{(f_j - \hat{f}_j)^2}{\hat{f}_j^2} + \frac{(f_j - \hat{f}_j)^2}{f_j^2} + \frac{(f_j - \hat{f}_j)^4}{f_j^2 \hat{f}_j^2} \right. \\
&\qquad \left. + \frac{R_j^2}{f_j^2} + \frac{R_j^2 (f_j - \hat{f}_j)^2}{f_j^2 \hat{f}_j^2} + Rest \right],
\end{aligned}
$$

where the term *Rest* includes all 15 intercross terms. For the brevity of exposition, we do not provide all 15 terms explicitly here. We will show that the first term is the leading term and the rest terms are negligible compared to the first main term. Now, our test statistic T_S^* can be written as

$$\begin{aligned} T_S^* &= \max_{1 \le m \le M_n} \frac{1}{\sqrt{8m}} \sum_{j=1}^{m} (\hat{\eta}_j^2 - 1) \\ &= \max_{1 \le m \le M_n} \left\{ \frac{1}{\sqrt{8m}} \sum_{j=1}^{m} (\eta_j^2 - 1) + \frac{1}{\sqrt{8m}} \sum_{j=1}^{m} \eta_j^2 \frac{(f_j - \hat{f}_j)^2}{\hat{f}_j^2} \right. \\ &\quad + \frac{1}{\sqrt{8m}} \sum_{j=1}^{m} \frac{(f_j - \hat{f}_j)^2}{f_j^2} + \frac{1}{\sqrt{8m}} \sum_{j=1}^{m} \frac{(f_j - \hat{f}_j)^4}{f_j^2 \hat{f}_j^2} \\ &\quad \left. + \frac{1}{\sqrt{8m}} \sum_{j=1}^{m} \frac{R_j^2}{f_j^2} + \frac{1}{\sqrt{8m}} \sum_{j=1}^{m} \frac{R_j^2 (f_j - \hat{f}_j)^2}{f_j^2 \hat{f}_j^2} + \frac{1}{\sqrt{8m}} \sum_{j=1}^{m} Rest \right\} \\ &= \max_{1 \le m \le M_n} \left\{ \frac{1}{\sqrt{8m}} \sum_{j=1}^{m} (\eta_j^2 - 1) + I_2 + I_3 + I_4 + I_5 + I_6 + I_7 \right\}. \end{aligned}$$

Let

$$T_{M_n}^* = \max_{1 \le m \le M_n} \frac{1}{\sqrt{8m}} \sum_{j=1}^{m} (\eta_j^2 - 1).$$

Thus, the proof of the Theorem follows from the following two Lemmas and Slutsky's theorem.

Lemma 6.1 *Under the conditions of Theorem 3.1, we have*

$$\begin{aligned} P\big(\sqrt{2 \log \log M_n}\, T_{M_n}^* - \{2 \log \log M_n + .5 \log \log \log M_n \\ - .5 \log (4\pi)\} < x\big) \to \exp(-\exp(-x)). \end{aligned}$$

Lemma 6.2 *Under the conditions of Theorem 3.1, we have*

$$\max_{1 \le m \le M_n} I_i = o_p((\log \log M_n)^{-1/2}),\ i = 2, 3, \ldots, 7.$$

In what follows in this section, we provide the proofs of the above two Lemmas.

Proof of Lemma 6.1: Let's consider statistic $T_{M_n}^{*a} := \max_{2 \le m \le M_n} (8m)^{-1/2} \sum_{j=2}^{m} (\eta_j^2 - 1)$ first. Under the assumption **A1**, by Theorem 10.3.2 of Brockwell and Davis (1991), we have $\eta_1 \sim \chi^2(1) - 1$ and $\eta_j \sim \chi^2(2)/2 - 1$, $j = 2, 3, \ldots, n$ (note those η_j, $j \ge 2$ are i.i.d. random variables). From the observation made by Fan and Huang (2001, p. 651), the maximization of m over $[2, M_n]$ cannot be achieved at $m < \log M_n$ and $T_{M_n}^{*a}$ is at least $(2 \log \log M_n)^{1/2}(1 + o_p(1))$. Now,

since $\eta_1 = O_p(1)$ and $m \geq \log M_n \to \infty$, we derive that $T^{*a}_{M_n}$ has the same limit distribution as that of $T^*_{M_n}$ from Slutsky theorem. In another words, the first term $\eta_1^2 - 1$ is negligible compared to the sum, since $m \geq \log M_n \to \infty$. From the same argument, the above statistic $T^{*a}_{M_n}$ has the same distribution as that of $T^{*b}_{M_n} := \max_{1\leq m\leq M_n}(8m)^{-1/2}\left[\sum_{j=2}^{m}(\eta_j^2-1)+(\eta_{1^*}^2-1)\right]$, where $\eta_{1^*} \sim \chi^2(2)/2-1$. Thus, $T^*_{M_n}$ and $T^{*b}_{M_n}$ have the same limit distributions. (i.e., η_1 can be replaced with η_{1^*}). Notice that the statistic $T^{*b}_{M_n}$ involves m i.i.d. random variables. From simple calculation, we have $E(\eta_j) = 0$, $E(\eta_j^2) = 1$ and $E(\eta_j^2 - 1)^2 = 8$ for $j = 1^*$ and $j \geq 2$. Apply the Theorem 1 of Darling and Erdős (1956) (here η_j have any finite moments), we derive that $T^{*b}_{M_n}$ has the limit distribution stated in Lemma 6.1. Since $T^*_{M_n}$ has the same limit distribution as $T^{*b}_{M_n}$, we complete the proof of the Lemma.

Remark 6.1 In the most frequency analysis of time series, one typically only considers those η_j, $j \geq 2$. However, in our testing hypothesis problem, η_1 provides useful information. This is understandable from the difference between null and alternatives. This observation is confirmed in our simulation studies with finite sample size. We find that the test statistic including the first term η_1 provides a little bit larger power than the test not including η_1. Therefore, we recommend the test statistic T^*_S which includes η_1 in practical implementation, although two test statistics ($T^*_{M_n}$ and $T^{*a}_{M_n}$) have the same limit distribution asymptotically.

In order to prove the the Lemma 6.2, we need following two auxiliary results. The first Proposition is from Kooperberg et al. (1995).

Proposition 6.1 *Under the conditions of Theorem 3.1, we have* $\max_{1\leq j\leq n}|R_j| = O_p(\log n/\sqrt{n})$.

Proposition 6.2 *Under the conditions of Theorem 3.1, we have* $\max_{1\leq j\leq n}\left|1/\hat{f}_j\right| = O_p(1)$.

Proof. Under the conditions of Theorem 3.1, we have $\inf_j f_j \geq c > 0$. Also, for all $M > 0$,

$$
\begin{aligned}
P\left(\max_j \left|\frac{1}{\hat{f}_j}\right| > M\right) &\leq P\left(\max_j \frac{1}{|f_j| - |f_j - \hat{f}_j|} > M\right) \\
&= P\left(\max_j \frac{1}{|f_j| - |f_j - \hat{f}_j|} > M, \max_j |f_j - \hat{f}_j| > \frac{c}{2}\right) \\
&\quad + P\left(\max_j \frac{1}{|f_j| - |f_j - \hat{f}_j|} > M, \max_j |f_j - \hat{f}_j| \leq \frac{c}{2}\right) \\
&\leq P\left(\max_j |f_j - \hat{f}_j| > \frac{c}{2}\right) + P\left(\max_j \frac{1}{|f_j| - c/2}\right. \\
&\qquad \left. > M, \max_j |f_j - \hat{f}_j| \leq \frac{c}{2}\right).
\end{aligned}
$$

Although one can show directly that the first probability in the right hand side (RHS) of the above inequality goes to zero from the explicit expression for $\hat{f}_j$ in Li and Ruppert (2008, p. 420), the details are very tedious. However, one can use the equivalence between penalized spline estimators and kernel estimators to derive the same conclusion. From Li and Ruppert (2008), one can see that $\hat{f}_j$ can be written as a kernel estimator (denoted with $\hat{f}_j^K$) with a Laplace kernel plus a negligible remainder term $O(n^{-2/5})$ which is from the modeling bias and the remainder R_j's in model (16.8). For the kernel estimator $\hat{f}_j^K$, one has a known result $\max_j |f_j - \hat{f}_j^K| = o_p(1)$. Thus, the first probability in RHS of the above inequality goes to zero. (Regarding the equivalence between penalized spline estimators and kernel estimators, see also Silverman (1984) and Wang et al. (2011)). As to the second probability in the RHS of the above inequality, it goes to 0 too from the assumption **A2** with $\inf_j f_j \geq c > 0$ and letting M go to infinity. Thus we prove the Proposition.

Now we are ready to prove the Lemma 6.2.

Proof of the Lemma 6.2. We will prove the Lemma for each value of $i = 2, 3, \ldots, 7$. Consider $i = 2$ first. By Cauchy-Schwartz inequality, we have

$$I_2^2 \leq \frac{1}{8m} \sum_{j=1}^{m} \eta_j^4 \sum_{j=1}^{m} \frac{(f_j - \hat{f}_j)^4}{\hat{f}_j^4}.$$

Under the conditions of Theorem 3.1, those conditions of Theorem 1 in Li and Ruppert (2008) are satisfied. Specifically, assumption **A1** implies that f has continuous second derivative. Also y_j has any finite moments from Gaussian assumptation. Thus, from Theorem 1 in Li and Ruppert (2008), we have $E(f_j - \hat{f}_j)^2 = O(n^{-4/5})$. Again, from the explicit expression for $\hat{f}_j$ in Li and Ruppert (2008, p.420), $f_j - \hat{f}_j$ can be written as a weighted sum of centered η_j's (they are χ^2 random variables) plus a negligible remainder term R_n with $E(R_n^2) = O(n^{-1})$. Therefore it can be shown that $E(f_j - \hat{f}_j)^4 = O(n^{-8/5})$. Thus, $\max_{1\leq m\leq M_n} \sum_{j=1}^{m} (f_j - \hat{f}_j)^4 = O_p(M_n\, n^{-8/5})$. Together with Proposition 6.2, we have $\max_{1\leq m\leq M_n} \sum_{j=1}^{m} (f_j - \hat{f}_j)^4/\hat{f}_j^4 = O_p(M_n\, n^{-8/5})$. From the similar argument as in Fan and Huang (2001, p. 651), we have $\max_{1\leq m\leq M_n} (8m)^{-1} \sum_{j=1}^{m} \eta_j^4 = O_p((\log\log M_n)^{1/2})$. Combining these two results together, we have $\max_{1\leq m\leq M_n} I_2^2 = O_p((\log\log M_n)^{1/2}\, M_n\, n^{-8/5})) = O_p((\log\log n)^{-3/2} n^{-4/5}))$. Thus, $\max_{1\leq m\leq M_n} I_2 = O_p((\log\log n)^{-3/4} n^{-2/5})) = o_p((\log\log n)^{-1/2}) = o_p((\log\log M_n)^{-1/2})$. Thus, we prove the term I_2. For the other terms, arguments are very similar to and simpler than the term I_2. As to $i = 3$, we have $I_3 \leq Cm^{-1/2} \sum_{j=1}^{m} (f_j - \hat{f}_j)^2$. Thus, $\max_{1\leq m\leq M_n} I_3 = O_p\left(M_n^{1/2} n^{-4/5}\right) = O_p\left((\log\log n)^{-1} n^{-2/5}\right) = o_p((\log\log n)^{-1/2}) = o_p((\log\log M_n)^{-1/2})$. For $i = 4$, we have $I_4 \leq Cm^{-1/2} \sum_{j=1}^{m} (f_j - \hat{f}_j)^4$. Thus, $\max_{1\leq m\leq M_n} I_4 = O_p\left(M_n^{1/2} n^{-8/5}\right) = O_p\left((\log\log n)^{-1} n^{-6/5}\right) = o_p((\log\log n)^{-1/2}) = o_p((\log\log M_n)^{-1/2})$. As to term I_5, we have $I_5 \leq Cm^{-1/2} \sum_{j=1}^{m} R_j^2$. From Proposition 6.1, we have $\max_{1\leq m\leq M_n} I_5 = O_p\left(M_n^{1/2} (\log n)^2\, n^{-1}\right) = O_p\left((\log\log n)^{-1} n^{-3/5} (\log n)^2\right) =$

$o_p((\log\log n)^{-1/2}) = o_p((\log\log M_n)^{-1/2})$. For the term I_6, we have

$$I_6 \le \frac{C}{\sqrt{m}} \sqrt{\sum_{j=1}^{m} R_j^4 \sum_{j=1}^{m} (f_j - \hat{f}_j)^4}.$$

Thus,

$$\max_{1 \le m \le M_n} I_6 = O_p\left(M_n^{1/2} \frac{(\log n)^2}{n} n^{-4/5}\right) = O_p\left((\log\log n)^{-1} n^{-7/5} (\log n)^2\right)$$
$$= o_p((\log\log n)^{-1/2}) = o_p((\log\log M_n)^{-1/2}).$$

For the intercross terms in I_7, we only provide one of these terms here for the brevity of the exposition. The proofs for the other terms follow from the similar arguments.

$$\max_{1 \le m \le M_n} \left| \frac{1}{\sqrt{8m}} \sum_{j=1}^{m} \eta_j \frac{f_j - \hat{f}_j}{f_j} \right| \le \max_{1 \le m \le M_n} \sqrt{\frac{1}{8m} \sum_{j=1}^{m} \eta_j^2 \sum_{j=1}^{m} \frac{(f_j - \hat{f}_j)^2}{f_j^2}}$$
$$= \sqrt{O_p((\log\log M_n)^{1/2})\, O_p(M_n n^{-4/5})}$$
$$= O_p\left((\log\log n)^{-3/4}\right) = o_p((\log\log n)^{-1/2})$$
$$= o_p((\log\log M_n)^{-1/2}).$$

Acknowledgements The first author of this chapter would like to express his sincere gratitude to his advisor, Prof. Hira L. Koul for his constant guidance, heart-warming encouragement, and generous support throughout author's doctoral study at Michigan State University and years later. Prof. Koul's extensive knowledge, insight, and dedication to statistics have been author's source of inspiration to work harder and make the best effort possible. Both authors of this chapter are grateful to one referee for his/her very helpful suggestions, which greatly improved the presentation and the content of the chapter.

References

Box GEP, Jenkins GM (1970) Times series analysis. Forecasting and control. Holden-Day, San Francisco, Calif.-London-Amsterdam

Brockwell PJ, Davis RA 1991 Time series: theory and methods. 2nd edn. Springer Series in Statistics. Springer-Verlag, New York

Darling DA, Erdős P (1956) A limit theorem for the maximum of normalized sums of independent random variables. Duke Math J 23:143–155

Eilers PHC, Marx BD. (1996) Flexible smoothing with B-splines and penalties. With comments and a rejoinder by the authors. Statist Sci 11:89–121

Fan J (1996) Test of significance based on wavelet thresholding and Neyman's truncation. J Amer Statist Assoc 91:674–688

Fan J, Huang L (2001) Goodness-of-fit tests for parametric regression models. J Amer Statist Assoc 96:640–652

Fan J, Yao Q (2003) Nonlinear time series: nonparametric and parametric methods. Springer Series in Statistics. Springer-Verlag, New York

Fan J, Zhang W (2004) Generalised likelihood ratio tests for spectral density. Biometrika 91:195–209

Giraitis L, Koul H, Surgailis D (2012) Large sample inference for long memory processes. Imperial College Press, London

Granger CWJ, Joyeux R (1980) An introduction to long-memory time series models and fractional differencing. J Time Ser Anal 1:15–29

Harris D, McCabe B, Leybourne S (2008) Testing for long memory. Econ Theor 24:143–175

Kooperberg C, Stone CJ, Truong YK. (1995) Rate of convergence for logspline spectral density estimation. J Time Ser Anal 16:389–401

Kwiatkowski D, Phillips P, Schmidt P, Shin Y (1992) Testing the null hypothesis of stationarity against the alternative of a unit root: how sure are we that economic time series have a unit root? J Econ 54:159–178

Lee D, Schmidt P (1996) On the power of the KPSS test of stationarity against fractionally-integrated alternatives. J Econ 73:285–302

Li Y, Ruppert D (2008) On the asymptotics of penalized splines. Biometrika 95:415–436

Newey WK, West KD (1987) A simple, positive semi-definite, hetcroskedasticity and autocorrelation consistent covariance matrix. Econometrica 55:703–708

Neyman J (1937). Smooth test for goodness of fit. Skandinavisk Aktuarietiskrift 20:149–199

Robinson PM. (1994). Efficient tests of nonstationary hypotheses. J Amer Statist Assoc 89:142–1437

Ruppert D, Wand MP, Carroll RJ (2003) Semiparametric regression. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, Cambridge

Ruppert D, Wand MP, Carroll RJ. (2009) Semiparametric regression during 2003–2007. Electron J Stat 3:1193–1256

Silverman BW. (1984) Spline smoothing: the equivalent variable kernel method. Ann Statist 12: 898–916

Tanaka K (1999) The nonstationary fractional unit root. Econ Theor 15:549–582

Wang X, Shen J, Ruppert D (2011) On the asymptotics of penalized spline smoothing. Electron J Stat 5:1–17

Chapter 17
On the Computation of R-Estimators

Kanchan Mukherjee and Yuankun Wang

17.1 Introduction

The idea of estimating location parameter based on rank statistics finds its root in the seminal work of Hodges and Lehmann (1963). Since then, a major branch of nonparametric statistics deals with the rank-estimation (R-estimation) of parameters by minimizing certain dispersions or equivalently, solving a system of equations based on the ranks of residual observations. In general, these equations are expressed in terms of linear rank statistics.

Hodges and Lehmann (1963) provided the R-estimator of the center of symmetry in the one-sample location model as well as that of location in the two-sample model and showed that they coincide with the corresponding Wilcoxon estimators. These estimators never have much lower but sometimes infinitely higher efficiencies than the sample mean or the difference of means in the case of two-sample location. Subsequently, Huber (1964) proposed the class of M-estimators and Bickel (1965) and Stigler (1974) discussed L-estimators of location.

The robust methods of estimating the location parameters can be extended naturally to the linear models where we are interested in estimating the unknown regression parameters based on very general score function. There is a vast literature on the rank estimation (R-estimation) of parameters in linear regression models. Major contributions include Adichie (1967), Sen (1969), Jurečková (1971), Koul (1971), Jaeckel (1972), and Heiler and Willers (1988), among others. R-estimators are sometimes preferable to their other competitors for their global robustness and efficiency considerations (classical Chernoff and Savage (1958) phenomenon). For details, see Hájek, Šidák and Sen (1999, Sect. 10.3) Koul (2002, Sect. 4.4) and Jurečková and Sen (1996, Sect. 3.4), among others.

K. Mukherjee (✉) · Y. Wang
Department of Mathematics and Statistics,
Lancaster University, Lancaster LA1 4YF, United Kingdom
e-mail: k.mukherjee@lancaster.ac.uk

Y. Wang
e-mail: y.wang4@lancaster.ac.uk

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_17,

Although R-estimators are useful robust estimators, unfortunately their computation is a challenging and long-standing problem. To date, working computational algorithm for these estimators is not well-developed and consequently their practical applications are restricted. The aim of this note is to propose a simple working algorithm for the computation of R-estimators for linear regression models and to discuss its various applications.

For notable previous attempts on the computation of R-estimators, we mention McKean and Hettmansperger (1978), Terpstra and McKean (2005), and Kloke and McKean (2012). McKean and Hettmansperger (1978) considered one-step R-estimates and illustrated its use in computing Wilcoxon R-estimator based on simulated data. Terpstra and McKean (2005) considered 'Wilcoxon weights' to compute an analogue of the Wilcoxon R-estimator and advocated its use through the R-code `ww`. Kloke and McKean (2012) used the R-function `optim` to compute the rank estimates with general score functions. In our opinion, Kloke and McKean (2012) has been one of the most important contributions in the computation that deals with general score function. Our algorithm of this chapter is in fact simpler than this and is easy to implement.

The rest of the chapter is organized as follows. We describe the algorithm in Sect. 17.2. In Sect. 17.3, this is applied to compute R-estimates of parameters when a simple linear regression model is fitted to a dataset. Similar applications were discussed in the context of multiple linear regression in Sect. 17.4. The concluding section describes applications to other area and plans for future research.

17.2 Algorithm

Consider the usual setup of a linear model where one observes $\{y_i; 1 \leq i \leq n\}$ with regressors $\{\boldsymbol{x}_i; 1 \leq i \leq n\}$ such that

$$y_i = \boldsymbol{x}_i^t \boldsymbol{\beta} + \varepsilon_i$$

or, in matrix notation,

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \varepsilon,$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)$ is the observation vector, $\boldsymbol{X}$ is the known $n \times p$ design matrix of rank p with i-th row $\boldsymbol{x}_i^t$, $\boldsymbol{\beta}$ is the unknown parameter vector and $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)^t$ is the error vector. Let $R_{i\boldsymbol{b}}$ denote the "$\boldsymbol{b}$-residual rank" of the i-th observation, that is

$$R_{i\boldsymbol{b}} = \sum_{j=1}^{n} I(y_j - \boldsymbol{x}_j^t \boldsymbol{b} \leq y_i - \boldsymbol{x}_i^t \boldsymbol{b}).$$

Jurečková (1971) defined the R-estimator of $\boldsymbol{\beta}$ as a solution of the equation

$$S(\boldsymbol{b}) = 0 \tag{17.1}$$

where

$$S(\boldsymbol{b}) = \sum_{i=1}^{n} (\mathbf{x_i} - \bar{\mathbf{x}})\varphi\left(\frac{R_{i\boldsymbol{b}}}{n+1}\right)$$

with $\varphi(u) : (0,1) \to I\!R$ is a score function. We assume that φ is nondecreasing with at most finite number of discontinuities but is not necessarily bounded. Let

$$\bar{\varphi} = \frac{1}{n}\sum_{i=1}^{n} \varphi(\frac{i}{n+1}) = \frac{1}{n}\sum_{i=1}^{n} \varphi\left(\frac{R_{i\boldsymbol{b}}}{n+1}\right).$$

To solve the Eq. (17.1) numerically, we write this as

$$\begin{aligned}
\mathbf{0} = S(b) &= \sum_{i=1}^{n} (x_i - \bar{x})\left\{\varphi\left(\frac{R_{i\boldsymbol{b}}}{n+1}\right) - \bar{\varphi}\right\} \\
&= \sum_{i=1}^{n} x_i \left\{\varphi\left(\frac{R_{i\boldsymbol{b}}}{n+1}\right) - \bar{\varphi}\right\} - \bar{x}\sum_{i=1}^{n}\left\{\varphi\left(\frac{R_{i\boldsymbol{b}}}{n+1}\right) - \bar{\varphi}\right\} \\
&= \sum_{i=1}^{n} x_i \left\{\frac{\varphi\left(\frac{R_{i\boldsymbol{b}}}{n+1}\right) - \bar{\varphi}}{y_i - x_i^t b}\right\}(y_i - x_i^t b) \\
&= \sum_{i=1}^{n} x_i w_i(b)(y_i - x_i^t b)
\end{aligned}$$

where

$$w_i(b) = \frac{\varphi(\frac{R_{i\boldsymbol{b}}}{n+1}) - \bar{\varphi}}{y_i - x_i^t b}, \quad 1 \le i \le n.$$

Let $W(b)$ be a diagonal matrix with (i,i)-th entry $w_i(b)$. Then

$$S(b) = X^t W(b)(Y - Xb).$$

Rearranging $\mathbf{0} = S(b)$, we obtain

$$b = [X^t W(b) X]^{-1}[X^t W(b) y].$$

This yields an iterative procedure

$$b_{i+1} = [X^t W(b_i) X]^{-1}[X^t W(b_i) y], \quad i \ge 0 \tag{17.2}$$

with a starting value $b_0 = [X^t X]^{-1}[X^t y]$, the least squares estimator of $\boldsymbol{\beta}$.

Remark The previous algorithm does not assume any specific form of the score function. It can be mechanically implemented for both bounded and unbounded score functions. If there is an intercept present in the linear regression model, we use R-estimates for the slope parameters and use the median of the R-residuals to estimate the intercept parameter.

We illustrate the use of the previous algorithm in the following sections with some real datasets.

Table 17.1 Number of international calls from Belgium

Year (x_i)	Number of calls (y_i)	Year (x_i)	Number of calls (y_i)
50	0.44	62	1.61
51	0.47	63	2.12
52	0.47	64	11.90
53	0.59	65	12.40
54	0.66	66	14.20
55	0.73	67	15.90
56	0.81	68	18.20
57	0.88	69	21.20
58	1.06	70	4.30
59	1.20	71	2.40
60	1.35	72	2.70
61	1.49	73	2.90

Table 17.2 Simple linear regression functions for the number of international calls from Belgium

Score function	Estimated regression function
Least squares	$\hat{y} = -26.01 + 0.504x$
Wilcoxon	$\hat{y} = -7.185 + 0.146x$
Van der Waerden	$\hat{y} = -6.731 + 0.138x$
Sign	$\hat{y} = -6.757 + 0.138x$

17.3 Simple Linear Regression

The following data (Belgian telephone data), published by the Belgian Statistical Survey, consist of the number of international phone calls $\{y_i\}$ made from Belgium (in tens of millions) over 24 different years $\{x_i\}$, $1 \leq i \leq n = 24$ where $x_1 = 50$ corresponds to the year 1950. The dataset contains some heavy contamination between 1964 and 1969 as a different recording system was used which recorded the total number of *minutes* of calls made rather than simply the numbers. The years 1963 and 1970 were partially affected as well since the transition between the recording systems did not occur on the New Year's day exactly. This dataset was discussed in Rousseeuw and Leroy (1987) also for demostrating the robustness of M-estimators in the linear regression model.

We fitted a simple linear regression model

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i, \quad 1 \leq i \leq n = 24,$$

to this and estimated the parameters using the least squares as well as different R-estimators based on different score functions φ. The estimated regression functions are given in the following Table 17.2.

As can be seen from Figs. 17.1, 17.2 and 17.3, the three R-estimate fits are not affected much by the contaminated observations, while the least squares solution is pulled towards the y values associated with the years 1964–1969. All three regression functions still mostly run through the uncontaminated data points, and provide similar good estimates of the number of calls for the contaminated years.

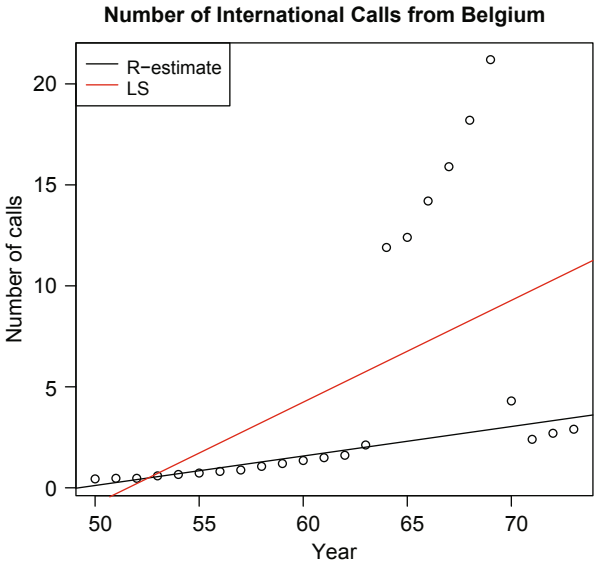


Fig. 17.1 Plot of the regression functions of the international calls based on the Wilcoxon score and least squares estimates

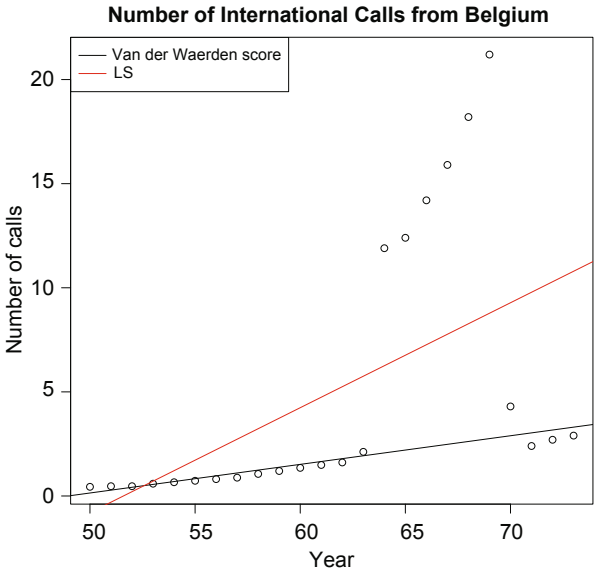


Fig. 17.2 Plot of the regression functions of the international calls based on the van der Waerden score and least squares estimates

Our results are comparable to the findings of Rousseeuw and Leroy (1987) who proposed least median squares estimators of parameters to obtain the regression function $y = -5.610 + 0.115x$.

We have also performed simple linear regression analysis of some standard datasets to demonstrate that R-estimates based on different score functions are similar to the least squares estimates for relatively clean data containing no outlier. The results will be reported elsewhere.

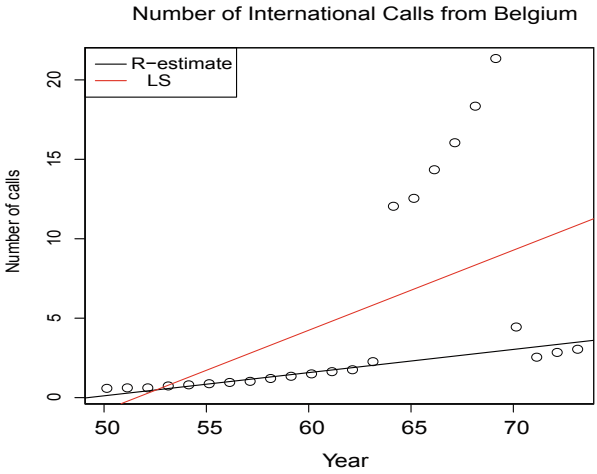


Fig. 17.3 Plot of the regression functions of the international calls based on the sign score and least squares estimates

17.4 Multiple Linear Regression

Next we consider examples of multiple linear regression models. The stackloss dataset has been examined by many methods. The dataset describes the operation of a plant for the oxidation of ammonia to nitric acid and consist of 21 four dimensional observations. The dependent variable stackloss is explained by the rate of operation, the cooling water inlet temperature and the acid concentration.

The least squares regression gives the regression function

$$y = -39.92 + 0.7156x_1 + 1.2953x_2 - 0.1521x_3$$

while the Wilcoxon and the van der Waerden R-estimates yield

$$y = -30.16 + 0.7789x_1 + 0.9853x_2 - 0.2340x_3$$

and

$$y = -37.61 + 0.7735x_1 + 1.0552x_2 - 0.1646x_3$$

respectively. The standardized residual plots for the different fits are shown in Figs. 17.4 and 17.5. The standardization of the residuals is performed based on the division of the raw residuals by the scale estimates corresponding to the fit. From Fig. 17.4 of the standardized least squares residuals, it might appear that the dataset only contains one outlier, observation 21. However, looking at Fig. 17.5 of the standardized R-residuals, we can see that observations 1, 3, and 4 are also outliers. This would not have been picked up if one had only used least squares analysis.

Next we consider an example of the polynomial regression. Table 17.4 shows cloud data, where the dependent variable is the cloud point of a liquid, a measure of the degree of crystallization in a stock. The independent variable is the percentage of I-8 in the base stock. This data was analyzed in Hettmansperger and McKean (2010) who used Wilcoxon estimates of the cubic polynomial regression.

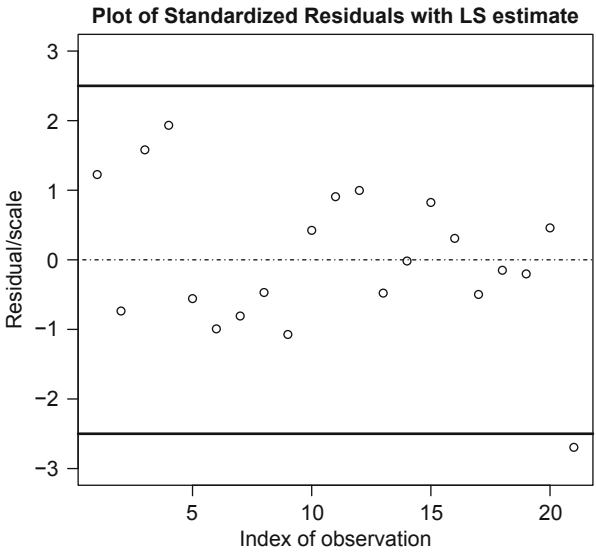


Fig. 17.4 Plot of standardized residuals for the least squares fit

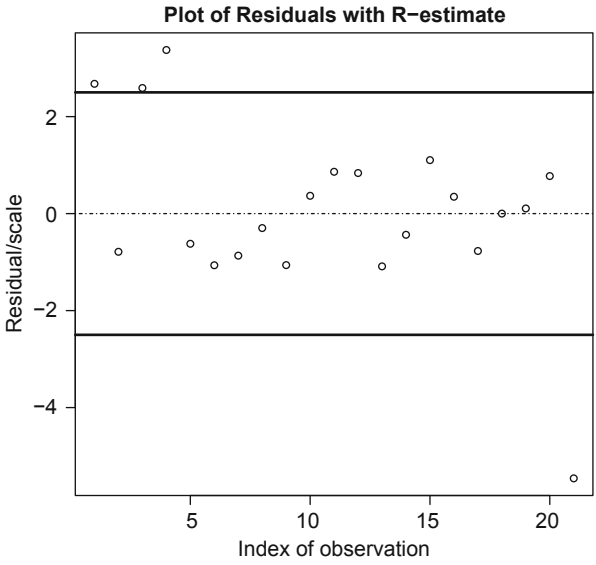


Fig. 17.5 Plot of standardized residuals for the Wilcoxon R-estimate fit

We fitted a linear, a quadratic and a cubic fit to this data and the corresponding Wilcoxon estimates are exhibited in Table 17.5. The rank estimates based on different score functions are quite similar to each other as well as to the least squares estimate. There is slight difference between the Wilcoxon estimates obtained using our algorithm and those reported in Hettmansperger and McKean (2010) where the corresponding estimates are 22.35, 2.24, −0.23, and 0.01. This is due to the use of different algorithms and some rounding off errors.

Table 17.3 Stackloss dataset

Index (i)	Rate (x_1)	Temperature (x_2)	Acid concentration (x_3)	Stackloss (y)
1	80	27	89	42
2	80	27	88	37
3	75	25	90	37
4	62	24	87	28
5	62	22	87	18
6	62	23	87	18
7	62	24	93	19
8	62	24	93	20
9	58	23	87	15
10	58	18	80	14
11	58	18	89	14
12	58	17	88	13
13	58	18	82	11
14	58	19	93	12
15	50	18	89	8
16	50	18	86	7
17	50	19	72	8
18	50	19	79	8
19	50	20	80	9
20	56	20	82	15
21	70	20	91	15

Table 17.4 Cloud data, CP=Cloud Point

% I–8	CP	% I–8	CP
0	22.1	2	26.1
1	24.5	4	28.5
2	26.0	6	30.3
3	26.8	8	31.5
4	28.2	10	33.1
5	28.9	0	22.8
6	30.0	3	27.3
7	30.4	6	29.8
8	31.4	9	31.8
0	21.9		

Table 17.5 Regression functions based on the least squares and different R-estimates for the Cloud data

Score function	Estimated regression function
Least squares	$\hat{y} = 22.31 + 2.22x - 0.22x^2 + 0.01x^3$
Wilcoxon	$\hat{y} = 22.38 + 2.21x - 0.22x^2 + 0.011x^3$
van der Waerden	$\hat{y} = 22.40 + 2.19x - 0.22x^2 + 0.01x^3$
Sign	$\hat{y} = 22.1 + 2.46x - 0.29x^2 + 0.015x^3$

Figure 17.6 exhibits different residual plots of the Wilcoxon estimates based on the linear, quadratic, and cubic fits as well as the normal q-q plot of the cubic fit residuals. This is a small dataset; nevertheless, the q-q plot suggests a slightly heavier tails than the normal distribution for errors and hence the use of the robust R-estimators is justified.

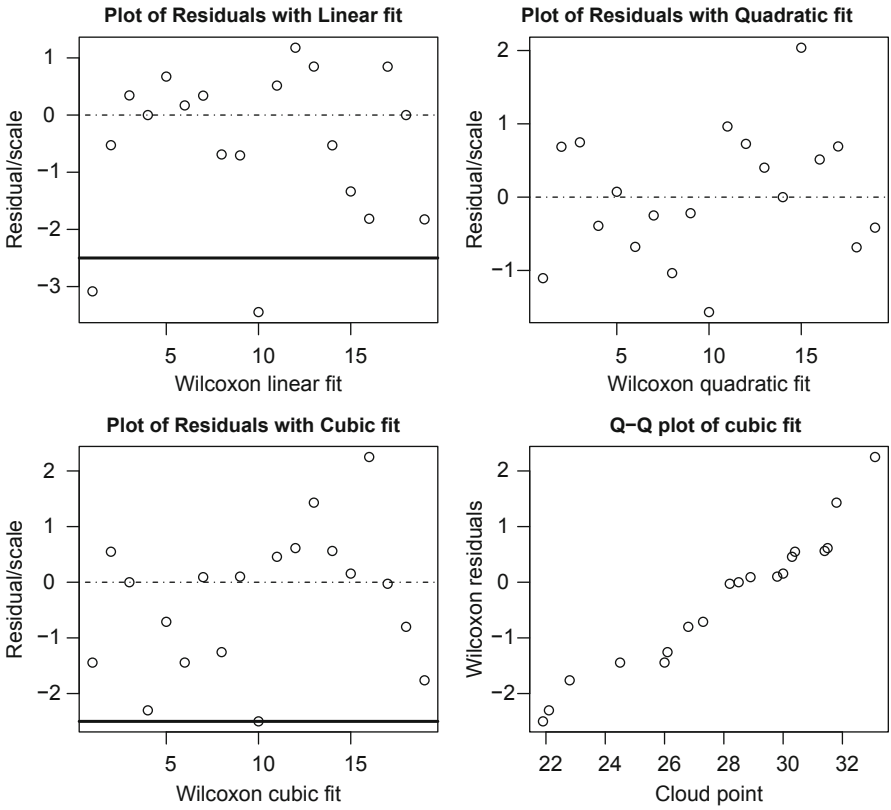


Fig. 17.6 Plot of standardized residuals for the Wilcoxon R-estimate fits

17.5 Conclusion

Computation of R-estimators has been a long-standing problem in the literature. In this chapter, we have proposed an iterative algorithm which can be applied routinely to compute R-estimates based on any score function. This algorithm depends on the form of a mean function that is linear in parameters. Therefore, the algorithm can be applied to compute R-estimators of the linear autoregressive models also and we plan to investigate this further. The algorithm yields convergent sequence of estimates for many well-known datasets considered in this chapter as well as elsewhere. In fact, we applied R-estimators to identify some outliers which would not have been detected using least squares.

Because of its simplicity, the algorithm can be applied to compute R-estimators from bootstrap samples and we plan to pursue this in future. In fact, it opens up various other possibilities and avenues to use R-estimators as one of the most competitive robust class of estimators in various fields of statistics.

Acknowledgement Kanchan Mukherjee would like to thank Prof. Koul for introducing him to the fascinating area of R-estimation and for his encouragement.

References

Adichie J (1967) Estimates of regression parameters based on rank tests Ann Math Statist 38: 894–904

Bickel P (1965) On some robust estimates of location. Ann Math Statist 36:847–858

Chernoff H, Savage I (1958) Asymptotic normality and efficiency of certain nonparametric test statistics. Ann Math Statist 29:972–994

Hájek J Šidák Z, Sen P (1999) Theory of Rank Tests. Academic Press, San Diego

Heiler S, Willers R (1988) Asymptotic normality of R-estimates in the linear model. Statistics 19:173–184

Hettmansperger T, McKean J (2010) Robust Nonparametric Statistical Methods. CRC Press, Boca Raton

Huber P (1964) Robust estimation of a location parameter. Ann Math Statist 35:73–101

Hodges J, Lehmann E (1963) Estimates of location based on rank tests. Ann Math Statist 34:598–611

Jaeckel L (1972) Estimating regression coefficients by minimizing the dispersion of the residuals. Ann Math Statist 43:1449–1458

Jurečková J (1971) Nonparametric estimates of regression coefficients. Ann Math Statist 42: 1328–1338

Jurečková J, Sen P (1996) Robust statistical procedures: asymptotics and interrelations. Wiley, New York

Kloke J, McKean J (2012) Rfit: Rank-based Estimation for Linear Models. The R Journal 4:57–64

Koul H (1971) Asymptotic behavior of a class of confidence region based on rank in regression. Ann Math Statist 42:466–476

Koul H (2002) Weighted Empirical processes in dynamic nonlinear models. Lecture notes in statistics, vol 166. Springer-Verlag, New York

McKean J, Hettmansperger T (1978) A robust analysis of the general linear model based on one step R -estimates. Biometrika 65:571–579

Rousseeuw P, Leroy A (1987) Robust Regression and Outlier Detection. Wiley, New York

Sen P (1969) On a class of rank order tests for the parallelism of several regression lines. Ann Math Statist 40:1668–1683

Stigler S (1974) Linear functions of order statistics with smooth weight functions. Ann Statist 2:676–693

Terpstra J, McKean J (2005) Rank-based analyses of linear models using R. J. Statistical Software (14):7

Chapter 18
Multiple Change-Point Detection in Piecewise Exponential Hazard Regression Models with Long-Term Survivors and Right Censoring

Lianfen Qian and Wei Zhang

18.1 Introduction

Hazard rate is an important function in survival analysis. It quantifies the instantaneous failure rate of a subject (component in reliability analysis) which has not failed at a given time point. In some real life applications, abrupt change in the hazard function is observed due to overhaul, major operation, or specific maintenance activity. This type of change could happen multiple times. In such situations one is interested to detect the locations where such changes occur and to estimate the sizes of the changes if detected.

18.1.1 A Single Change-Point Hazard Model

In the last century, the main stream of the change-point detection in hazard function is for a single change-point piecewise constant hazard model. That is, let T be the failure time and suppose the hazard rate of T is a constant α_1 until a change-point τ, at which point the hazard rate makes a jump and stays another constant thereafter. The problem of interest is to detect the location of the change and to estimate the size if τ exists. To be more precise, the hazard rate of T is

$$\lambda_T(t) = \alpha_1 I(0 \le t < \tau) + \alpha_2 I(t \ge \tau). \tag{18.1}$$

There are mainly three different approaches for inference and estimation under the single change-point model (18.1). They are parametric, semi-parametric, and

L. Qian (✉) · W. Zhang
Florida Atlantic University, 777 Glades Road, Boca Raton, FL 33431, USA
e-mail: lqian@fau.edu

L. Qian
Wenzhou University, Zhejiang 325035, China

W. Zhang
e-mail: drunkway@gmail.com

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_18,

Bayesian approaches. More approaches can be found in the review paper on a single change point in a hazard rate by Anis (2009). The parametric approach is mainly the likelihood ratio based methods considered by Matthews and Farewell (1982); Nguyen et al. (1984); Matthews et al. (1985); Yao (1986); Worsley (1988); Henderson (1990) and Loader (1991).

Matthews and Farewell (1982) first noted the existence of the change-point in the hazard rate when analyzing the failure times of nonlymphoblastic leukemia patients. Nguyen et al. (1984) discussed the unboundedness feature of the likelihood function when the change-point approaches to the maximum observation of the failure times. Matthews et al. (1985) suggested to normalize the score statistic and showed that the asymptotic limiting process of the normalized score process is related to the Ornstein–Uhlenbeck process and the standard Brownian bridge. If $T_{(n-1)}$ is the second largest observation, Yao (1986) suggested to maximize the log-likelihood function in the change-point over $[0, T_{(n-1)}]$ and derived the asymptotic properties of the estimators for both the change-point and the piecewise exponential hazard rates. Worsley (1988) gave the exact critical values of the maximum likelihood estimator over three intervals: i.e. $[0, T_{(n-1)}]$, from pth to $(1-p)$th sample quantiles, and artificially censor the largest observation so that the likelihood function in the change-point is finite. Pham and Nyugen (1990) extended Yao's result to a random compact set. However, it is shown that the result is not better than Yao's. Henderson (1990) noticed that the likelihood ratio test is not sufficient and derived exact critical values for a weighted and standardized likelihood ratio. Loader (1991) derived the approximate confidence regions and joint confidence regions for the change-point and the size of change over another interval.

The semi-parametric approach is studied by Chang et al. (1994), Gijbels and Gürler (2003). This approach is a hybrid martingale based method. The formal one combines the score function with the martingale approach, while the latter one combines the least squared principle with the martingale approach. They assume that the unknown change-point τ belongs to a certain known interval $[0, B]$. Specifically, Chang et al. (1994) constructed the following process

$$X(t) = \left[\frac{\Lambda(B) - \Lambda(t)}{B - t} - \frac{\Lambda(t)}{t}\right] k(t(B-t)), \text{for } 0 < t < B$$

where $k(t) = t^d, 0 \le d \le 1$ and $\Lambda(t)$ is the cumulative hazard rate. The estimator of τ is defined as the smallest maximizer of $X_n(t)$, which is defined as the process when $\Lambda(t)$ in $X(t)$ is replaced by its Nelson–Aalen estimator. They obtain the consistency and the limiting distribution of the estimator of τ using a martingale inequality and Poisson approximation.

On the other hand, Gijbels and Gürler (2003) considered the following process

$$Y(t) = \frac{\Lambda(t)}{t}.$$

Once again replacing Λ by its Nelson–Aalen estimator to obtain the empirical hazard rate process $Y_n(t)$. Then the splitting point which gives the best least square fit

between $Y_n(t)$ and $Y(t)$ over a set of pre-chosen grid points is defined as the estimator of τ.

The last method is Bayesian approach. Achcar and Bolfarine (1998) first examined the Bayes estimator assuming a discrete random change-point. Achcar and Loibel (1998) extended the method to noninformative reference priors. No asymptotic results are obtained. Ebrahimi et al. (1997) proposed a Bayes estimator avoiding asymptotics to provide a more reliable inference conditional only upon the data actually observed. Xu and Qian (2013) discussed the method to accommodate censored data in the presence of both covariates and long-term survivors.

18.1.2 Multiple Change-Points Hazard Model with Covariates

Assuming the existence of the change-points in a hazard function, Pons (2003) considered estimation in a Cox regression model with a change-point according to a threshold in a covariate; Dupuy (2006) studied the estimation in a change-point right-censored exponential model with covariates. Goodman et al. (2011) proposed a Wald-type test statistic for multiple change-points in a piecewise linear hazard model incorporating covariates. They assumed that the changes only affect parameters of the baseline hazard function. Properties of the suggested test and estimators of the change-points are investigated via simulations, but no theoretical results are available. Dupuy (2009) studied the likelihood ratio type test for the existence of a change in both baseline hazard rate and covariate effects in an exponential regression model with right-censoring. Non-asymptotic bounds for the type II error probability are obtained. Li et al. (2013) extended the work to include long-term survivors, but all, except Goodman et al. (2011), of above literatures still assumed single change-point.

In this chapter, we are interested in multiple change-points detection problem. To be more precise, suppose that the failure times of the subjects under study in medical applications (or components in reliability analysis) are independent. Let T^* be the failure time of a subject with hazard rate λ^* and Z be the covariate vector. We consider a piecewise exponential hazard model with multiple change-points both in the baseline hazard and in the covariates as follows:

$$\lambda^*(t|Z) = \sum_{i=1}^{q+1} \alpha_i e^{\beta_i' Z} I(\tau_{i-1} \le t < \tau_i), \tag{18.2}$$

where q is the number of change-points, $0 < \tau_1 < \tau_2 < \ldots < \tau_q < \infty$ are the change-points, $\tau_0 = 0$ and $\tau_{q+1} = \infty$. For $i = 1, \ldots, q+1$, denote $I_i(t) = I(\tau_{i-1} \le t < \tau_i)$ and

$$c_i(t|Z) = \beta_i' Z - \alpha_i (t - \tau_{i-1}) e^{\beta_i' Z} - \sum_{j=1}^{i-1} a_j (\tau_j - \tau_{j-1}) e^{\beta_j' Z}.$$

Then, the corresponding density and survival functions of model (18.2) are:

$$f^*(t|Z) = \sum_{i=1}^{q+1} \alpha_i e^{c_i(t|Z)} I_i(t) \text{ and } S^*(t|Z) = \sum_{i=1}^{q+1} \exp\left\{c_i(t|Z) - \beta_i' Z\right\} I_i(t).$$

Let $\tau = (\tau_1, \ldots, \tau_q)'$ and $\theta' = (\alpha_1, \ldots, \alpha_{q+1}, \beta_1', \ldots, \beta_{q+1}')$. For $i = 1, 2, 3, \ldots, q+1$, denote $b_i(Z) = -\alpha_i \exp(\beta_i' Z)$. Then, for given (q, τ'), the cumulative hazard function is

$$\Lambda^*(t|Z) \equiv \Lambda^*(t, \theta|Z) = \sum_{i=1}^{q+1} \left[b_i(Z)(t - \tau_{i-1}) + \sum_{j=1}^{i-1} b_j(Z)(\tau_j - \tau_{j-1}) \right] I_i(t). \tag{18.3}$$

Therefore, $\Lambda^*(t, \theta|Z)$ is a piecewise linear function in t for a given Z.

18.1.3 Hazard with a Change Point in the Presence of Long-Term Survivors

On the other hand, it has been noticed that survival data often show long-term survivors. For example, Matthews and Farewell (1982) noticed that the Kaplan–Meier estimator of the failure time distribution function of leukemia patients levels off significantly below one, which indicates the presence of long-term survivors. Taylor (1995) observed that only some of the patients with tumors in head and neck will experience local recurrences after radiation therapy. The remaining patients will not have recurrences because all of the tumor cells would have been killed by the radiation. For survival data in a mixture model with long-term survivors, there are two types of individuals: susceptibles and long-term survivors. The susceptibles are at risk of developing the event under consideration, and the event would be observed with certainty if complete follow-up were possible. The long-term survivors will never experience the event.

The existence of long-term survivors leads to estimates of both the probability of being a susceptible and the failure time distribution for susceptibles. Though survival models with long-term survivors have been studied for decades and many applications have been reported in Maller and Zhou (1996), these models have not been considered for possible change-point phenomena. In reality, however, change-point may well exist in a survival data with long-term survivors. Zhao et al. (2009) studied parameter estimation for a piecewise constant hazard model with one change-point in the presence of long-term survivors via the non-parametric Nelson–Aalen estimator. They didn't detail out the case when covariate effects exist. Li et al. (2013) derived similar results via maximum likelihood estimation in the presence of both covariates and long-term survivors for right censored failure time data. However, all those results are limited to one change-point hazard model. While earlier studies focus on the single change-point hazard model, the attention has shifted to the

multiple change-points in recent years. Goodman et al. (2011) studied the multiple change-points detection problem without considering long-term survivors. Zhang et al. (2013) proposed a hybrid sequential martingale based maximum likelihood approach to detect multiple change-points for a piecewise constant hazard model in the presence of long-term survivors. However, Zhang et al. (2013) did not include possible covariate effects.

In this chapter, we consider the multiple change-points detection for hazard rates in the presence of both covariates and long-term survivors for right censored failure times. The rest of this chapter is structured as follows. Section 18.2 introduces the piecewise constant hazard regression model with multiple change-points. Section 18.3 proposes a multiple change-points detection algorithm and gives estimators based on the Nelson–Aalen estimator and weighted least squares principle. Section 18.4 reports the finite sample performance, sensitivity, and reliability analyses of the proposed method through a simulation study. We apply the proposed method to test for change-points in the hazard rates of prostate cancer patients in Sect. 18.5. We conclude our results in Sect. 18.6.

18.2 Multiple Change-Points Hazard Regression Model with Long-Term Survivors

In this section, we introduce a piecewise constant hazard model with multiple change-points. In the presence of long-term survivors and covariates, we notice that the susceptible proportion may depend on the covariates. Let $p(Z)$ be the probability of being susceptible subjects and $1-p(Z)$ is the probability being long-term survivors for a given covariate Z. Let T and T^* be the failure times of each and a susceptible subject with survival functions S and S^*, respectively. Then, for $t < \infty$, the survival function of T is

$$S(t|Z) = P(T > t|Z) = 1 - p(Z) + p(Z)P(T^* > t|Z) = 1 - p(Z) + p(Z)S^*(t|Z)$$

with the hazard function

$$\lambda(t|Z) = \frac{-S'(t|Z)}{S(t|Z)} = \frac{-p(Z)(S^*(t|Z))'}{S(t|Z)} = \frac{p(Z)f^*(t|Z)}{1 - p(Z) + p(Z)S^*(t|Z)}. \quad (18.4)$$

In this chapter, we assume that the susceptible hazard rate of T^* satisfies Eq. (18.2). Then the hazard function of T for a given Z is

$$\lambda(t|Z) = \sum_{i=1}^{q+1} \frac{p(Z)\alpha_i e^{c_i(t|Z)}}{1 - p(Z) + p(Z)e^{c_i(t|Z) - \beta_i' Z}} I_i(t)$$

and the corresponding cumulative hazard function is

$$\Lambda(t|Z) = -\sum_{i=1}^{q+1} \log \left[1 - p(Z) + p(Z) \exp \left\{ -b_i(t - \tau_{i-1}) - \sum_{j=1}^{i-1} b_j(\tau_j - \tau_{j-1}) \right\} \right] I_i(t).$$

Combining with (18.3), one implies the following relationship

$$\Lambda^*(t|Z) = -\log\left\{\frac{1}{p(Z)}\left[-1 + p(Z) + e^{-\Lambda(t|Z)}\right]\right\}. \tag{18.5}$$

18.3 Change-Points Detection and Estimation

Let C be a censoring variable. Under non-informative censoring, we observe (X_i, δ_i, Z_i), $i = 1, \ldots, n$, where $X_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$. We propose a semi-parametric detection algorithm for change-points including the number of change-points. As in Maller and Zhou (1996), we first define the estimator of the susceptible proportion $p(Z)$ for each given $Z = z$ as:

$$\widehat{p}_z = \widehat{F}_n(X_{(n)}|Z = z),$$

where $X_{(n)} = \max_{\delta_i=1, i=1,\ldots,n} X_i$ is the largest uncensored observation and $\widehat{F}_n(\cdot\,|Z)$ denotes the Kaplan–Meier estimator of $F(\cdot|Z) = 1 - S(\cdot|Z)$.

18.3.1 A Weighted Average Hazard Process

We construct a weighted average hazard function in time t for a given Z:

$$Y^*(t|Z) = \frac{\Lambda^*(t|Z)}{t}k(t),$$

where $k(t) = t^d, 0 \leq d \leq 1$. For $d = 0$, Eq. (18.3) implies that

$$Y^*(t|Z) = \sum_{i=1}^{q+1}\left[b_i(Z) + \frac{\sum_{j=1}^{i-1} b_j(Z)(\tau_j - \tau_{j-1}) - b_i(Z)\tau_{i-1}}{t}\right] I_i(t).$$

This function has a simple structure: it remains constant up to time τ_1 and thereafter changes as a function of t^{-1} before τ_2, and another function of t^{-1} after τ_2 and so on. So the estimating procedure now consists of fitting a constant line up to τ_1, and from τ_1 to τ_2 a function of the second form in above equation and from τ_2 on another function of t^{-1} of the third form and so on.

Gijbels and Gürler (2003) studied a single change-point estimation using $d = 0$. However, when $d = 0$, a dramatic boundary effect near 0 is present due to the denominator, hence the fraction $Y^*(t|Z)$ is unstable. This brings some difficulty in estimation if we choose the first few grid points relatively small. To overcome this difficulty, we introduce the weight function k for an appropriate non-zero $d > 0$ to adjust the boundary effect. Simulation results show the adjustment is effective.

18.3.2 Least Squared Approach

Let $\Lambda_n^*(t|Z)$ be the Nelson-Aalen estimator of $\Lambda^*(t|Z)$. Denote the empirical version $Y_n^*(t|Z)$ of $Y^*(t|Z)$ by replacing $\Lambda^*(t|Z)$ with the Nelson-Aalen estimator $\Lambda_n^*(t|Z)$. Let G_Z be the cumulative distribution function of Z. We assume that all potential change-points lie in a certain known interval $[B_1, B_2]$. We choose grid points $\{t_g\}$ such that $B_1 \le t_1 < t_2 < \ldots < t_g \le B_2$, where g is the number of grid points greater than the true number of change-points.

For each integer $q \in \{1, 2, \ldots, g\}$ and a given subset $\{i_1, i_2, \ldots, i_q\}$ such that $0 = t_{i_0} < t_1 \le t_{i_1} < t_{i_2} < \ldots < t_{i_q} \le t_g$, we carry out least square fit by minimizing the following object function first:

$$
\begin{aligned}
&S(\theta|i_1, i_2, \ldots, i_q) \\
&= \int \left\{ \sum_{j=1}^{i_1} \left[Y_n^*(t_j|z) - b_1(z) \right]^2 + \sum_{i_1+1}^{i_2} \left[Y_n^*(t_j|z) - b_2(z) - \{b_1(z) - b_2(z)\} \frac{t_{i_1}}{t_j} \right]^2 \right. \\
&\quad + \cdots + \sum_{i_k+1}^{i_{k+1}} \left[Y_n^*(t_j|z) - b_{k+1}(z) + b_{k+1}(z) \frac{t_{i_k}}{t_j} - \sum_{m=1}^{k} b_m(z) \frac{t_{i_m} - t_{i_{m-1}}}{t_j} \right]^2 \\
&\quad \left. + \cdots + \sum_{i_q+1}^{g} \left[Y_n^*(t_j|z) - b_{q+1}(z) + b_{q+1}(z) \frac{t_{i_q}}{t_j} - \sum_{m=1}^{q} b_m(z) \frac{t_{i_m} - t_{i_{m-1}}}{t_j} \right]^2 \right\} dG_Z(z).
\end{aligned}
$$

Now we are ready to describe the proposed change-points detection algorithm. This algorithm includes four steps:

Step 1. For a given integer $q \in \{1, \ldots, g\}$ and indices $i_1, i_2, \ldots, i_q$ such that $1 \le t_1 \le t_{i_1} \le t_{i_2} \le \ldots \le t_{i_q} \le g$, we define $\widehat{\theta}(i_1, i_2, \ldots, i_q) = \arg\min_\theta S(\theta|i_1, i_2, \ldots, i_q)$. Substitute $\widehat{\theta}(i_1, i_2, \ldots, i_q)$ into $S(\theta|i_1, i_2, \ldots, i_q)$ to obtain

$$\tilde{S}(i_1, \ldots, i_q) = S(\widehat{\theta}(i_1, i_2, \ldots, i_q)|i_1, \ldots, i_q).$$

Step 2. Minimize $\tilde{S}(i_1, \ldots, i_q)$ over all possible combinations of q chosen indices $1 \le i_1 \le i_2 \le \cdots \le i_q \le g$. Denote

$$
\begin{aligned}
(\tilde{i_1}, \tilde{i_2}, \ldots, \tilde{i_q}) &= argmin_{(1 \le i_1 \le i_2 \le \ldots \le i_q \le g)} \tilde{S}(i_1, \ldots, i_q) \text{ and } \widehat{S}(q) \\
&= \tilde{S}(\tilde{i_1}, \tilde{i_2}, \ldots, \tilde{i_q}).
\end{aligned}
$$

Step 3. Define $\tilde{q} = argmin_{1 \le q \le g} \widehat{S}(q)$.
Use $\chi^2_{s+1} = -2\left[\widehat{S}(\tilde{q}) - \widehat{S}(\tilde{q} - 1)\right]$ to test $H_{0q} : q = \tilde{q} - 1$ versus $H_{1q} : q = \tilde{q}$, where s is the dimension of Z. If the null hypothesis is not rejected, we continue the hypothesis test for one less change-points until to reach rejection of null hypothesis. If we can not reject the null hypothesis for $\tilde{q} = 1$, we conclude there is no change-point. Otherwise, we choose the

smallest $\tilde{q} \geq 1$ which leads to the rejection as the estimator of the true number of change-points, denoted by $\hat{q}$.

Step 4. If there is no change-point, the estimation of parameters are well worked out. If $\hat{q} \geq 1$, we define our estimators as $\widehat{\tau}' = (t_{\tilde{i}_1}, \ldots, t_{\tilde{i}_{\hat{q}}})$ and $\widehat{\theta} = \widehat{\theta}(\tilde{i}_1, \ldots, \tilde{i}_{\hat{q}})$.

18.4 A Simulation Study

In this section, we conduct a simulation study to examine the finite sample performance, sensitivity, and reliability analyses for our algorithm. To examine the effect of the model misspecification, we generate data from three models: model 1 has no change-point; model 2 has one change-point; and model 3 has two change-points. The sample sizes range from 50 to 500 with 1000 replications. For each given sample size n, the covariate sample $z = \{z_1, \ldots, z_n\}$ is generated from a Bernoulli random variable with the success probability 0.5. This is of practical meaning since we are often dealing with treatment effect for comparable studies, so we consider 0 as treatment (group) 1 and 1 as treatment (group) 2. Denote $n_0 = \sum_{i=1}^{n} I(z_i = 0)$ and $n_1 = \sum_{i=1}^{n} I(z_i = 1)$. The susceptible proportion takes two values: $p(0)$ and $p(1)$. In our simulation, we consider $(p(0), p(1)) = (0.8, 0.8), (0.8, 0.9), (0.9, 0.8)$ and $(0.9, 0.9)$. Denote $n_s = [n_0 p(0) + n_1 p(1)]$, the number of susceptible subjects. For each model, we use two parameter settings. That is:

For model 1 with no change-point, we set

$$\theta' = (\alpha_1, \beta_1) = (0.50, -0.40) \text{ and } (0.80, -0.50);$$

For model 2 with one change-point, we set $\tau_1 = 1$,

$$\theta' = (\alpha_1, \alpha_2, \beta_1, \beta_2) = (0.80, 0.20, -0.15, -0.70) \text{ and } (0.60, 0.10, -0.40, 0.70).$$

For model 3 with two change-points, we set $\tau_1 = 1$ and $\tau_2 = 2$,

$$\begin{aligned} \theta' \quad &= \quad (\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3) = (0.90, 0.30, 0.10, -0.10, -0.40, -0.70) \\ &\text{and} \quad (0.60, 0.10, 0.30, 0.40, 0.70, 0.50). \end{aligned}$$

We generate the failure time samples with right censoring and long-term survivors using the following algorithm for model 3. For models 1 and 2, set $\tau_1 = \tau_2 = \infty$ for model 1 and $\tau_2 = \infty$ for model 2, respectively.

1. Generate a simple random sample $u_1, \ldots u_{n_s}$ from uniform distribution $U(0, 1)$.
2. Denote $J_1(u|z) = I\left(-\alpha_1 \tau_1 e^{\beta_1 z} \leq \ln u < 0\right)$,

$$J_2(u|z) = I\left(-\alpha_1 \tau_1 e^{\beta_1 z} - \alpha_2(\tau_2 - \tau_1) e^{\beta_2 z} \leq \ln u < -\alpha_1 \tau_1 e^{\beta_1 z}\right)$$

and

$$J_3(u|z) = I\left(\ln u < -\alpha_1 \tau_1 e^{\beta_1 z} - \alpha_2(\tau_2 - \tau_1) e^{\beta_2 z}\right).$$

Generate the failure time sample $t_1, \ldots, t_{n_s}$ by the inverse function of $S^*(t_i|z_i)$. That is:

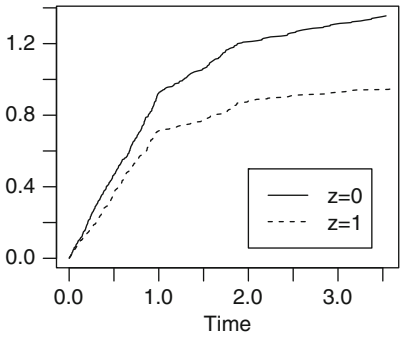


Fig. 18.1 The Nelson–Aalen estimator of $\Lambda_n^*(t|z)$

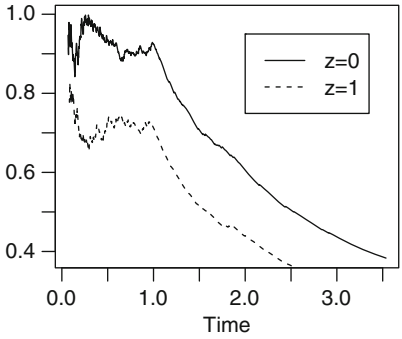


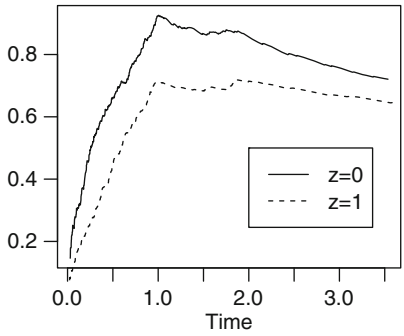
Fig. 18.2 $Y_n^*(t)$ when $d = 0$

$$t_i = -\left[\frac{\ln u_i}{\alpha_1 e^{\beta_1 z_i}}\right] J(u_i|z_i) - \left[\frac{\ln u_i + \alpha_1\tau_1 e^{\beta_1 z_i}}{\alpha_2 e^{\beta_2 z_i}} + \tau_1\right] J_2(u_i|z_i)$$
$$- \left[\frac{\ln u_i + \alpha_1\tau_1 e^{\beta_1 z_i} + \alpha_2(\tau_2 - \tau_1)e^{\beta_2 z_i}}{\alpha_3 e^{\beta_3 z_i}} + \tau_2\right] J_3(u_i|z_i).$$

3. Notice that the long-term survivor proportion is $n - n_s$. So we add $n - n_s$ number of the largest observation of $\{t_1, \ldots, t_{n_s}\}$ as the long-term survivor observations.
4. Generate the censoring sample $c_1, \ldots, c_n$ from uniform distribution from 0 to the largest failure time.

We illustrate the estimated cumulative hazard for data from two change-points model with $\theta' = (\alpha_1, \alpha_2, \alpha_3, \beta_1, \beta_2, \beta_3) = (0.90, 0.30, 0.10, -0.10, -0.40, -0.70)$, $p(0) = 0.9$, $p(1) = 0.9$, $\tau_1 = 1$, $\tau_2 = 2$ and $n = 200$. Figure 18.1 shows the graph of $\Lambda_n^*(t|z)$, while Figs. 18.2 and 18.3 show the plots of $Y_n^*(t|z)$ when setting $d = 0$ and $d = 0.5$, respectively. Compared with $\Lambda_n^*(t|z)$, the function $Y_n^*(t|z)$ magnifies the changes in shape before and after the change-points. However, under our simulation setting, there exists some boundary effect when $d = 0$. To adjust the boundary effect, we introduce the weight function $k(x) = x^d$. It is observed that by setting $d = 0.5$, the boundary effect is adjusted and the changes of shapes are magnified even more as desired. By using our algorithm, we found $\hat{q} = 2$ and numerical results are given in Table 18.3. From Table 18.3, one observes that, as the sample size increases, our estimates shrink to the true parameters with small standard error and almost ignorable miscalssification error for both parameter settings. Tables 18.1–18.3

Fig. 18.3 $Y_n^*(t)$ when $d = 0.5$



report the simulation results on the sample mean (Mean), the sample standard error (Se), and the over-fitting error (Oe) for the two parameter settings (Set 1 and Set 2) for data generated from the models 1–3, respectively. Our algorithm allows model misspecification.

Table 18.3 shows that our algorithm works well if the susceptible proportion depends on the covariate. Figures 18.4–18.6 show comparison between the true $\lambda^*(t|z)$ and its estimate based on one realization for the parameter settings (Set 1) for sample size 500. From these three figures, one observes that our estimates of change-points and other parameters are very accurate.

18.5 Real Data Analysis

We consider data from a retrospective study of 45 women who had surgery for breast cancer. Tumor cells, surgically removed from each woman, were classified according to the results of staining on a marker taken from the Roman snail, the Helix pomatia agglutinin (HPA). The marker binds to cancer cells associated with metastasis to nearby lymph nodes. Upon microscopic examination, the cancer cells stained with HPA are classified as positive, corresponding to a tumor with the potential for metastasis. Otherwise is negative. Eight individuals in the negative stained group, and eleven in the positive stained group are censored. It is of interest to determine the relationship of HPA staining and the survival of women with breast cancer. The survival times for the positive stained group are:

5; 8; 10; 13; 18; 24; 26; 26; 31; 35; 40; 41; 48; 50; 59; 61; 68; 71; 76+; 105+; 107+; 109+; 113; 116+; 118; 143; 154+; 162+; 188+; 212+; 217+; 225+;

where + denotes a right-censored observation. The survival times for the negative stained group are:

23; 47; 69; 70+; 71+; 100+; 101+; 148; 181; 198+; 208+; 212+; 224+.

First of all, we estimate the susceptible proportion. Here the stain group is the covariate. Using the KM-estimator, we have $p(1) = 0.4609$ and $p(2) = 0.6906$, where $p(1)$ and $p(2)$ are the susceptible proportions corresponding to the negative and positive stained group respectively. Figure 18.7 shows the graph of Nelson-Aalen estimator $\Lambda_n^*(t|Z)$. From the relation of $\Lambda_n(t|Z)$ and $\Lambda_n^*(t|Z)$, we can graph the figure of $\Lambda_n(t|Z)$, as shown in Fig. 18.8. Figure 18.7 indicates the existence of

Table 18.1 The summary statistics for data generating from no change-point model

Sample size	Parameter	Set 1	Mean	Se	Oe	Set 2	Mean	Se	Oe
50	p	0.90	0.8714	0.0188	0.026	0.80	0.7759	0.0205	0.018
	α	0.50	0.5020	0.0785		0.80	0.7954	0.0689	
	β	−0.40	−0.3859	0.1156		−0.50	−0.4952	0.1659	
100	p	0.90	0.8899	0.009	0.025	0.80	0.7859	0.0169	0.014
	α	0.50	0.4995	0.0568		0.80	0.8025	0.0468	
	β	−0.40	−0.3864	0.1454		−0.50	−0.5145	0.1382	
200	p	0.90	0.8944	0.006	0.015	0.80	0.7921	0.0085	0.011
	α	0.50	0.4974	0.0374		0.80	0.7986	0.0268	
	β	−0.40	−0.4166	0.1254		−0.50	−0.5086	0.0964	
500	p	0.90	0.8971	0.001	0.009	0.80	0.7964	0.0033	0.006
	α	0.50	0.4989	0.0259		0.80	0.7994	0.0198	
	β	−0.40	−0.4059	0.0062		−0.50	−0.5065	0.0073	

Table 18.2 The summary statistics for data generating from one change-point model

Sample size	Parameter	Set 1	Mean	Se	Oe	Set 2	Mean	Se	Oe
50	p	0.80	0.7806	0.0078	0.031	0.90	0.8924	0.0042	0.036
	τ	1	0.9930	0.2527		1	1.0235	0.1825	
	α_1	0.80	0.8180	0.1419		0.60	0.5684	0.1008	
	α_2	0.20	0.2093	0.0859		0.10	0.1105	0.0457	
	β_1	−0.15	−0.1259	0.2598		−0.40	−0.3689	0.1986	
	β_2	−0.70	−0.6784	0.1956		0.70	0.7126	0.2564	
100	p	0.80	0.7910	0.004	0.026	0.90	0.8936	0.0035	0.018
	τ	1	1.0026	0.2026		1	0.9568	0.1865	
	α_1	0.80	0.8203	0.1103		0.60	0.5875	0.0849	
	α_2	0.20	0.2019	0.0387		0.10	0.0986	0.0356	
	β_1	−0.15	−0.1359	0.2105		−0.40	−0.3853	0.1854	
	β_2	−0.70	−0.6825	0.1882		0.70	0.6958	0.1987	
200	p	0.80	0.7935	0.002	0.019	0.90	0.8964	0.0028	0.012
	τ	1	0.9985	0.1548		1	0.9689	0.1321	
	α_1	0.80	0.8086	0.0569		0.60	0.5964	0.0549	
	α_2	0.20	0.2011	0.0256		0.10	0.1105	0.0259	
	β_1	−0.15	−0.1398	0.1758		−0.40	−0.3868	0.1221	
	β_2	−0.70	−0.6882	0.1569		0.70	0.7052	0.1524	
500	p	0.80	0.7973	0.0008	0.009	0.90	0.8995	0.0010	0.005
	τ	1	1.0011	0.0956		1	0.9865	0.1102	
	α_1	0.80	0.8009	0.0368		0.60	0.5987	0.0465	
	α_2	0.20	0.2007	0.0095		0.10	0.1028	0.0208	
	β_1	−0.15	−0.1425	0.1185		−0.40	-0.3964	0.0952	
	β_2	−0.70	−0.6912	0.0906		0.70	0.7014	0.0684	

more than one change-points and piecewise linear property in $\Lambda_n^*(t|Z)$. Hence, our algorithm is suitable to model this data. Using our algorithm, we found $\hat{q} = 2$. The estimation for the rest parameters are:

$$\begin{aligned} & (\hat{\tau}_1, \hat{\tau}_2, \hat{\alpha}_1, \hat{\alpha}_2, \hat{\alpha}_3, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3) \\ = \; & (71, 105, 0.009, 0.0001, 0.0171, 0.6931, 3.5556, 1.2730). \end{aligned}$$

Table 18.3 The summary statistics for data generating from two change-points model

Sample size	Parameter	Set 1	Mean	Se	Oe	Set 2	Mean	Se	Oe
50	p(0)	0.80	0.7928	0.0065	0.011	0.90	0.8962	0.0031	0.013
	p(1)	0.80	0.7928	0.0065		0.80	0.7964	0.0048	
	τ_1	1	0.8864	0.2008		1	0.9548	0.2159	
	τ_2	2	2.1058	0.3548		2	2.0345	0.2015	
	α_1	0.90	0.9478	0.0756		0.60	0.5725	0.0854	
	α_2	0.30	0.3517	0.1622		0.10	0.0825	0.0454	
	α_3	0.10	0.0966	0.0278		0.30	0.3191	0.1025	
	β_1	−0.10	−0.0828	0.0598		0.40	0.3462	0.0846	
	β_2	−0.40	−0.3857	0.2105		0.70	0.7251	0.1964	
	β_3	−0.70	−0.6748	0.2958		0.50	0.4751	0.1326	
100	p(0)	0.80	0.7954	0.0058	0.008	0.90	0.8938	0.0036	0.006
	p(1)	0.80	0.7954	0.0058		0.80	0.7971	0.0038	
	τ_1	1	0.9124	0.1954		1	0.9452	0.1764	
	τ_2	2	2.0824	0.2759		2	1.9624	0.2185	
	α_1	0.90	0.9328	0.0648		0.60	0.5824	0.0576	
	α_2	0.30	0.3324	0.1198		0.10	0.0862	0.0368	
	α_3	0.10	0.0948	0.0195		0.30	0.2915	0.0675	
	β_1	−0.10	−0.0915	0.0359		0.40	0.3654	0.1124	
	β_2	−0.40	−0.3912	0.1541		0.70	0.7104	0.1346	
	β_3	−0.70	−0.7059	0.1654		0.50	0.4822	0.1124	
200	p(0)	0.80	0.7969	0.0043	0.005	0.90	0.8954	0.0038	0.004
	p(1)	0.80	0.7969	0.0043		0.80	0.7974	0.0036	
	τ_1	1	0.9459	0.1548		1	0.9457	0.1344	
	τ_2	2	2.0359	0.1258		2	2.0247	0.1547	
	α_1	0.90	0.9254	0.0569		0.60	0.5871	0.0476	
	α_2	0.30	0.3259	0.0954		0.10	0.1085	0.0249	
	α_3	0.10	0.0971	0.0154		0.30	0.2964	0.0568	
	β_1	−0.10	−0.0955	0.0295		0.40	0.3794	0.0956	
	β_2	−0.40	−0.3964	0.1056		0.70	0.7105	0.0906	
	β_3	−0.70	−0.6954	0.1259		0.50	0.4862	0.0854	
500	p(0)	0.80	0.7985	0.0028	0.001	0.90	0.8975	0.0027	0.003
	p(1)	0.80	0.7985	0.0028		0.80	0.7982	0.0021	
	τ_1	1	0.9851	0.0684		1	0.9907	0.0615	
	τ_2	2	2.0124	0.0589		2	2.0217	0.0708	
	α_1	0.90	0.9117	0.0359		0.60	0.5938	0.0412	
	α_2	0.30	0.2958	0.0548		0.10	0.0957	0.0217	
	α_3	0.10	0.0974	0.0085		0.30	0.2918	0.0386	
	β_1	−0.10	−0.1059	0.0219		0.40	0.3815	0.0518	
	β_2	−0.40	−0.4021	0.0658		0.70	0.7077	0.0734	
	β_3	−0.70	−0.7021	0.0705		0.50	0.4938	0.0615	

Figure 18.9 shows our estimated $\lambda^*(t|Z)$ for both positive and negative stained groups. One observes that the estimated hazard rate shows similar pattern for both groups. That is, the hazard rate starts moderate, followed by a stable stage, then high risk. One also notices that the negative group shows lower hazard rate than positive stained group all the time, specifically significant during the periods of the beginning and the end of the study.

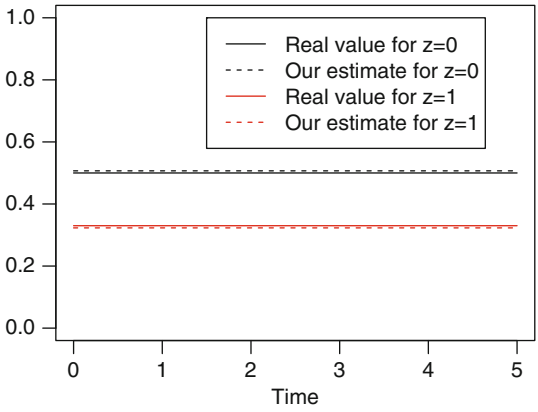


Fig. 18.4 Comparison of $\lambda^*(t|z)$ and $\hat{\lambda}^*(t|z)$ for no change-point model

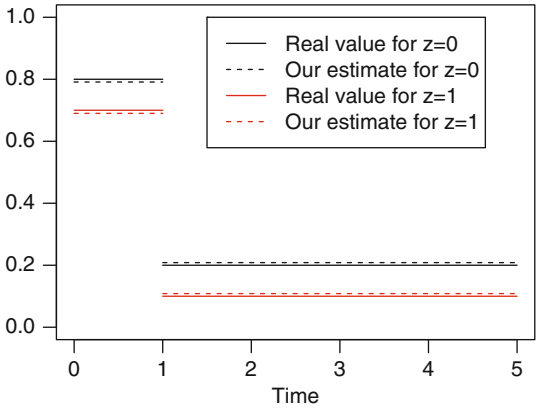


Fig. 18.5 Comparison of $\lambda^*(t|z)$ and $\hat{\lambda}^*(t|z)$ for one change-point model

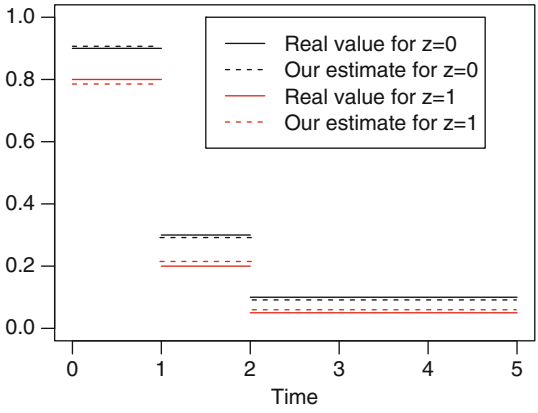


Fig. 18.6 Comparison of $\lambda^*(t|z)$ and $\hat{\lambda}^*(t|z)$ for two change-points model

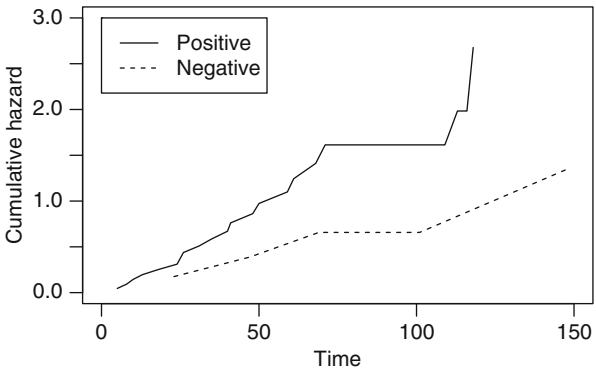


Fig. 18.7 $\Lambda_n^*(t)$ for breast cancer data

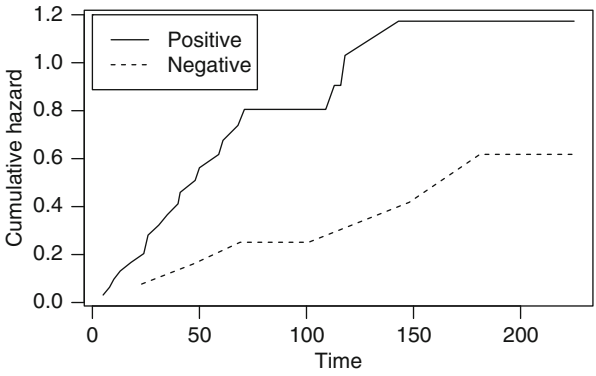


Fig. 18.8 $\Lambda_n(t)$ for breast cancer data

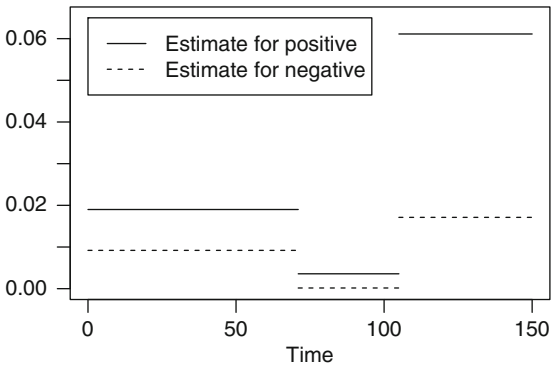


Fig. 18.9 Our estimated $\lambda^*(t|Z)$ for positive and negative stained group

18.6 Conclusions

In this chapter, we introduce a method for fitting failure times in a mixture model that allows the existence of both susceptibles and long-term survivors with covariates observed. We propose an algorithm to fit this kind of data through a grid search

weighted least squared method. The simulation study shows that the misclassification rate is almost ignorable, especially when sample size is relatively large. It also confirms that our algorithm works superiorly on detecting the number of change-points. Furthermore, the estimation for other model parameters are robust and accuracy against various model parameter settings, hence the algorithm is effective. The real data analysis shows the existence of multiple change-points problems with covariate effects and long-term survivors in real life, thus illustrates the importance of the research topic.

Acknowledgements We would like to dedicate this paper to Prof. Hira H. Koul and express our appreciation for his endless encouragement and professional support. We wish him continuing success and health in many years to come! Additionally, we thank the anonymous referees for their helpful comments.

References

Achcar JA, Bolfarine H (1998) Constant hazard against a change-point alternative: a Baysian approach with censored data. Commun Stat—Theory Methods 18(10):3801–3819

Achcar JA, Loibel S (1998) Constant hazard function model with a change-point: a Bayesian analysis using Markov Chain Monte Carlo methods. Biom J 40:543–555

Anis MZ (2009). Inference on a sharp jump in hazard rate: a review. Econ Qual Control 24 (2): 213–229.

Chang IS, Chen CH, Hsiung CA (1994). Estimation in change-point hazard rate models with random censorship. Change-point Problems, IMS Lecture Notes Monograph, Ser. 23, Inst Math Statist, Hayward, CA: 78–92

Dupuy JF (2009) Detecting change in a hazard regression model with rightcensoring. J Stat Plan Infer 139:1578–1586

Dupuy JF (2006) Estimation in a change-point hazard regression model. Stat Probab Lett 76:182–190

Ebrahimi N, Gelfand AE, Ghosh SK, Ghosh M (1997). Bayesian analysis of change point hazard rate problem. Technical Report, University of Connecticut, 97–08

Gijbels I, Gürler U (2003) Estimation of a change point in a hazard function based on censored data. Lifetime Data Anal 9:395–411

Goodman MS, Li Y, Tiwari RC (2011). Detecting multiple change points in piecewise constant hazard functions. J Appl Stat 38(11):2523–2532

Henderson R (1990) A problem with the likelihood ratio test for a change-point hazard rate model. Biometrika 77:835–843

Li YX, Qian LF, Zhang W (2013). Estimation in a change-point hazard regression model with long-term survivors. Stat Probab Lett 83:1683–1691

Loader CR (1991) Inference for a hazard rate change point. Biometrika 78:749–757

Maller RA, Zhou X (1996) Survival Analysis with long-term survivors. Wiley, New York

Matthews DE, Farewell VT (1982) On testing for constant hazard against a change-point alternative. Biometrics 38:463–468

Matthews DE, Farewell VT, Pyke R (1985) Asymptotic score-statistic processes and tests for constant hazard against a change-point alternative. Ann Statist 13:583–591

Nguyen HT, Rogers GS, Walker EA (1984) Estimation in change-point hazard rate models. Biometrika 71:299–304

Pham TD, Nguyen HT (1990) Strong consistency of the maximum likelihood estimators in the change-point hazard rate model. Statistics 21:203–216

Pons O (2003) Estimation in a Cox regression model with a change-point according to a threshold in a covariate. Ann Statist 31(2):442–463

Taylor JMG (1995) Semiparametric estimation in failure time mixture models. Biometrics 51:899–907

Worsley WJ (1988) Exact percentage points of the likelihood ratio test for a change point hazard rate model. Biometrics 44:259–263

Xu AC, Qian LF. (2013). Bayesian estimation in a change-point hazard regression model with long-term survivors. Comput Stat Data Anal. (Submitted)

Yao YC (1986) Maximum likelihood estimation in hazard rate models with a change point. Commun Stat- Theory Methods 15(8):2455–2466

Zhao XB, Wu XY, Zhou X (2009) A change-point model for survival data with long-term survivors. Stat Sinica 19:377–390

Zhang W, Qian LF, Li YX. (2013). Semiparametric sequential testing for multiple change points in piecewise constant hazard functions with long-term survivors. Commun Stat—Simul Comput. DOI:10.1080/03610918.2012.742106

Chapter 19
How to Choose the Number of Gradient Directions for Estimation Problems from Noisy Diffusion Tensor Data

Lyudmila Sakhanenko

19.1 Introduction

Low angular resolution diffusion tensor imaging (DTI) is en vivo brain imaging technique. It is based on measurements of water diffusion on a grid of points. Locally the relative amount of water diffusion along a spatial direction $g \in R^3, \|g\| = 1$, at a voxel x, $S(x, g)$, is estimated as follows:

$$\log\left(\frac{S(x,g)}{S_0(x)}\right) = -cg^* M(x)g + \sigma(x,g)\xi_g, \tag{19.1}$$

where $S_0(x)$ is the amount of water diffusion without gradient application; $\sigma(x, g) > 0$; ξ_g describes noise; the constant c depends only on the proton gyromagnetic ratio, the gradient pulse sequence shape, duration and other timing parameters of the imaging procedure; see (Basser and Pierpaoli 1998). Here and throughout the paper all vectors are columns. $M(x)$ is a diffusion tensor, which is a positive definite 3×3 matrix. In the absence of noise this relationship is known as the Stejskal-Tanner equation. If measurements of S along at least six directions g are recorded, and an ellipsoidal spatial distribution of water diffusion is assumed at each voxel x, then there is sufficient data to estimate the diffusion tensor $M(x)$. Its eigenvector field corresponding to the largest eigenvalue models the gradient field along neural fibers when the tensor is anisotropic. Tracing the fibers is then usually done by following the gradient field in small steps; see (Assemlal et al. 2011) and (Koltchinskii et al. 2007).

DT-MRI data sets along prominent water diffusion directions can provide a geometric representation of those fibers. The axon fibers are scientifically important because they provide pathways through which brain regions communicate, and their integrity is compromised by a variety of diseases as well as invasive neurosurgery. Tracing fiber pathways through DT-MRI data sets allows neuroscientists to study the effects of diseases and treatments on inter-regional communication in the brain, and

L. Sakhanenko (✉)
Department of Statistics and Probability, Michigan State University,
East Lansing, MI 48824–1027, USA
e-mail: luda@stt.msu.edu

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_19,

it allows neurosurgeons to plan procedures such that these communication pathways are preserved; see (Assemlal et al. 2011).

Suppose, at a fixed location x, we observe S along $N \geq 6$ gradients. Using a vector representation $\text{vec}_M(x) = (M_{1,1}, M_{1,2}, M_{1,3}, M_{2,2}, M_{2,3}, M_{3,3})^*(x) \in R^6$ of the symmetrical tensor $M(x) \in R^{3\times 3}$ for a fixed x we observe

$$Z(x) = B\text{vec}_M(x) + \Sigma^{1/2}(x)\Xi_x, \tag{19.2}$$

where $B \in R^{N\times 6}$ is a fixed matrix, $Z(x), \Xi_x \in R^N$ are vectors, and the $N \times N$-tensor $\Sigma(x)$ is symmetric positive definite. Here the entries of Z are estimated by (19.1). The rows of the matrix B are obtained from gg^* for the corresponding vectors g. We assume $E\,\Xi_x\Xi_x^* = I$. Note that this is a linear model with the fixed design.

First, one needs to estimate the tensor $M(x)$ at a fixed location $x \in G$ from the raw Z measurements. There are various ways to do so. The popular approach is to use the ordinary least squares estimator of $M(x)$ for x in some set $G \subset R$

$$\tilde{\text{vec}}_M(x) = (B^*B)^{-1}B^*Z(x), \tag{19.3}$$

provided that $(B^*B)^{-1}$ exists. Another estimator is the weighted least squares estimator of $M(x), x \in G$, which is studied extensively in the work of (Zhu et al. 2007, 2009). This estimator has an improved statistical efficiency and is defined as follows:

$$\tilde{\text{vec}}_{M_w}(x) = (B^*\Sigma^{-1}(x)B)^{-1}B^*\Sigma^{-1}(x)Z(x). \tag{19.4}$$

Note that formula (19.3) can be rewritten as

$$\tilde{\text{vec}}_M(x) = \text{vec}_M(x) + \Gamma, \quad \Gamma = (B^*B)^{-1}B^*\Sigma^{1/2}(x)\Xi_x, \tag{19.5}$$

where Γ denotes a 6D-vector representation of a random tensor in $R^{3\times 3}$.

Note that $E\,\Gamma = 0$ and

$$\begin{aligned} E(\Gamma\Gamma^*) &= (B^*B)^{-1}B^*E[\Sigma^{1/2}(x)\Xi_x\Xi_x(\Sigma^{1/2}(x))^*]B(B^*B)^{-1} \\ &= (B^*B)^{-1}B^*\Sigma(x)B(B^*B)^{-1} =: C(x), \end{aligned}$$

where $C : R^3 \to R^{6\times 6}$ is a tensor field.

Quite similarly, define

$$\tilde{\text{vec}}_{M_w}(x) = \text{vec}_M(x) + \Gamma_w, \quad \Gamma_w = (B^*\Sigma^{-1}(x)B)^{-1}B^*\Sigma^{-1/2}(x)\Xi_x, \tag{19.6}$$

where Γ_w denotes a 6D-vector representation of a random tensor in $R^{3\times 3}$. Note that $E\,\Gamma_w = 0$ and $E(\Gamma_w\Gamma_w^*) = (B^*\Sigma^{-1}(x)B)^{-1} =: C_w(x)$, where $C_w : R^3 \to R^{6\times 6}$ is a tensor field. Since the expression for C_w contains the inverse of Σ the arguments of this paper do not apply readily and go beyond the scope of this work.

The uncertainty in tensor components of M propagates to the eigenvectors and then into trajectories. This uncertainty is often characterized in terms of covariances of estimators of tensors, eigenvectors and trajectories. These covariances are complicated functionals of $C(x)$; see (Koltchinskii et al. 2007) for example. It is of practical

interest to select gradient directions, and thus define B, in a somewhat optimal way in order to make the Lebesgue 2-norm of this matrix $C(x)$ as small as possible for a fixed $x \in G$. This is the goal of this paper.

We show that m independent repetitions using one set of six directions yield smaller norms of $C(x)$ than designs where a large set of 6 m directions is used, assuming that norms of covariances of image components Z_i are similar for both designs. The difference is of the order m^{-1}. On practice m is often 10 or so. This holds for each location on the grid where a brain image is obtained. When a fiber track is estimated, a functional of this covariance integrated along the track gives the covariance of the fiber estimator. This essentially means that the covariances $C(x)$ are accumulated along the track. So the savings in the covariance of fiber track estimator are again of the order m^{-1}. It would mean, in particular, that the confidence ellipsoids are m times wider for a general design of 6 m directions compared to the confidence ellipsoids for a design based on m independent repetitions of a set of just six directions.

We need to keep in mind that this conclusion holds for so-called low resolution case, when the diffusion is adequately described by a 3×3 tensor. This corresponds to locations where diffusion is highly anisotropic and there is just one dominant fiber. However, for locations where several fibers cross or branch the previous model is not sufficient. There the diffusion can be described by means of a $\underbrace{3 \times \ldots \times 3}_{r}$ tensor of a higher order $r > 2$. Then in this paper we show that designs with m independent repetitions using one set of $J_r = (r+1)(r+2)/2$ directions yield smaller norms of $C(x)$ than designs where a large set of $J_r m$ directions is used, assuming that norms of covariances of image components Z_i are similar for both designs. The difference is of the order m^{-1}. On practice $J_r m$ is often 60 or more. For $r = 4$ we have $J_r = 15$, so $m = 4$. This holds for each location on the grid where a brain image is obtained. When a fiber track is estimated, a functional of this covariance integrated along the track gives the covariance of the fiber estimator. Again this essentially means that the covariances $C(x)$ are accumulated along the track. So the savings in the covariance of fiber track estimator are of the order m^{-1}. As in the low resolution case, the confidence ellipsoids based on a general design of $J_r m$ directions, are m times wider than those confidence ellipsoids that are based on a design with m independent repetitions of a set of J_r directions.

The rest of the chapter is split into three sections. We study the low angular resolution case in Sect. 19.2, and the high angular resolution case in Sect. 19.3. Section 19.4 contains conclusions.

19.2 Low Angular Resolution Case

Let $m \geq 6$ be an integer, and let $N = 6\,m$. Let us present matrices B and Σ as

$$B = (B_1, \ldots, B_m)^*, \quad B_k \in R^{6\times 6}, \quad \Sigma = (\Sigma_{kl})_{k,l=1,\ldots,m}, \quad \Sigma_{kl} \in R^{6\times 6}.$$

Typically, B_1 is constructed from two orthogonal systems of vectors in R^3. For example the oblique double gradient encoding uses directions $g_1 = 2^{-1/2}(1,0,1)$, $g_2 = 2^{-1/2}(1,0,-1)$, $g_3 = 2^{-1/2}(0,1,1)$, $g_4 = 2^{-1/2}(0,-1,1)$, $g_5 = 2^{-1/2}(1,1,0)$, $g_6 = 2^{-1/2}(-1,1,0)$. The matrix B_1 has the k-th row that is proportional to $(g_{k,1}^2, 2g_{k,1}g_{k,2}, 2g_{k,1}g_{k,3}, g_{k,2}^2, 2g_{k,2}g_{k,3}, g_{k,3}^2)$ for $k = 1,\ldots,6$; see (Basser and Pierpaoli 1998) for details. Hence, the matrix B_1 corresponding to this set of gradient directions is proportional to

$$B_1 = \begin{pmatrix} -1 & 0 & -2 & 0 & 0 & -1 \\ -1 & 0 & 2 & 0 & 0 & -1 \\ 0 & 0 & 0 & -1 & 2 & -1 \\ 0 & 0 & 0 & -1 & -2 & -1 \\ -1 & 2 & 0 & -1 & 0 & 0 \\ -1 & -2 & 0 & -1 & 0 & 0 \end{pmatrix}.$$

Often the other parts $B_k, k = 2,\ldots,m$, are constructed from the same systems by means of rotations. So more precisely $B_k = A_k B_1$ for some finite matrices A_k with bounded norms for $k = 2,\ldots,m$. Let $A_1 = I$. Then,

$$B^*B = \sum_{k=1}^{m} B_k^* B_k = B_1^* \left(I + \sum_{k=2}^{m} A_k^* A_k \right) B_1,$$

$$B^*\Sigma(x)B = \sum_{i=1}^{m}\sum_{k=1}^{m} B_k^* \Sigma_{ki}(x) B_i = B_1^* \left(\sum_{i=1}^{m}\sum_{k=1}^{m} A_k^* \Sigma_{ki}(x) A_i \right) B_1.$$

We obtain

$$C(x) = B_1^{-1} \left(I + \sum_{k=2}^{m} A_k^* A_k \right)^{-1} \left(\sum_{i=1}^{m}\sum_{k=1}^{m} A_k^* \Sigma_{ki}(x) A_i \right) \left(I + \sum_{k=2}^{m} A_k^* A_k \right)^{-1} (B_1^*)^{-1}. \tag{19.7}$$

Proposition. *Let $\Sigma_{kl} = 0$ for all but $C_1 m^\alpha$ of the matrices, where $\alpha \in [0,2]$ and $C_1 > 0$. Then $\|C(x)\|_2 = C_2 m^{\alpha-2}$ for some finite positive constant C_2.*

The proof is obvious. Now let us consider several cases.

Case 1. Let $\Sigma_{kl} = 0$ for all but a fixed number p of the matrices. Then $\|C(x)\|_2 = C_p m^{-2}$, where the constant depends on p only. In this case the log-losses of signal can be observed without noise in most of the experiments. This is not a realistic scenario.

Case 2. Let $\Sigma_{kl} = 0$ for all but m of the matrices. For example, $\Sigma_{kl} = 0$ for all $k \neq l$. Then $\|C(x)\|_2 = Cm^{-1}$. This is a typical scenario when six directions are chosen and fixed. Then the measurements along these directions are repeated independently m times. In this case $A_k = I, k = 1,2,\ldots,m$ and $S(x) = m^{-2} B_1^{-1} \sum_{k=1}^{m} \Sigma_{kk}(x)(B_1^*)^{-1}$. Then $\|C(x)\| = m^{-1}\|B_1^{-1}\|^2 \max_{k=1,\ldots,m}\|\Sigma_{kk}(x)\|$.

Case 3. Let $\Sigma_{kl} = 0$ for all but γm^2 of the matrices, where $\gamma \in (0, 1]$. For example, $\Sigma_{kl} = 0$ for all $k > l$. Then $\|C(x)\|_2$ is a finite constant. This is a typical scenario when $N = 6\,m$ directions are chosen and no independence is assumed.

Once B is selected and the measurements $Z(x_i), i = 1, \ldots, n$, are collected on a grid $x_i, i = 1, \ldots, n$, one can estimate the tensor M, its main eigenvector and the trajectory along the eigenvector field. Under proper assumptions all these entities can be estimated by asymptotically normal estimators. Thus, the covariances of these estimators are an appropriate gauge of the uncertainty in image measurements. Since these covariances are functionals of $C(x)$, the optimal designs should have small norms of $C(x)$. Then case 2 is an appealing design.

19.3 High Angular Resolution Case

The model for low angular resolution DTI does not allow branching or crossing of fibers. To generalize to these scenarios several other models have been suggested. We consider only nonparametric models in order to be fair in comparison with the model (19.8). (Özarslan and Mareci 2003) and (Descoteaux et al. 2006) propose to model the relative amount of water diffusion along a spatial direction $g \in R^3, \|g\| = 1$, at a voxel x by a tensor of higher order $r > 2$:

$$\log\left(\frac{S(x,g)}{S_0(x)}\right) = -c \sum_{i_1=1}^{3} \cdots \sum_{i_r=1}^{3} T_{i_1 \ldots i_r}(x) g_{i_1} \ldots g_{i_r} + \sigma(x,g)\xi_g,$$

$\sigma(x, g) > 0$; ξ_g describes noise; the constant c depends only on the proton gyromagnetic ratio, the gradient pulse sequence shape, duration and other timing parameters of the imaging procedure. $T(x)$ with components $T_{i_1 \ldots i_r}(x)$ is a diffusion tensor, which is a supersymmetrical positive definite $\underbrace{3 \times \ldots \times 3}_{r}$ tensor. So r is an even number. Due to symmetry $T(x)$ can be represented by a vector $\mathrm{vec}_T \in R^{J_r}$ with the dimension $J_r = (r + 1)(r + 2)/2$. We repeat the procedure in the previous section where we replace 6 by J_r and M by T. So with a slight abuse of notation, let $N = J_r m$ for some $m \geq 1$ and then at a fixed location x we observe

$$Z(x) = B\mathrm{vec}_T(x) + \Sigma^{1/2}(x)\Xi_x \tag{19.8}$$

where $B \in R^{N \times J_r}$ is a fixed matrix, $Z(x), \Xi_x \in R^N$ are vectors, the $N \times N$-tensor $\Sigma(x)$ is symmetric positive definite. We estimate $\mathrm{vec}_T(x)$ by $\tilde{\mathrm{vec}}_T(x) := (B^*B)^{-1}B^*Z(x)$. As in the previous section we represent B and Σ as

$$B = (B_1, \ldots, B_m)^*, \quad B_k \in R^{J_r \times J_r}, \quad \Sigma = (\Sigma_{kl})_{k,l=1,\ldots,m}, \quad \Sigma_{kl} \in R^{J_r \times J_r},$$

so that the covariance of $\tilde{\mathrm{vec}}_T(x)$ is written in (19.7). Then, the proposition also holds for the high angular resolution case. Thus, designs based on independent repetitions of the same system of gradients would yield smaller norms of covariance $C(x)$, so

the tensor T would be estimated better. Currently, there is no statistical work on how the uncertainty in Z propagates to pseudo-eigenvectors and to trajectories, but we conjecture that better estimators of tensor T would yield better estimators for pseudo-eigenvectors and for trajectories.

19.4 Conclusion

Right now most practitioners agreed that it is better to increase the number of distinct directions rather than to increase m. Pretty much everybody is using the largest N possible with $m = 1$. Contrary to the common practice, in both cases $r = 2$ and $r > 2$ the designs with m independent repetitions of a set of J_r directions lead to smaller norms of covariance $C(x)$ at all locations x, than those obtained by means of general designs with $J_r m$ directions. Smaller norms of covariance $C(x)$ translate into tighter confidence cones for eigenvectors and tighter confidence ellipsoids along estimated fibers. Thus, we would recommend to use these special designs with m independent repetitions on practice. From the statistical point of view these designs yield better estimators for all the objects of interest, including tensor components, eigenvectors, and fibers.

Acknowledgements Research is partially supported by NSF grant DMS-1208238. The author is grateful to professor Hira Koul for his guidance, help, and contagious enthusiasm for statistics.

References

Assemlal H-E, Tschumperle D, Brun L, Siddiqi K (2011) Recent advances in diffusion MRI modeling: angular and radial reconstruction. Med Image An 15:369–396

Basser P, Pierpaoli C A simplified method to measure the diffusion tensor from seven MR images. Magn Reson Med 39:928–934

Descoteaux M, Angelino E, Fitzgibbons S, Deriche R (2006) Apparent diffusion coefficients from high angular resolution diffusion imaging: estimation and applications. Magn Reson Med 56 (2):395–410

Koltchinskii V, Sakhanenko L, Cai S (2007) Integral curves of noisy vector fields and statistical problems in diffusion tensor imaging: nonparametric kernel estimation and hypotheses testing. Ann Stat 35:1576–1607

Özarslan E, Mareci T (2003) Generalized diffusion tensor imaging and analytical relationships between diffusion tensor imaging and high angular resolution diffusion imaging. Magn Reson Med 50(5):955–965

Zhu H, Zhang H, Ibrahim J, Peterson B (2007) Statistical analysis of diffusion tensors in diffusion-weighted magnetic resonance image data. J. Amer Stat Assoc102:1081–1110

Zhu H, Li Y, Ibrahim I, Shi X, An H, Chen Y, Gao W, Lin W, Rowe D, Peterson B (2009) Regression models for identifying noise sources in magnetic resonance images. J Amer Stat Assoc 104:623–637

Chapter 20
Efficient Estimation in Two-Sided Truncated Location Models

Weixing Song

20.1 Introduction

The adaptiveness and the asymptotic efficiency are very important concepts in the theory of statistical estimation. Extensive research has been done when the underlying distributions have common support. See Ibragimov and Hasminski (1981); Akahira and Takeuchi (1981), and Bickel et al. (1998) for detailed discussion on these topics. Starting from 1970s, the research on the adaptiveness and the asymptotic efficiency in nonregular models, in particular, when the underlying distributions are not commonly supported, began to emerge. Early works on this topic were summarized in Akahira and Takeuchi (2003) and the references therein. For the unknown parameter θ in a class of uniformly distributed distribution family, Akahira (1982) successfully constructed an upper bound of the asymptotic distributions of $n(\hat{\theta}_n - \theta)$ for all asymptotically median unbiased (AMU, which will be defined later) estimators $\hat{\theta}_n$ of θ using the Neyman–Pearson testing framework. The concept of two-sided asymptotic efficiency is thus defined based on this bound and some examples were supplied. Akahira (1982) also noticed that for some examples, the proposed AMU estimators only attain the bound at one point, or are uniformly "close" to the bound. It is not clear, however, whether there exist any AMU estimators to be two-sided asymptotically efficient.

To be specific, let X be a random variable with distribution $P_\theta, \theta \in \Theta$. The parameter space Θ is assumed to be an open set in $\mathbb{R}$. Denote $\hat{\theta}_n$ an estimator of θ based on a sample $X_1, X_2, \ldots, X_n$ of size n from X. Let $\{c_n\}$ be a sequence of positive numbers tending to infinity as $n \to \infty$. Then $\hat{\theta}_n$ is called a consistent estimator of order $\{c_n\}$ if for every $\varepsilon > 0$ and every $\vartheta \in \Theta$ there exists a sufficiently small number $\delta > 0$ and a sufficiently large number L satisfying the following inequality

$$\limsup_{n\to\infty} \sup_{\theta:|\theta-\vartheta|<\delta} P_\theta\{c_n|\hat{\theta}_n - \theta| \geq L\} < \varepsilon. \tag{20.1}$$

W. Song (✉)
Department of Statistics, Kansas State University, Manhattan, Kansas, USA
e-mail: weixing@ksu.edu

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_20,

A cumulative distribution function $F_\theta(\cdot)$ is called the asymptotic distribution function of $c_n(\hat{\theta}_n - \theta)$, if for each real number t, $F_\theta(t)$ is continuous in θ, and for any $\vartheta \in \Theta$ there exists a positive number δ such that for any continuity point t of $F_\theta(\cdot)$,

$$\limsup_{n\to\infty} \sup_{\theta:|\theta-\vartheta|<\delta} \left| P_\theta\{c_n(\hat{\theta}_n - \theta) \le t\} - F_\theta(t) \right| < \varepsilon.$$

Note that some requirements, such as the uniform requirement of $\sup_{\theta:|\theta-\vartheta|<\delta}$, do not present in the usual definitions of consistency and asymptotic distribution with order $\{c_n\}$. An estimator $\hat{\theta}_n$ is called to be an AMU if for every $\vartheta \in \Theta$, there exists a positive number δ such that

$$\limsup_{n\to\infty} \sup_{\theta:|\theta-\vartheta|<\delta} \left| P_\theta\{\hat{\theta}_n \le \theta\} - \frac{1}{2} \right| = 0,$$

$$\limsup_{n\to\infty} \sup_{\theta:|\theta-\vartheta|<\delta} \left| P_\theta\{\hat{\theta}_n \ge \theta\} - \frac{1}{2} \right| = 0.$$

For the class of AMU estimators of θ, Akahira (1982) proposed the following left-hand side and right-hand side asymptotic efficiency.

Definition 1. *An AMU estimator $\hat{\theta}_n$ is called right-hand side (left-hand side) asymptotically efficient if for any AMU estimator $\tilde{\theta}_n$,*

$$\liminf_{n\to\infty} \left[P_\theta\{c_n(\hat{\theta}_n - \theta) \le t\} - P_\theta\{c_n(\tilde{\theta}_n - \theta) \le t\} \right] \ge 0, \quad \text{for all } t > 0$$

$$(\liminf_{n\to\infty} \left[P_\theta\{c_n(\tilde{\theta}_n - \theta) \le t\} - P_\theta\{c_n(\hat{\theta}_n - \theta) \le t\} \right] \ge 0, \quad \text{for all } t < 0).$$

The above definition is intuitively well-defined, but in practice, to use the definition as a criterion to check an AMU estimator to be left-hand or right-hand side asymptotically efficient, we have to find a tangible upper bound for $P_\theta\{c_n(\tilde{\theta}_n - \theta) \le t\}$ when $t > 0$, $P_\theta\{c_n(\tilde{\theta}_n - \theta) \ge t\}$ when $t < 0$, for all AMU estimators $\tilde{\theta}_n$ of θ. For some particular distribution families, such an upper bound is constructed based on the Neyman–Pearson lemma after properly setting up a simple versus simple hypothesis testing problem about the unknown parameter. The detailed derivation of the upper bound can be found in (Akahira 1982). In some nonregular cases, Akahira (1982) also showed that there exist either right-hand side asymptotically efficient estimator or left-hand side asymptotically efficient estimators. However, in general there are no AMU estimators to be both right-hand and left-hand side asymptotically efficient. See Takeuchi (1974) for some examples.

A weaker version than both right-hand and left-hand side asymptotic efficiency is the following two-sided asymptotic efficiency.

Definition 2. *An AMU estimator $\hat{\theta}_n$ of θ is called two-sided asymptotically efficient if for any AMU estimator $\tilde{\theta}_n$ and $t > 0$,*

$$\liminf_{n\to\infty} \left[P_\theta\{c_n|\hat{\theta}_n - \theta| \le t\} - P_\theta\{c_n|\tilde{\theta}_n - \theta| \le t\} \right] \ge 0.$$

For a special case in which X, or some transformation of X, is uniformly distributed over $[a(\theta), b(\theta)]$, $a(\theta) < b(\theta)$ and $a'(\theta) \leq b'(\theta) < 0$, where $a'(\theta)$ and $b'(\theta)$ are the derivatives of $a(\theta)$ and $b(\theta)$ with respect to θ, respectively, Akahira (1982) proposed an upper bound for $\limsup_{n\to\infty} P_\theta\{n|\hat{\theta}_n - \theta| \leq t\}$, based on the generalized Neyman–Pearson lemma. For a uniform distribution and a symmetric truncated normal distribution, Akahira (1982) showed that $[X_{(1)} + X_{(n)}]/2$ is two-sided asymptotically efficient, while in a truncated exponential distribution case and an asymmetric truncated normal distribution case, two estimators are considered, but they are not asymptotically efficient, although for some t values, $\limsup_{n\to\infty} P_\theta\{n|\hat{\theta}_n - \theta| \leq t\}$ attains the upper bound. The existence of asymptotically efficient estimator in such distribution families has not been answered.

In this paper, we shall focus on a class of truncated location family which subsume the truncation family discussed in (Akahira 1982). By adopting the Neyman–Pearson testing framework, the left-hand side, right-hand side, and two-sided efficiency are discussed. The question of the existence of two-sided efficient estimation will be completely addressed in the paper, based on the newly defined notion of the asymptotically weak inadmissible median unbiased estimator. In fact, it is shown that in this particular distribution family, whether or not there exist asymptotically efficient estimators is totally determined by whether or not the density function at both ends are equal.

The paper is organized as follows. The model of interest and technical assumptions are laid out in Sect. 20.2, together with an important lemma and two examples. Section 20.3 includes the main results on one-sided asymptotic efficiency and the two-sided efficiency is discussed in Sect. 20.4; all the technical proofs are deferred to Sect. 20.5.

20.2 Models and Assumptions

Let $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ be a sample of size n from a truncated location distribution

$$dP_\theta(x) = f(x-\theta) I_{[\theta+a,\theta+b]}(x)dx, \tag{20.2}$$

where a and b are two known real numbers, and θ is the unknown location parameter to be estimated. Here we do not have to assume f is known, even at the truncation points. Hence the model discussed here is essentially a semiparametric model and is more flexible than the one discussed in Akahira (1982).

The following regularity conditions on model (2) are needed to set up the bounds for defining the one-sided and two-sided efficiencies.

Assumptions:

(A1). $f(x)$ is twice continuously differentiable on $[a, b]$, and $f(x) > 0$ on $[a, b]$.
(A2). $f(a)f(b) > 0$.
(A3). $0 \leq I = \int_a^b [f'(x)]^2/f(x)dx < \infty$.

(A1) guarantees that a Taylor expansion of $f(x)$ up to second order can be implemented and the second order term can be neglected in some integrals when n is large. Similar to Akahira (1982), we only consider all consistent estimators of θ of order $c_n = n$, see (20.1) for the definition of consistency. In fact, one can show that for the truncated distribution family (20.2), n is the largest consistency order under assumption (A2). In condition (A3), if the integration is positive, then the central limit theorem can be applied to derive the asymptotic power functions of the tests for the hypothesis proposed in the following discussion; if the integration is 0, which implies that $f(x)$ is constant almost everywhere with respect to the Lebesgue measure on $[a, b]$, that is, the underlying distribution of X is uniform on $[a, b]$. In this case, a randomized most powerful test is needed to calculate the power function. This is also the case discussed by (Weiss and Wolfowitz 1968), and Akahira (1982).

The following lemma will be frequently used in the proof of the main results stated in the next two sections.

Lemma 1 *Suppose the two-sided truncation model (2) satisfies condition (A1), (A2) and (A3) with strict inequality. Then for any real number $t > 0$,*

$$\prod_{i=1}^{n} \frac{f(X_i - \theta - t/n)}{f(X_i - \theta)} \to \exp\left(t[f(a) - f(b)]\right) \tag{20.3}$$

in probability conditioning on $A_n = \cap_{i=1}^{n} A_{ni}$, where $A_{ni} = \{a + \theta + t/n < X_i < b + \theta\}$ no matter the true parameter is θ or $\theta + t/n$, and

$$\prod_{i=1}^{n} \frac{f(X_i - \theta + t/n)}{f(X_i - \theta)} \to \exp\left(-t[f(a) - f(b)]\right) \tag{20.4}$$

in probability conditioning on $B_n = \cap_{i=1}^{n} B_{ni}$, where $B_{ni} = \{a+\theta < X_i < b+\theta-t/n\}$ no matter the true parameter is θ or $\theta - t/n$. Furthermore, for $t > 0$,

$$\sqrt{n}\left(\sum_{i=1}^{n} \log\left[\frac{f(X_i - \theta - t/n)}{f(X_i - \theta)} I_{A_{ni}}\right] - t[f(a) - f(b)]\right) \Longrightarrow N(0, \sigma^2(t)), \tag{20.5}$$

$$\sqrt{n}\left(\sum_{i=1}^{n} \log\left[\frac{f(X_i - \theta + t/n)}{f(X_i - \theta)} I_{B_{ni}}\right] + t[f(a) - f(b)]\right) \Longrightarrow N(0, \sigma^2(t)) \tag{20.6}$$

conditioning on A_n and B_n, respectively, where $\sigma^2(t) = t^2[I - (f(b) - f(a))^2]$. Two examples are given below to illustrate the validity of (20.3) and (20.4) in Lemma 1.

Example 1. Suppose $X_1, X_2, \ldots, X_n$ are independent and identically distributed (i.i.d.) from a truncated exponential distribution $f(x - \theta) = ce^{-(x-\theta)} I(a \le x - \theta \le b)$, where $a < b$, and $c = (e^{-a} - e^{-b})^{-1}$. Note that $f(a) = ce^{-a}$, $f(b) = ce^{-b}$ and

$f(a)-f(b)=1$. For $t>0$, and conditioning on $\{a+\theta+t/n < X_{(1)} \le X_{(n)} < b+\theta\}$, we simply have

$$\prod_{i=1}^{n} \frac{f(X_i-\theta-t/n)}{f(X_i-\theta)} = e^t = e^{t[f(a)-f(b)]},$$

no matter the true parameter is θ or $\theta+t/n$, and conditioning on $\{a+\theta < X_{(1)} \le X_{(n)} < b+\theta-t/n\}$, we simply have

$$\prod_{i=1}^{n} \frac{f(X_i-\theta+t/n)}{f(X_i-\theta)} = e^{-t} = e^{-t[f(a)-f(b)]},$$

no matter the true parameter is θ or $\theta-t/n$.

Example 2. Suppose $X_1, X_2, \ldots, X_n$ are i.i.d. from a truncated normal distribution distribution $f(x-\theta) = ce^{-(x-\theta)^2/2} I(a < x-\theta < b)$, where $a < b$ and $c = (\Phi(b)-\Phi(a))^{-1}$, Φ is the CDF of the standard normal variable. Note that $f(a) = ce^{-a^2/2}$, $f(b) = ce^{-b^2/2}$. For $t>0$, and conditioning on $\{a+\theta+t/n < X_{(1)} \le X_{(n)} < b+\theta\}$, we can show that, in probability,

$$\prod_{i=1}^{n} \frac{f(X_i-\theta-t/n)}{f(X_i-\theta)} = \exp\left(-\frac{t^2}{2n} + \frac{t}{n}\sum_{i=1}^{n}(X_i-\theta)\right) \to e^{t[f(a)-f(b)]},$$

since, by law of large numbers, no matter the true parameter is θ or $\theta+t/n$, $n^{-1}\sum_{i=1}^{n}(X_i-\theta) \to f(a)-f(b)$ in probability as $n\to\infty$. Similarly one can obtain (20.4).

20.3 Left-hand and Right-hand Side Asymptotic Efficiency

Let $\hat{\theta}_n$ be a AMU estimator of θ, and define

$$B(t) = \begin{cases} \liminf_{n\to\infty} P_\theta\{n(\hat{\theta}_n-\theta) \ge t\}, & t<0, \\ \liminf_{n\to\infty} P_\theta\{n(\hat{\theta}_n-\theta) \le t\}, & t \ge 0. \end{cases}$$

It is easy to see that the larger the value of $B(t)$, the smaller the deviation of the AMU estimator $\hat{\theta}_n$ from θ. To find an upper bound for $B(t)$. we consider the most powerful test for the following hypothesis,

$$H_0 : \theta = \theta_0 + \frac{t}{n}, \quad \text{v.s.} \quad H_1 : \theta = \theta_0.$$

Let $\phi_n(X)$ be the most powerful test determined by the Neyman–Pearson lemma with asymptotic significance level $1/2$, and denote $\beta(t)$ the asymptotic power function of ϕ_n. Then the optimality of most powerful test implies that $B(t) \le \beta(t)$ for any $\theta \in \Theta$ and $t \in \mathbb{R}$. Thus, an AMU estimator $\hat{\theta}_n$ is left-hand side or right-hand side asymptotically efficient if and only if $B(t) = \beta(t)$ for all t.

The following theorem provides an explicit form for $\beta(t)$.

Theorem 1. *Suppose the two-sided truncation model (2) satisfies condition (A1), (A2) and (A3). Then*

$$\beta(t) = \begin{cases} 1, & t < -\frac{\log 2}{f(a)}, \\ 1 - e^{f(b)t} + \frac{1}{2}e^{[f(b)-f(a)]t}, & -\frac{\log 2}{f(a)} \le t \le 0, \\ 1 - e^{-f(a)t} + \frac{1}{2}e^{[f(b)-f(a)]t}, & 0 \le t \le \frac{\log 2}{f(b)}, \\ 1, & t > \frac{\log 2}{f(b)}. \end{cases} \tag{20.7}$$

The condition in Theorem 1 is similar to the one used in Weiss and Wolfowitz (1968). They also provide a sufficient condition on f to ensure the validity of (20.3). It is not difficult to modify their sufficient condition to fit the current setup.

In fact, the above result is also true when $f(a) = 0$, $f(b) > 0$, or $f(a) > 0$, $f(b) = 0$, see the proof of Theorem 1. In these cases, one can easily construct left-hand side or right-hand side asymptotically efficient estimators. However, there is no AMU estimator to be both left-hand side and right-hand side asymptotically efficient. Also see Takeuchi (1974) for some interesting examples. In particular, for our current setup, we can show that

Corollary 1. *If $f(a) = f(b)$ in model (2), then $\beta(t) = 1.5 - e^{f(a)|t|}$ for $|t| \le log2/f(a)$, and 1 otherwise. Furthermore, we claim that $\hat{\theta}_{1/2} = (X_{(1)}+X_{(n)}-a-b)/2$ is not left-hand side or right-hand side asymptotically efficient.*

20.4 Two-Sided Asymptotic Efficiency

In this section we shall discuss the two-sided asymptotic efficiency of AMU estimators in model (2). Denote θ_0 the true value of θ, and for each $t > 0$, $\theta_1 = \theta_0 + t/n$, $\theta_2 = \theta_0 - t/n$. Then in the neighborhood of θ_0, by a similar argument as in Akahira (1982), we can show that an upper bound for $\limsup_{n\to\infty} P_\theta\{n|\hat{\theta}_n - \theta_0| < t\}$ is given by $\beta(t) = \limsup_{n\to\infty}(E_{\theta_2}\phi_n(X) - E_{\theta_1}\phi_n(X))$, and $\phi_n(X)$ is defined as

$$\phi_n(X) = \begin{cases} 1, & \prod_{i=1}^n f(X_i-\theta_2)I_{A_i} - \prod_{i=1}^n f(X_i-\theta_1)I_{B_i} > \lambda_n \prod_{i=1}^n f(X_i-\theta_0)I_{C_i}, \\ r, & \prod_{i=1}^n f(X_i-\theta_2)I_{A_i} - \prod_{i=1}^n f(X_i-\theta_1)I_{B_i} = \lambda_n \prod_{i=1}^n f(X_i-\theta_0)I_{C_i}, \\ 0, & \prod_{i=1}^n f(X_i-\theta_2)I_{A_i} - \prod_{i=1}^n f(X_i-\theta_1)I_{B_i} < \lambda_n \prod_{i=1}^n f(X_i-\theta_0)I_{C_i}, \end{cases} \tag{20.8}$$

where λ_n satisfies

$$\lim_{n\to\infty} E_{\theta_0}\phi_n(X) = \frac{1}{2} \tag{20.9}$$

and $A_i = \{X_i : a \le X_i - \theta_2 \le b\}$, $B_i = \{X_i : a \le X_i - \theta_1 \le b\}$, $C_i = \{X_i : a \le X_i - \theta_0 \le b\}$, $i = 1, 2, \ldots, n$. For the sake of completeness, we restate the definition of the two-sided asymptotically efficiency below.

Definition 3. *An estimator $\hat{\theta}_n$ is said to be a two-sided asymptotic efficient estimator of θ in the distribution (20.2) if it satisfies*

(1). $\hat{\theta}_n$ is an AMU estimator of θ with order n,
(2). $\limsup_{n\to\infty} P_\theta\{n|\hat{\theta}_n - \theta| < t\} = \beta(t)$ for all $t > 0$ and all $\theta \in \Theta$.

For the sake of brevity, denote $f_1 = \min\{\mathrm{f(a)}, \mathrm{f(b)}\}$, $f_2 = \max\{\mathrm{f(a)}, \mathrm{f(b)}\}$, $f^- = f(a) - f(b)$ and $f^+ = f(a) + f(b)$. The upper bound $\beta(t)$ is provided in the following theorem.

Theorem 2. *For the distribution family (20.1), when (A1), (A2) and (A3) hold, we have*

$$\beta(t) = \begin{cases} 1 - e^{-tf_2} + \frac{1}{2}e^{-tf^-}, & \text{if} \quad e^{-tf_1} - e^{-tf^+} \ge \frac{1}{2}, \\ 1 - e^{-2tf_2}, & \text{if} \quad e^{-tf_1} - e^{-tf^+} \le \frac{1}{2},\ t \ge \frac{\log 2}{f_2}, \\ 1 - e^{-tf_2} + (1 - 2e^{-tf_2})\sinh(tf^-), & \text{if} \quad e^{-tf_1} - e^{-tf^+} \le \frac{1}{2}, \\ & e^{-tf_2} - e^{-tf^+} \le \frac{1}{2},\ t < \frac{\log 2}{f_2}, \\ 1 - e^{-tf_1} + \frac{1}{2}e^{tf^-}, & \text{if} \quad e^{-tf_1} - e^{-tf^+} \le \frac{1}{2}, \\ & e^{-tf_2} - e^{-tf^+} > \frac{1}{2},\ t < \frac{\log 2}{f_2}. \end{cases}$$

where $\sinh(x) = (e^x - e^{-x})/2$ is the hyperbolic sine function.

In particular, if $f(a) = f(b) > 0$, we have

Corollary 2. *In addition to the conditions in Theorem 2, we further assume that $f(a) = f(b) > 0$ in the distribution family (20.1). Then $\beta(t) = 1 - e^{-2tf(a)}$.*

Now, define

$$\hat{\theta}_n^* = \frac{1}{2}[X_{(1)} + X_{(n)} - a - b].$$

By the asymptotic independence of $X_{(1)}$ and $X_{(n)}$, one can show that, for any $t > 0$,

$$\lim_{n\to\infty} P_\theta\{n|\hat{\theta}_n^* - \theta_0| < t\} = 1 - e^{-2\,f(a)t} + \frac{f(a)}{f(a) + f(b)}[e^{-2\,f(a)t} - e^{-2\,f(b)t}].$$

In the case of $f(a) = f(b) > 0$, we can see that, for any $t > 0$,

$$\lim_{n\to\infty} P_\theta\{n|\hat{\theta}_n^* - \theta_0| < t\} = 1 - e^{-2\,f(a)t}$$

which is equal to $\beta(t)$ given in Corollary 2. That is, $\hat{\theta}_n^*$ is two-sided asymptotically efficient when $f(a) = f(b)$. However, one can check that $\hat{\theta}_n^*$ is not two-sided asymptotically efficient in the case of $f(a) \ne f(b)$. In fact, we will show that no AMU estimator attains the upper bound $\beta(t)$, that is, no AMU estimator is two-sided asymptotically efficient in the sense of Definition 3. To prove our claim, the following definition is needed and the definition itself may not be just limited to the location model (2).

Definition 4. Let $X_1, X_2, \ldots, X_n$ *be a sample from the distribution family* $f(x,\theta)$, *and denote* $\mathcal{A}$ *as the set of all the AMU estimators of* θ. *An estimator* $\hat{\theta}_n \in \mathcal{A}$ *is called to be an asymptotically weak inadmissible median unbiased estimator, if there exists an estimator* $\tilde{\theta}_n \in \mathcal{A} - \{\hat{\theta}_n\}$ such that

$$\liminf_{n\to\infty} P_\theta\{c_n|\tilde{\theta}_n - \theta| < t\} \geq \limsup_{n\to\infty} P_\theta\{c_n|\hat{\theta}_n - \theta| < t\}$$

holds for all $\theta \in \Theta$, and $t > 0$; moreover, for every $\theta \in \Theta$, there exists a set A_θ with positive Lebesgue measure, the strict inequality holds for all $\theta \in \Theta$ and $t \in A_\theta$.

Without loss of generality, we shall assume that $f(a) > f(b)$. The main result we obtained is the following theorem.

Theorem 3. *Let* $\mathbf{X} = \{X_1, X_2, \ldots, X_n\}$ *be a sample from the distribution family* (20.2),*then the solution* $T_n(\mathbf{X})$ *of the following equation*

$$\frac{\int_{X_{(n)}-b}^{T_n(\mathbf{X})} \prod_{i=1}^n f(X_i - \theta)d\theta}{\int_{X_{(n)}-b}^{X_{(1)}-a} \prod_{i=1}^n f(X_i - \theta)d\theta} = \frac{1}{2} \tag{20.10}$$

is an asymptotically weak admissible median unbiased estimator of θ, *and* $T_n(\mathbf{X})$ *is equivalent to the solution* $T_n^*(\mathbf{X})$ *of the following equation*

$$e^{nkT_n^*(\mathbf{X})} = \frac{1}{2}\left[e^{nk(X_{(n)}-b)} + e^{nk(X_{(1)}-a)}\right] \tag{20.11}$$

in the sense that $T_n(\mathbf{X})$ *and* $T_n^*(\mathbf{X})$ *have the same asymptotic distribution, where* $k = f(a) - f(b) > 0$. *Moreover, we claim that there is no AMU estimator in* $\mathcal{A}$ *which is two sided asymptotically efficient*.

As an example, we consider the following two-sided truncated exponential distribution family

$$f(x) = \begin{cases} ce^{-(x-\theta)}, & \theta \leq x \leq \theta + 1, \\ 0, & \text{otherwise}, \end{cases}$$

where $c = (1 - e^{-1})^{-1}$. Then the solution $T_n(\mathbf{X})$ defined in (20.10) has the form

$$T_n(\mathbf{X}) = \frac{1}{n}\log\frac{1}{2}\left[e^{nX_{(1)}} + e^{n(X_{(n)})-1}\right].$$

We can show that, as $n \to \infty$, $P_\theta\{|T_n(\mathbf{X}) - \theta| < t\}$ converges to

$$W(t) = \begin{cases} 1 - \frac{1}{2e^t}(2e^t - 1)^{-ce^{-1}} - \frac{1}{2e^{-t}}(2e^{-t} - 1)^c, & 0 \leq t \leq \log 2, \\ 1 - \frac{1}{2e^t}(2e^t - 1)^{-ce^{-1}}, & t > \log 2. \end{cases}$$

Akahira (1982) showed that among all AMU estimators having the form of $\hat{\theta}_p = pX_{(1)} + (1-p)(X_{(n)} - 1)$, the one with $p = e/(1+e)$ is the best in the sense that the

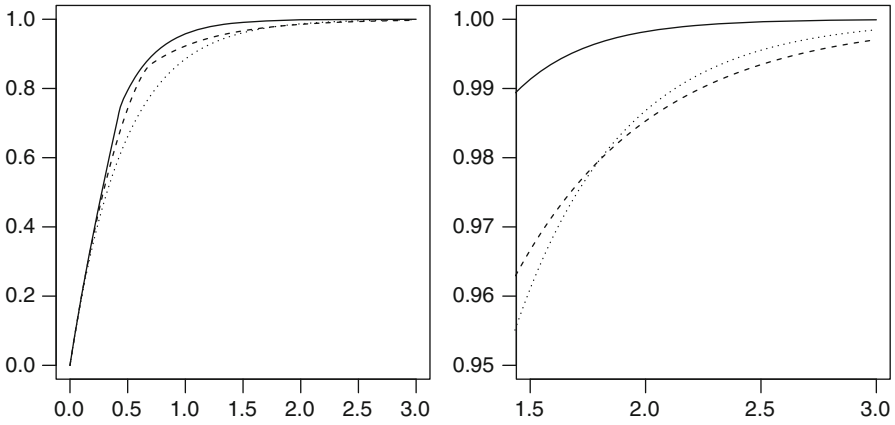


Fig. 20.1 x-Axis represents the t-values; y-axis represents the values of $\beta(t)$ (*solid line*), $W(t)$ (*dashed line*) and $L(t)$ (*dotted line*)

larger the probability $P_\theta\{n|\hat{\theta}_p - \theta| < t\}$, the better the estimator, and for all $t \geq 0$,

$$L(t) = \lim_{n\to\infty} P_\theta\{n|\hat{\theta}_{e/(1+e)} - \theta| < t\} = 1 - e^{-c(1+e^{-1})t}.$$

The upper bound $\beta(t)$ in this two-sided truncated exponential distribution family is given by

$$\beta(t) = \begin{cases} 1 - e^{-2ct} - (1 - 2e^{-ct})\sinh(t), & 0 \leq t \leq \log 2/c, \\ 1 - e^{-2ct}, & t > \log 2/c. \end{cases}$$

For ease of comparison, the plots of the function $\beta(t)$, $W(t)$ and $L(t)$ for $t \in [0, 3]$ are drawn in the left panel of Fig. 20.1. As expected, $\beta(t)$, denoted by the solid line, is uniformly higher than the other two curves. The curves of $W(t)$, denoted by the dashed line, and $L(T)$, denoted by the dotted line, show that no one is uniformly higher than the other. This is clearer from the right panel of Fig. 20.1, which magnifies the upper-right corner of the plot in the left panel. It is also evident that $W(t)$ is closer to the upper bound $\beta(t)$ than $L(t)$.

The estimator $T_n^*(\mathbf{X})$ defined in (20.11) only depends on the values of the density function f at both ends, and does not rely on the true form of f on (a, b), so it is an adaptive estimator. If both $f(a)$ and $f(b)$ are unknown, then we can estimate them by kernel technique. In fact, let K be a symmetric kernel density function, $K(a, b) = \int_0^{b-a} K(x)dx$, and define

$$\hat{f}_n(a) = \frac{1}{nhK(a,b)} \sum_{i=1}^{n} K\left(\frac{X_i - X_{(1)}}{h}\right), \quad \hat{f}_n(b) = \frac{1}{nhK(a,b)} \sum_{i=1}^{n} K\left(\frac{X_{(n)} - X_i}{h}\right).$$

Then one can show that $\hat{f}_n(a)$ and $\hat{f}_n(b)$ are consistent estimators of $f(a)$ and $f(b)$, respectively. Replace $f(a)$ and $f(b)$ in (20.11), one can get a two-sided adaptive admissible estimator of θ.

20.5 Proofs

Let's prove Lemma 1 first.

Proof of Lemma 1: Let us show (20.5) first. For convenience, let $Z_{ni} = \log \frac{f(X_i-\theta-t/n)}{f(X_i-\theta)} I_{A_{ni}}$. Then for each n, $\sum_{i=1}^n Z_{ni}$ is a sum of i.i.d. random variables no matter the true parameter is θ or $\theta + t/n$, conditioning on $a+\theta+t/n \le X_{(1)} \le X_{(n)} \le b+\theta$ and $a+\theta \le X_{(1)} \le X_{(n)} \le b+\theta-t/n$. Let's assume the true parameter is θ. Then conditioning $a+\theta+t/n \le X_{(1)} \le X_{(n)} \le b+\theta$, $X_i, i = 1, 2, \ldots, n$ are i.i.d. and have density function $cf(x-\theta)I(a+\theta+t/n \le x \le b+\theta)$, where $c^{-1} = P(a+\theta+t/n \le X \le b+\theta) = 1+O(1/n)$. In the following discussion, all expectations are calculated under this conditional distribution. For $k = 1, 2, 3$, we have

$$\begin{aligned}
EZ_{ni}^k &= c\int_{a+\theta+\frac{t}{n}}^{b+\theta} \left[\log \frac{f(x-\theta-t/n)}{f(x-\theta)}\right]^k f(x-\theta)dx \\
&= c\int_{a+\frac{t}{n}}^{b} \left[\log \frac{f(x-t/n)}{f(x)}\right]^k f(x)dx \\
&= c\int_{a+\frac{t}{n}}^{b} \left[\log f(x) - \frac{t}{n}\frac{f'(x)}{f(x)} + O\left(\frac{1}{n^2}\right) - \log f(x)\right]^k f(x)dx \\
&= \left(1+O\left(\frac{1}{n}\right)\right)\cdot\left[\int_{a+\frac{t}{n}}^{b}\left[-\frac{t}{n}\frac{f'(x)}{f(x)}\right]^k f(x)dx + O\left(\frac{1}{n^{2k}}\right)\right] \\
&= \frac{(-t)^k}{n^k}\int_a^b \left[\frac{f'(x)}{f(x)}\right]^k f(x)dx + O\left(\frac{1}{n^{k+1}}\right).
\end{aligned}$$

In particular,

$$EZ_{n1} = \frac{t[f(a)-f(b)]}{n} + O\left(\frac{1}{n^2}\right),\quad EZ_{n1}^2 = \frac{t^2 I}{n^2} + O\left(\frac{1}{n^3}\right),\quad EZ_{ni}^3 = O\left(\frac{1}{n^3}\right).$$

Therefore,

$$\frac{\sum_{i=1}^n E[Z_{ni}-EZ_{ni}]^3}{\left(\sqrt{\mathrm{Var}(\sum_{i=1}^n Z_{ni})}\right)^3} = \frac{nE[Z_{n1}-EZ_{n1}]^3}{\left(\sqrt{n\mathrm{Var}(Z_{ni})}\right)^3} = \frac{n\cdot O(n^{-3})}{[n\cdot O(n^{-2})]^{3/2}} = o(1).$$

By Lyapunov central limit theorem, we have

$$\frac{\sum_{i=1}^n Z_{ni} - nEZ_{n1}}{\sqrt{\mathrm{Var}(\sum_{i=1}^n Z_{ni})}} \Longrightarrow N(0,1)$$

in distribution. Further note that

$$\frac{nEZ_{n1} - t[f(a)-f(b)]}{\sqrt{\mathrm{Var}(\sum_{i=1}^n Z_{ni})}} = \frac{O(n^{-1})}{O(n^{-1/2})} = o(1),\quad \frac{\mathrm{Var}(\sum_{i=1}^n Z_{ni})}{n^{-1}\sigma^2(t)} \to 1,$$

we obtain that

$$\sqrt{n}\left[\sum_{i=1}^{n} Z_{ni} - t[f(a) - f(b)]\right] \Longrightarrow N(0, \sigma^2(t)).$$

So, (20.5) is proved, and hence (20.3). If the true parameter is $\theta + t/n$, then still $c = 1 + O(1/n)$. So for $k = 1, 2, 3$,

$$\begin{aligned} EZ_{ni}^k &= c\int_{a+\frac{t}{n}}^{b}\left[\log\frac{f(x-t/n)}{f(x)}\right]^k f(x-t/n)dx \\ &= c\int_{a+\frac{t}{n}}^{b}\left[\log f(x) - \frac{t}{n}\frac{f'(x)}{f(x)} + O\left(\frac{1}{n^2}\right) - \log f(x)\right]^k\left[f(x) + O\left(\frac{1}{n}\right)\right]dx \\ &= \int_{a+\frac{t}{n}}^{b}\left[-\frac{t}{n}\frac{f'(x)}{f(x)}\right]^k f(x)dx + O\left(\frac{1}{n^{k+1}}\right) \\ &= \frac{(-t)^k}{n^k}\int_{a}^{b}\left[\frac{f'(x)}{f(x)}\right]^k f(x)dx + O\left(\frac{1}{n^{k+1}}\right). \end{aligned}$$

Therefore, the proofs of (20.5) and (20.3) are the same as the one when θ is the true parameter. Similarly, one can show (20.6) and (20.4). The details are omitted for the sake of brevity. □

Proof of Theorem 1: First assume that $t > 0$. For a sample $X_1, X_2, \ldots, X_n$ from model (2), denote $C_n = \{X_i, i = 1, 2, \ldots, n : X_{(1)} < \theta_0 + t/n + a\}$ and A_n as in Lemma 20.1. By Neyman–Pearson lemma, the uniformly most powerful (UMP) test of the hypothesis $H_0 : \theta_n = \theta_0 + t/n$ versus $H_1 : \theta_0$ has the following form

$$\varphi_n(X) = \begin{cases} 1, & \text{if} X \in C_n \bigcup \left[A_n \bigcap \left\{\prod_{i=1}^n f(X_i - \theta_0) > k_n \prod_{i=1}^n f(X_i - \theta_n)\right\}\right] \\ r_n, & A_n \bigcap \left\{\prod_{i=1}^n f(X_i - \theta_0) = k_n \prod_{i=1}^n f(X_i - \theta_n)\right\} \\ 0, & \text{otherwise} \end{cases}$$

where k_n is chosen so that

$$\lim_{n\to\infty} E_{\theta_n}\varphi_n(X) = \frac{1}{2}. \tag{20.12}$$

Based on the asymptotic result (20.4) in Lemma 1, a proper k_n can be chosen so that $r_n = 0$, that is, a nonrandomized test can be constructed. Also $\lim_{n\to\infty} E_{\theta_n}\varphi_n(X)$ equals

$$\lim_{n\to\infty}\left[P_{\theta_n}(C_n) + P_{\theta_n}(A_n)P_{\theta_n}\left(\prod_{i=1}^{n} f(X_i - \theta_0) > k_n\prod_{i=1}^{n} f(X_i - \theta_n)\Big|A_n\right)\right] = \frac{1}{2}.$$

Since $P_{\theta_n}(C_n) = 0$, $\lim_{n\to\infty} P(A_n) = e^{-f(b)t}$, so, we can choose k_n so that

$$P_{\theta_n}\left(\prod_{i=1}^{n} f(X_i - \theta_0) > k_n\prod_{i=1}^{n} f(X_i - \theta_n)\Big|A_n\right) \to \frac{1}{2}e^{f(b)t}, \tag{20.13}$$

if $0 \le t \le (\log 2)/f(b)$. Also note that

$$\lim_{n\to\infty} P_{\theta_0}(C_n) = 1 - e^{-f(a)t}, \quad \lim_{n\to\infty} P_{\theta_0}(A_n) = e^{-f(a)t},$$

and by Lemma 20.1, (20.13) also holds when the true parameter is θ_0. Therefore,

$$\lim_{n\to\infty} E_{\theta_0}\varphi_n(X) = 1 - e^{-f(a)t} + \frac{1}{2}e^{[f(b)-f(a)]t}. \tag{20.14}$$

If $t > \log 2/f(b)$, a most powerful test can be defined as

$$\varphi_n(X) = \begin{cases} 1, & \text{if} X \in B_n \bigcup \left\{\theta_0 + t/n + a \le X_{(1)} \le X_{(n)} \le \theta_0 + u/n + b\right\} \\ 0, & \text{otherwise,} \end{cases}$$

where $0 < u < t$. Note that

$$\begin{aligned} E_{\theta_n}\varphi_n(X) &= P_{\theta_n}\left\{\theta_0 + t/n + a \le X_{(1)} \le X_{(n)} \le \theta_0 + u/n + b\right\} \\ &= \left[\int_{\theta_0+t/n+a}^{\theta_0+u/n+b} f(x - \theta_0 - t/n)dx\right]^n = \left[1 - \int_{(u-t)/n+b}^{b} f(x)dx\right]^n. \end{aligned}$$

So $\lim_{n\to\infty} E_{\theta_n}\varphi_n(X) = e^{-f(b)(t-u)}$. Choose $u = t - \log 2/f(b)$, then $\varphi_n(X)$ has the desired asymptotic level 1/2. Accordingly, with such a u, the asymptotic power of $\varphi_n(X)$ at θ_0 is given by

$$\lim_{n\to\infty} E_{\theta_0}\varphi_n(X) = 1. \tag{20.15}$$

For $t < 0$, the MPT for testing $H_0 : \theta_n = \theta_0 + t/n$ versus $H_1 : \theta_0$ has the following form

$$\varphi_n(X) = \begin{cases} 1, & \text{if} X \in D_n \bigcup \left[B_n \bigcap \left\{\prod_{i=1}^n f(X_i - \theta_0) > k_n \prod_{i=1}^n f(X_i - \theta_n)\right\}\right] \\ r_n, & B_n \bigcap \left\{\prod_{i=1}^n f(X_i - \theta_0) = k_n \prod_{i=1}^n f(X_i - \theta_n)\right\} \\ 0, & \text{otherwise,} \end{cases}$$

where $D_n = \{X_i : i = 1, 2, \ldots, n : X_{(n)} > \theta_0 + t/n + b\}$, and $B_n = \{X_i : i = 1, 2, \ldots, n : \theta_0 + a \le X_{(1)} \le X_{(n)} \le \theta_0 + t/n + b\}$. A similar argument as before shows that, under the constraint (20.12), for $-\log 2/f(a) \le t \le 0$, the power function satisfies

$$\lim_{n\to\infty} E_{\theta_0}\varphi_n(X) = 1 - e^{f(b)t} + \frac{1}{2}e^{[f(b)-f(a)]t}. \tag{20.16}$$

If $t < -\log 2/f(a)$, we can choose $u = t + \log 2/f(a)$ and the following test

$$\varphi_n(X) = \begin{cases} 1, & \text{if} X \in D_n \bigcup \left\{\theta_0 + u/n + a \le X_{(1)} \le X_{(n)} \le \theta_0 + t/n + b\right\} \\ 0, & \text{otherwise,} \end{cases}$$

is most powerful with the asymptotic level 1/2 and the asymptotic power 1. This, together with the results (20.14), (20.16), and (20.15), completes the proof of Theorem 1. □

Proof of Corollary 1: The form of $\beta(t)$ is an immediate consequence of Theorem 1 with $f(a) = f(b)$ in (20.7). When $f(a) = f(b)$, we can also show that

$$\lim_{n\to\infty} P_\theta\{n(\hat{\theta}_{1/2} - \theta) \le t\} = \begin{cases} \frac{1}{2}e^{2\,f(a)t}, & t < 0, \\ 1 - \frac{1}{2}e^{-2\,f(a)t}, & t \ge 0. \end{cases}$$

It is easy to see that $\hat{\theta}_{1/2}$ is an AMU estimator, and for $0 < t < \log 2/f(a)$,

$$\lim_{n\to\infty} P_\theta\{n(\hat{\theta}_{1/2} - \theta) \le t\} = 1 - \frac{1}{2}e^{-2\,f(a)t} < 1.5 - e^{-f(a)t},$$

and for $t \ge \log 2/f(a)$, $\lim_{n\to\infty} P_\theta\{n(\hat{\theta}_{1/2} - \theta) \le t\} = 1 - \frac{1}{2}e^{-2\,f(a)t} < 1$. For $-\log 2/f(a) < t < 0$,

$$\lim_{n\to\infty} P_\theta\{n(\hat{\theta}_{1/2} - \theta) \ge t\} = 1 - \frac{1}{2}e^{2\,f(a)t} < 1.5 - e^{f(a)t},$$

and for $t < -\log 2/f(a)$, $\lim_{n\to\infty} P_\theta\{n(\hat{\theta}_{1/2} - \theta) \ge t\} = 1 - \frac{1}{2}e^{2\,f(a)t} < 1$. That is, for all $t > 0$, $\lim_{n\to\infty} P_\theta\{n(\hat{\theta}_{1/2} - \theta) \le t\} < \beta(t)$, and for all $t < 0$, $\lim_{n\to\infty} P_\theta\{n(\hat{\theta}_{1/2} - \theta) \ge t\} < \beta(t)$. One can see that the equality holds only at $t = 0$. This concludes the proof of Corollary 1. □

Proof of Theorem 2: Without loss of generality, we only prove the result for $f(a) > f(b) > 0$.

The proof will be divided into three parts based on $\lambda_n < 0$, $\lambda_n = 0$ and $\lambda_n > 0$ in (20.8). Since the proofs are similar, only the proof for $\lambda_n > 0$ is present here for the sake of brevity. Denote

$$A_n = \{X_{(1)} < \theta_0 + a, X_{(n)} < \theta_2 + b\}, \quad B_n = \{\theta_0 + a < X_{(1)} < \theta_1 + a, X_{(n)} < \theta_0 + b\},$$

$$C_n = \{X_{(1)} > \theta_1 + a, \theta_2 + b < X_{(n)} < \theta_0 + b\}, \quad D_n = \{\theta_1 + a < X_{(1)} \le X_{(n)} < \theta_2 + b\},$$

and $E_n = \{\theta_1 + a < X_{(1)} \le X_{(n)} < \theta_0 + b\}$. Then, $\phi_n^* = 1$ if and only if $X_1, X_2, \ldots, X_n$ belongs to

$$A_n \bigcup B_n \bigcup \left[D_n \bigcap \left\{ \frac{\prod_{i=1}^n f(X_i - \theta_2)}{\prod_{i=1}^n f(X_i - \theta_0)} - \frac{\prod_{i=1}^n f(X_i - \theta_1)}{\prod_{i=1}^n f(X_i - \theta_0)} > \lambda_n \right\} \right]$$

$$\bigcup \left[C_n \bigcap \left\{ \frac{\prod_{i=1}^n f(X_i - \theta_1)}{\prod_{i=1}^n f(X_i - \theta_0)} < -\lambda_n \right\} \right].$$

Similarly to the proof of Lemma 1, we can show that under θ_0, θ_1, and θ_2, in probability,

$$\frac{\prod_{i=1}^n f(X_i - \theta_2)}{\prod_{i=1}^n f(X_i - \theta_0)} I[\theta_1 + a < X_i < \theta_2 + b] \to e^{-tf^-},$$

$$\frac{\prod_{i=1}^n f(X_i - \theta_1)}{\prod_{i=1}^n f(X_i - \theta_0)} I[\theta_1 + a < X_i < \theta_2 + b] \to e^{tf^-}.$$

as $n \to \infty$. This is also true after replacing $I[\theta_1 + a < X_i < \theta_2 + b]$ with $I[\theta_1 + a < X_i < \theta_0 + b]$. After certain normalization, asymptotical normalities can also be achieved as in Lemma 1.

If $t < \log 2/f(a)$ and $e^{-f(a)t} - e^{-tf^+} > 1/2$, we can choose λ_n such that $\lim_{n\to\infty} \lambda_n = -e^{-tf^-}$ and

$$\lim_{n\to\infty} P_{\theta_0}\left\{\frac{\prod_{i=1}^n f(X_i-\theta_1)}{\prod_{i=1}^n f(X_i-\theta_0)} \le -\lambda_n \middle| C_n\right\} = \frac{1-e^{f(a)t}/2 - e^{-f(b)t}}{1-e^{-f(b)t}}.$$

Since $e^{tf^-} - e^{-tf^-} < e^{tf^-}$, with such a choice of λ_n, we will have

$$\lim_{n\to\infty} P_{\theta_0}\left\{\frac{\prod_{i=1}^n f(X_i-\theta_2)}{\prod_{i=1}^n f(X_i-\theta_0)} - \frac{\prod_{i=1}^n f(X_i-\theta_1)}{\prod_{i=1}^n f(X_i-\theta_0)} > \lambda_n \middle| D_n\right\} = 1.$$

Therefore,

$$\begin{aligned}\lim_{n\to\infty} E_{\theta_0}\phi_n^*(X) &= 1 - e^{-f(a)t} \\ &\quad + e^{-tf^+} \cdot \lim_{n\to\infty} P_{\theta_0}\left\{\frac{\prod_{i=1}^n f(X_i-\theta_2)}{\prod_{i=1}^n f(X_i-\theta_0)} - \frac{\prod_{i=1}^n f(X_i-\theta_1)}{\prod_{i=1}^n f(X_i-\theta_0)} > \lambda_n \middle| D_n\right\} \\ &\quad + \left[e^{-f(a)t} - e^{-tf^+}\right] \cdot \lim_{n\to\infty} P_{\theta_0}\left\{\frac{\prod_{i=1}^n f(X_i-\theta_1)}{\prod_{i=1}^n f(X_i-\theta_0)} \le -\lambda_n \middle| C_n\right\} \\ &= \frac{1}{2}.\end{aligned}$$

It is also easy to check that

$$\begin{aligned}&\lim_{n\to\infty} (E_{\theta_2}\phi_n^*(X) - E_{\theta_1}\phi_n^*(X)) = 1 - e^{-f(a)t} + e^{-f(a)t} - e^{-2\,f(a)t} \\ &+ (e^{-2(a)t} - e^{-2\,f(b)t}) \cdot \lim_{n\to\infty} P_{\theta_1}\left\{\frac{\prod_{i=1}^n f(X_i-\theta_2)}{\prod_{i=1}^n f(X_i-\theta_0)} - \frac{\prod_{i=1}^n f(X_i-\theta_1)}{\prod_{i=1}^n f(X_i-\theta_0)} > \lambda_n \middle| D_n\right\} \\ &- (e^{-f(b)t} - e^{-2\,f(b)t}) \cdot \lim_{n\to\infty} P_{\theta_1}\left\{\frac{\prod_{i=1}^n f(X_i-\theta_1)}{\prod_{i=1}^n f(X_i-\theta_0)} \le -\lambda_n \middle| C_n\right\} \\ &= 1 - e^{-f(b)t} + \frac{1}{2}e^{tf^-}. \qquad (20.17)\end{aligned}$$

If $t < \log 2/f(a)$ and $e^{-f(a)t} - e^{-tf^+} \le 1/2$, we can choose λ_n such that $\lim_{n\to\infty} \lambda_n = e^{tf^-} - e^{-tf^-}$ and

$$\lim_{n\to\infty} P_{\theta_0}\left\{\frac{\prod_{i=1}^n f(X_i-\theta_2)}{\prod_{i=1}^n f(X_i-\theta_0)} - \frac{\prod_{i=1}^n f(X_i-\theta_1)}{\prod_{i=1}^n f(X_i-\theta_0)} > \lambda_n \middle| D_n\right\} = e^{f(b)t} - \frac{1}{2}e^{tf^+}.$$

Since $e^{tf^-} - e^{-tf^-} < e^{tf^-}$, with such a choice of λ_n, we will have

$$\lim_{n\to\infty} P_{\theta_0}\left\{\frac{\prod_{i=1}^n f(X_i-\theta_1)}{\prod_{i=1}^n f(X_i-\theta_0)} \le -\lambda_n \middle| C_n\right\} = 0.$$

Therefore, we still have $\lim_{n\to\infty} E_{\theta_0}\phi_n^*(X) = 1/2$, but

$$\lim_{n\to\infty} (E_{\theta_2}\phi_n^*(X) - E_{\theta_1}\phi_n^*(X)) = 1 - e^{-2\,f(a)t} + (1-2e^{-f(a)t})\sinh[tf^-]. \quad (20.18)$$

The similar argument can be extended to the cases of $\lambda_n = 0$ and $\lambda_n > 0$ by looking for a proper critical function ϕ_n^*. For the sake of brevity, the proofs are left out. □

Remark: If the asymptotic distribution of the likelihood ratios, after normalization, are degenerated, then for $t < \log 2/f(a)$ and $e^{-f(a)t} - e^{-tf^+} > 1/2$, we can choose $\phi_n^*(X) = 1 - I[F_n \cup G_n]$, where $F_n = \{X_{(1)} > \theta_0 + a + u/n, \theta_2 + b < X_{(n)} < \theta_0 + b\}$, $G_n = \{X_{(1)} > \theta_0 + a, \theta_0 + b < X_{(n)} < \theta_1 + b\}$, and $0 < u < t$ such that $E_{\theta_0}\phi_n^* \to 1/2$. Then we also have (20.17). If $t < \log 2/f(a)$ and $e^{-f(a)t} - e^{-tf^+} \le 1/2$, then we can choose $\phi_n(X)$ to be the indicator function of the union of three sets: $\{X_{(1)} < \theta_0 + a, X_{(n)} < \theta_2 + b\}$, $G_n = \{\theta_0 + a < X_{(1)} < \theta_1 + a, X_{(n)} < \theta_0 + b\}$ and $\{\theta_1 + a < X_{(1)} < \theta_0 + u/n + a, X_{(n)} < \theta_2 + b\}$, where u is chosen so that $u \ge t$ and $E_{\theta_0}\phi_n^* \to 1/2$. With such choices of ϕ_n^* and u, we can obtain (20.18).

Proof of Theorem 3: By Fox and Rubin (1964), the roof $T_n(\mathbf{X})$ of Eq. (20.10) is an AMU and an admissible estimator of θ under the absolute deviation loss $L(\theta, t) = |t - \theta|$. Suppose there is another AMU estimator $\hat{\theta}_n$ of θ such that

$$\liminf_{n\to\infty} P_\theta\{n|\hat{\theta}_n - \theta| < t\} \ge \limsup_{n\to\infty} P_\theta\{n|T(X_{(1)}, X_{(n)}) - \theta| < t\}$$

or equivalently,

$$\limsup_{n\to\infty} P_\theta\{n|\hat{\theta}_n - \theta| > t\} \le \liminf_{n\to\infty} P_\theta\{n|T(X_{(1)}, X_{(n)}) - \theta| > t\}$$

holds for all $\theta \in \Theta$ and $t > 0$. Moreover, for each $\theta \in \Theta$, there exists a set A_θ with positive Lebesgue measure such that the strict inequality holds for all $t \in A_\theta$. Therefore,

$$\begin{aligned}
&\limsup_{n\to\infty} E_\theta[n|\hat{\theta}_n - \theta|] = \limsup_{n\to\infty} \int_0^\infty P_\theta\{n|\hat{\theta}_n - \theta| > t\}dt \\
&\le \int_0^\infty \limsup_{n\to\infty} P_\theta\{n|\hat{\theta}_n - \theta| > t\}dt < \int_0^\infty \liminf_{n\to\infty} P_\theta\{n|T_n(\mathbf{X}) - \theta| > t\}dt \\
&\le \liminf_{n\to\infty} \int_0^\infty P_\theta\{n|T_n(\mathbf{X}) - \theta| > t\}dt = \liminf_{n\to\infty} E_\theta[n|T_n(\mathbf{X}) - \theta|]
\end{aligned}$$

which implies, when n is large enough, $E_\theta|\hat{\theta}_n - \theta| < E_\theta|T_n(\mathbf{X}) - \theta|$. This contradicts the admissibility of $T_n(\mathbf{X})$, and therefore, $T_n(\mathbf{X})$ must be an asymptotically weak admissible median unbiased estimator.

To show the equivalence of $T_n(\mathbf{X})$ and $T_n^*(\mathbf{X})$ defined in (20.11), note that by Taylor expansion

$$\frac{\prod_{i=1}^n f(X_i - \theta)}{\prod_{i=1}^n f(X_i)} = \exp\left[\sum_{i=1}^n \log \frac{f(X_i - \theta)}{f(X_i)}\right] = \exp\left[\theta \sum_{i=1}^n \frac{f'(X_i)}{f(X_i)} + o_p(n\theta)\right],$$

then from $\theta \in [X_{(n)} - b, X_{(1)} - a]$ which implies that $\theta = O_p(1/n)$, then $o_p(n\theta) = o_p(1)$, therefore, we obtain

$$\begin{aligned}\frac{1}{2} &= \frac{\int_{X_{(n)}-b}^{T_n(\mathbf{X})} \prod_{i=1}^n f(X_i - \theta)d\theta}{\int_{X_{(n)}-b}^{X_{(1)}-a} \prod_{i=1}^n f(X_i - \theta)d\theta} \\ &= \frac{\int_{X_{(n)}-b}^{T_n(\mathbf{X})} \exp[-\theta \sum_{i=1}^n f'(X_i)/f(X_i)]d\theta}{\int_{X_{(n)}-b}^{X_{(1)}-a} \exp[-\theta \sum_{i=1}^n f'(X_i)/f(X_i)]d\theta}[1 + o_p(1)].\end{aligned}$$

Denote $M_n(X) = n^{-1} \sum_{i=1}^n f'(X_i)/f(X_i)$, then the above equality indeed is equivalent to

$$\frac{\exp[-n(X_{(n)} - b)M_n(X)] - \exp[-nT_n(\mathbf{X})M_n(X)]}{\exp[-n(X_{(n)} - b)M_n(X)] - \exp[-n(X_{(1)} - a)M_n(X)]}[1 + o_p(1)] = \frac{1}{2},$$

or

$$e^{-nT_n(\mathbf{X})M_n(X)} = \frac{1}{2}\left[e^{-n(X_{(n)}-b)M_n(X)} + e^{-n(X_{(1)}-a)M_n(X)}\right] + o_p(1).$$

The the desired result follows from the facts that $M_n(X) \to E_0[f'(X)/f(X)] = f(b) - f(a)$ in probability.

The proof of the theorem will be complete if we can show that $T_n(\mathbf{X})$ defined above is not two-sided asymptotic efficient. For this purpose, the asymptotic distribution of $T_n(\mathbf{X})$ should be derived. The equivalence of $T_n(\mathbf{X})$ and $T_n^*(\mathbf{X})$ implies that it suffices to discuss $T_n^*(\mathbf{X})$ only. Without loss of generality, let us assume that $\theta = 0$. It is easy to see that

$$P_0\left\{2^{-1}e^{nk[X_{(n)}-b]} < x\right\} = \begin{cases} 0, & x < 0, \\ \left[1 + \frac{f(b)\log 2x}{nk} + o\left(\frac{1}{n}\right)\right]^n \to (2x)^{f(b)/k}, & 0 \le x \le 1/2, \\ 1, & x > 1/2. \end{cases}$$

and

$$P_0\left\{2^{-1}e^{nk[X_{(1)}-a]} < x\right\} = \begin{cases} 1 - \left[1 - \frac{f(a)\log 2x}{nk} + o\left(\frac{1}{n}\right)\right]^n \to (2x)^{-f(a)/k}, & x \ge 1/2, \\ 0, & x < 1/2. \end{cases}$$

Based on these two probabilities, and also the asymptotic independence of $X_{(1)}$ and $X_{(n)}$, we can obtain

$$\begin{aligned}&\lim_{n\to\infty} P_\theta\{n|T_n(\mathbf{X}) - \theta| < t\} = \lim_{n\to\infty} P_\theta\{nk|T_n^*(\mathbf{X}) - \theta| < kt\} \\ = &\begin{cases} 1 - \frac{1}{2e^{kt}}(2e^{kt} - 1)^{-f(b)/f^-} - \frac{1}{2e^{-kt}}(2e^{-kt} - 1)^{f(a)/f^-}, & 0 \le t \le \frac{\log 2}{k}, \\ 1 - \frac{1}{2e^{kt}}(2e^{kt} - 1)^{-f(b)/f^-}, & t > \frac{\log 2}{k}. \end{cases}\end{aligned}$$

We can check that this limit indeed does not exceed the $\beta(t)$ function for all $t \ge 0$, and when t is big enough, it is also easy to see that this limit is strictly less than

$\beta(t)$. That is, we proved that $T_n(\mathbf{X})$ is not two-sided asymptotically efficient. This, together with the fact that $T_n(\mathbf{X})$ is an asymptotically weak admissible median unbiased estimator, implies that there is no AMU in the location family (20.2) to be two-sided asymptotically efficient. □

Acknowledgement Dr. Hira L. Koul is always an outstanding scientist in my eyes. I am truly honored and lucky to be one of his many Ph.D. students, and no words can express my sincere gratitude for his academic guidance in my career.

References

Akahira M (1982) Asymptotic optimality of estimators in non-regular cases. Ann Inst Statist Math PartA 34:69–82

Akahira M, Takeuchi K(1981) Asymptotic efficiency of statistical estimators concepts and higher order asymptotic efficiency. Lecture Notes in Statistics, (7). Springer-Verlag

Akahira M Takeuchi K (2003) Joint statistical papersof AkahiraandTakeuchi. World Scientific Pub. Co. Inc.

Bickel PJ, Klaassen Chris AJ, Ritov Y, Wellner JA (1998) Efficient and adaptive estimation for semiparametric models. Springer,

Fox M Rubin H (1964) Admissibility of quantile estimates of a single location parameter. Ann Math Statist 35:1019–1031

Ibragimov IA Has'minskii RZ (1981) Statistical estimation, asymptotic theory. Applacations of Mathetmatics, Volume 16, Springer

Takeuchi K (1974) Asymptotic theoryof statistical inference. Education Press, Tokyo

Weiss L Wolfowitz J (1968) Generalized maxium likelihood estimatorsina particular case. Theory Prob. Applications 13:622–627

Chapter 21
Semiparametric Analysis of Treatment Effect via Failure Probability Ratio and the Ratio of Cumulative Hazards

Song Yang

21.1 Introduction

For clinical trials with time to event data, often proportional hazards (Cox 1972) is assumed when comparing two treatment arms, and a single value of the hazard ratio is used to describe the group difference. When the proportionality assumption may not hold true, a natural approach to assess the time-dependency of the treatment effect is to analyze the hazard ratio function. For example, a conventional method is to give a hazard ratio estimate over each of a few time intervals, by fitting a piecewise proportional hazards model. Alternatively, a "defined" time-varying covariate can be used in a Cox regression model, resulting in a parametric form for the hazard ratio function (e.g., Kalbfleisch and Prentice 2002, Chap. 6). With these approaches, it may not be easy to pre-specify the partition of the time axis or the parametric form of the hazard ratio function. Also, although the hazard ratio provides a nice display of temporal pattern of the treatment effect, it may not directly translate to the survival experience. It is possible for the hazard ratio to be less than 1 in a region where there is no improvement in the survival probability, or more than 1 in a region where the survival probability is not reduced. Similar phenomena also exists for the average of hazard ratio. Thus to assess the cumulative treatment effect, other measures can be used to supplement the hazard ratio.

Let $F_T(t)$ and $F_C(t)$ be the cumulative distribution functions of the two comparison groups, named treatment and control, respectively. The failure probability ratio

$$RR(t) = \frac{F_T(t)}{F_C(t)}$$

is the process version of relative risk, a measure often used in epidemiology. It directly indicates if the failure probability in the time interval $(0, t]$ is lower in the treatment group than in the control group, regardless of the possible up and down pattern

S. Yang (✉)
Office of Biostatistics Research, National Heart, Lung, and Blood Institute,
6701 Rockledge Dr. MSC 7913, Bethesda, MD, 20892, USA
e-mail: yangso@nhlbi.nih.gov

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_21,

of the hazard ratio within $(0, t]$. Let $\Lambda_T(t)$ and $\Lambda_C(t)$ be the cumulative distribution functions of the two comparison groups respectively. The ratio of cumulative hazards

$$CHR(t) = \frac{\Lambda_T(t)}{\Lambda_C(t)}$$

also indicates the cumulative treatment effect, taking value < 1 if and only if $F_T(t) < F_C(t)$. Unlike the failure probability ratio, a value 0.8 for the ratio of cumulative hazards does not translate to a 20 % reduction of the failure probability. However, there is a nice property that if one adopts a proportional hazards adjustment for baseline covariates, then the ratio of cumulative hazards remains the same while the failure probability ratio depends on those covariates.

Although measures such as the failure probability ratio and the ratio of cumulative hazards provide usual supplementary information in addition to the hazard ratio, and the non-parametric estimators are easily available via the Nelson–Aalen estimator for the cumulative hazard function (Nelson 1969; Aalen 1975) and the Kaplan–Meier estimator of the survival function (Kaplan and Meier 1958), the non-parametric inference procedures are not used frequently, as the estimates are often not very smooth and the confidence intervals can be quite wide near the beginning of the data range. In this chapter, we consider semiparametric inference on the two ratios under a sufficiently flexible model. Assume that the failure times are absolutely continuous. The short-term and long-term hazards model proposed in Yang and Prentice (2005) postulates that

$$\lambda_T(t) = \frac{1}{e^{-\beta_2} + \left(e^{-\beta_1} - e^{-\beta_2}\right) S_C(t)} \lambda_C(t), \; t < \tau_0, \tag{21.1}$$

where β_1, β_2 are scalar parameters, S_C the survivor function of the control group, $\lambda_T(t)$, $\lambda_C(t)$ the hazard function of the two groups respectively, and

$$\tau_0 = \sup \left\{ x : \int_0^x \lambda_C(t) dt < \infty \right\}. \tag{21.2}$$

Under this model, $\lim_{t \downarrow 0} \lambda_T(t)/\lambda_C(t) = e^{\beta_1}$, $\lim_{t \uparrow \tau_0} \lambda_T(t)/\lambda_C(t) = e^{\beta_2}$. Thus, various patterns of the hazard ratio can be realized, including proportional hazards, no initial effect, disappearing effect, and crossing hazards. In particular, model (21.1) includes the proportional hazards model and the proportional odds model as special cases. There is no need to pre-specify a partition of the time axis or a parametric form of the hazard ratio function. For this model, Yang and Prentice (2005) proposed a pseudo-likelihood method for estimating the parameters, and Yang and Prentice (2011) studied inference procedures on the hazard ratio function and the average of the hazard ratio function. Extension of model (21.1) to the regression setting was also studied for current status data in Tong et al. (2007).

In the sections to follow, we first obtain the estimates and point-wise confidence intervals of the two ratios under model (21.1). Since the ratios are functions of time, simultaneous confidence intervals, or confidence bands, of the ratios are more

appropriate than the point-wise confidence intervals. We will employ a resampling scheme to obtain the confidence bands of the ratios. Such semiparametric inference procedures are applicable in a wide range of applications due to the properties of model (21.1) mentioned before. They will be illustrated through applications to data from two clinical trials.

Some previous work is related to the problems considered here. Dong and Matthews (2012) developed empirical likelihood estimator for the ratio of cumulative hazards with covariate adjustment. Schaubel and Wei (2011) considered several measures under dependent censoring and non-proportional hazards, and point-wise confidence intervals were constructed. In earlier works, Dabrowska et al. (1989) introduced a relative change function defined in terms of cumulative hazards and found simultaneous bands for this function under the assumption of proportional hazards. Parzen et al. (1997) constructed nonparametric simultaneous confidence bands for the survival probability difference. Cheng et al. (1997) proposed point-wise and simultaneous confidence interval procedures for the survival probability under semiparametric transformation models. McKeague and Zhao (2002) proposed simultaneous confidence bands for ratios of survival functions via the empirical likelihood method.

The article is organized as follows. In Sect. 21.2 the short-term and long-term hazard ratio model and the parameter estimator are described. Point-wise confidence intervals are established for the failure probability ratio and the ratio of cumulative hazards. In Sect. 21.3, confidence bands are developed. Simulation results are presented in Sect. 21.4. Applications to data from two clinical trials are given in Sect. 21.5. Some discussion is given in Sect. 21.6.

21.2 The Estimators and Point-Wise Confidence Intervals

Denote the pooled lifetimes of the two groups by $T_1, \cdots, T_n$, with $T_1, \cdots, T_{n_1}$, $n_1 < n$, constituting the control group. Let $C_1, \cdots, C_n$ be the censoring variables, and $Z_i = I(i > n_1)$, $i = 1, \cdots, n$, where $I(\cdot)$ is the indicator function. The available data consist of the independent triplets (X_i, δ_i, Z_i), $i = 1, \ldots, n$, where $X_i = \min(\mathrm{T_i}, \mathrm{C_i})$ and $\delta_i = I(T_i \leq C_i)$. We assume that T_i, C_i are independent given Z_i. The censoring variables (C_i's) need not be identically distributed, and in particular the two groups may have different censoring patterns. For $t < \tau_0$ with τ_0 defined in (21.2), let $R(t)$ be the odds function $1/S_C(t) - 1$ of the control group. The model of Yang and Prentice (2005) can be expressed as

$$\lambda_i(t) = \frac{1}{e^{-\beta_1 Z_i} + e^{-\beta_2 Z_i} R(t)} \frac{dR(t)}{dt}, \; i = 1, \ldots, n, t < \tau_0,$$

where $\lambda_i(t)$ is the hazard function for T_i given Z_i.

Under model (21.1), $RR(t)$ and $CHR(t)$ depends on the parameter $\beta = (\beta_1, \beta_2)^T$ and the baseline function $R(t)$, where "^{T}" denotes transpose. Yang and Prentice (2005) studied a pseudo likelihood estimator $\hat{\beta}$ of β which we describe below.

Let $\tau < \tau_0$ be such that

$$\lim_n \sum_{i=1}^{n} I\big(X_i \geq \tau\big) > 0, \tag{21.3}$$

with probability 1. For $t \leq \tau$, define

$$\begin{aligned}
\hat{P}(t;\mathbf{b}) &= \prod_{s\leq t}\left(1 - \frac{\sum_{i=1}^{n}\delta_i e^{-b_2 Z_i} I(X_i = s)}{\sum_{i=1}^{n} I(X_i \geq s)}\right), \\
\hat{R}(t;\mathbf{b}) &= \frac{1}{\hat{P}(t;\mathbf{b})}\int_0^t \frac{\hat{P}_-(s;\mathbf{b})}{\sum_{i=1}^{n} I\big(X_i \geq s\big)} d\left(\sum_{i=1}^{n}\delta_i e^{-b_1 Z_i} I\big(X_i \leq s\big)\right),
\end{aligned}$$

where $\hat{P}_-(s;\mathbf{b})$ denotes the left continuous (in s) version of $\hat{P}(s;\mathbf{b})$. Let $L(\beta, R)$ be the likelihood function of β under model (21.1) when the function $R(t)$ is known, with the corresponding score vector $S(\beta, R) = \partial \ln L(\beta, R)/\partial\beta$. Define $Q(\mathbf{b}) = S(\mathbf{b}, R)|_{R(t)=\hat{R}(t;\mathbf{b})}$. Then the pseudo maximum likelihood estimator $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)^T$ of β is the zero of $Q(\mathbf{b})$. Note that the use of $\hat{R}(t;\mathbf{b})$ results in the estimating function $Q(\mathbf{b})$ which does not involve the infinite dimensional nuisance parameter $R(t)$, thus the finite dimensional parameter β can be estimated much more easily.

Once $\hat{\beta}$ is obtained, $R(t)$ can be estimated by $\hat{R}\big(t;\hat{\beta}\big)$. Thus under model (21.1), plugging-in the estimators $\hat{\beta}$ and $\hat{R}\big(t;\hat{\beta}\big)$, we can estimate the failure probability ratio $RR(t)$ and the ratio of cumulative hazards $CHR(t)$ by

$$\widehat{RR}(t) = \frac{1 + \hat{R}\big(t;\hat{\beta}\big)}{\hat{R}\big(t;\hat{\beta}\big)}\left(1 - \left\{1 + e^{-\hat{\beta}_2+\hat{\beta}_1}\hat{R}\big(t;\hat{\beta}\big)\right\}^{-e^{\hat{\beta}_2}}\right), \tag{21.4}$$

and

$$\widehat{CHR}(t) = \frac{e^{\hat{\beta}_2}\ln\left\{1 + e^{-\hat{\beta}_2+\hat{\beta}_1}\hat{R}\big(t;\hat{\beta}\big)\right\}}{\ln\left\{1 + \hat{R}\big(t;\hat{\beta}\big)\right\}}, \tag{21.5}$$

respectively. Note that under the model and with the pseudo likelihood estimator, the distributions of the two groups share a common baseline function $R(t)$ which is estimated using pooled data. Thus the resulting estimators for $RR(t)$ and $CHR(t)$ are expected to be smoother and more stable than the nonparametric estimators. In Appendix A, we show that, under certain regularity conditions, the two estimators in (21.4) and (21.5) are strongly consistent under model (21.1). To study the distributional properties of the estimators, let

$$U_n(t) = \sqrt{n}(\widehat{RR}(t) - RR(t)),\ t \leq \tau,$$

$$V_n(t) = \sqrt{n}(\widehat{CHR}(t) - CHR(t)),\ t \leq \tau,$$

and

$$\Omega = \Big\{ -\frac{1}{n}\frac{\partial Q(\beta)}{\partial \beta} \Big\}^{-1}.$$

Let $\hat{\Omega}$ be an estimator of Ω defined by replacing β with $\hat{\beta}$ and $R(t)$ with $\hat{R}\big(t;\hat{\beta}\big)$.

In Appendix B we show that, for $t \le \tau$, the processes U_n and V_n are asymptotically equivalent to, respectively,

$$\begin{aligned} \tilde{U}_n(t) &= \frac{A_{RR}^T(t)\Omega}{\sqrt{n}} \left(\sum_{i \le n_1} \int_0^\tau \mu_1 dM_i + \sum_{i > n_1} \int_0^\tau \mu_2 dM_i \right) \\ &\quad + \frac{B_{RR}(t)}{\sqrt{n}} \left(\sum_{i \le n_1} \int_0^t \nu_1 dM_i + \sum_{i > n_1} \int_0^t \nu_2 dM_i \right) \end{aligned} \tag{21.6}$$

and

$$\begin{aligned} \tilde{V}_n(t) &= \frac{A_{CHR}^T(t)\Omega}{\sqrt{n}} \left(\sum_{i \le n_1} \int_0^\tau \mu_1 dM_i + \sum_{i > n_1} \int_0^\tau \mu_2 dM_i \right) \\ &\quad + \frac{B_{CHR}(t)}{\sqrt{n}} \left(\sum_{i \le n_1} \int_0^t \nu_1 dM_i + \sum_{i > n_1} \int_0^t \nu_2 dM_i \right), \end{aligned} \tag{21.7}$$

where A_{RR}, A_{CHR}, μ_1, μ_2 are appropriately defined 2×1 vector functions and B_{RR}, B_{CHR}, ν_1, ν_2 scalar functions given in Appendix B. It will then be shown that U_n and V_n converge weakly to some zero-mean Gaussian processes U^* and V^* respectively. With estimators $\hat{B}_{RR}(t)$, $\hat{A}_{RR}(t)$, ... , given in Appendix B, it will be shown that the limiting covariance functions of U^* and V^* can be consistently estimated, respectively, by

$$\begin{aligned} \hat{\sigma}_{RR}(s,t) = \hat{A}_{RR}^T(s)\hat{\Omega} &\left(\int_0^\tau \frac{\hat{\mu}_1(w)\hat{\mu}_1^T(w)K_1(w)d\hat{R}\big(w;\hat{\beta}\big)}{n\big(1+\hat{R}\big(w;\hat{\beta}\big)\big)} \right. \\ &\left. + \int_0^\tau \frac{\hat{\mu}_2(w)\hat{\mu}_2^T(w)K_2(w)d\hat{R}\big(w;\hat{\beta}\big)}{n\big(e^{-\hat{\beta}_1}+e^{-\hat{\beta}_2}\hat{R}\big(w;\hat{\beta}\big)\big)} \right) \hat{\Omega}^T \hat{A}_{RR}(t) \\ &+ \hat{B}_{RR}(s)\hat{B}_{RR}(t) \left(\int_0^s \frac{\hat{\nu}_1^2(w)K_1(w)d\hat{R}\big(w;\hat{\beta}\big)}{n\big(1+\hat{R}\big(w;\hat{\beta}\big)\big)} \right. \\ &\left. + \int_0^s \frac{\hat{\nu}_2^2(w)K_2(w)d\hat{R}\big(w;\hat{\beta}\big)}{n\big(e^{-\hat{\beta}_1}+e^{-\hat{\beta}_2}\hat{R}\big(w;\hat{\beta}\big)\big)} \right) \\ &+ \hat{B}_{RR}(t)\hat{A}_{RR}^T(s)\hat{\Omega} \left(\int_0^t \frac{\hat{\mu}_1(w)\hat{\nu}_1(w)K_1(w)d\hat{R}\big(w,\hat{\beta}\big)}{n\big(1+\hat{R}\big(w;\hat{\beta}\big)\big)} \right. \end{aligned}$$

$$+\int_0^t \frac{\hat{\mu}_2(w)\hat{v}_2(w)K_2(w)d\hat{R}\big(w,\hat{\beta}\big)}{n\big(e^{-\hat{\beta}_1}+e^{-\hat{\beta}_2}\hat{R}\big(w;\hat{\beta}\big)\big)}\Bigg)$$
$$+\hat{B}_{RR}(s)\hat{A}_{RR}^T(t)\hat{\Omega}\left(\int_0^s \frac{\hat{\mu}_1(w)\hat{v}_1(w)K_1(w)d\hat{R}\big(w,\hat{\beta}\big)}{n\big(1+\hat{R}\big(w;\hat{\beta}\big)\big)}\right.$$
$$\left.+\int_0^s \frac{\hat{\mu}_2(w)\hat{v}_2(w)K_2(w)d\hat{R}\big(w,\hat{\beta}\big)}{n\big(e^{-\hat{\beta}_1}+e^{-\hat{\beta}_2}\hat{R}\big(w;\hat{\beta}\big)\big)}\right), \tag{21.8}$$

and

$$\hat{\sigma}_{CHR}(s,t)=\hat{A}_{CHR}^T(s)\hat{\Omega}\left(\int_0^\tau \frac{\hat{\mu}_1(w)\hat{\mu}_1^T(w)K_1(w)d\hat{R}\big(w;\hat{\beta}\big)}{n\big(1+\hat{R}\big(w;\hat{\beta}\big)\big)}\right.$$
$$\left.+\int_0^\tau \frac{\hat{\mu}_2(w)\hat{\mu}_2^T(w)K_2(w)d\hat{R}\big(w;\hat{\beta}\big)}{n\big(e^{-\hat{\beta}_1}+e^{-\hat{\beta}_2}\hat{R}\big(w;\hat{\beta}\big)\big)}\right)\hat{\Omega}^T\hat{A}_{CHR}(t)$$
$$+\hat{B}_{CHR}(s)\hat{B}_{CHR}(t)\left(\int_0^s \frac{\hat{v}_1^2(w)K_1(w)d\hat{R}\big(w;\hat{\beta}\big)}{n\big(1+\hat{R}\big(w;\hat{\beta}\big)\big)}\right.$$
$$\left.+\int_0^s \frac{\hat{v}_2^2(w)K_2(w)d\hat{R}\big(w;\hat{\beta}\big)}{n\big(e^{-\hat{\beta}_1}+e^{-\hat{\beta}_2}\hat{R}\big(w;\hat{\beta}\big)\big)}\right)$$
$$+\hat{B}_{CHR}(t)\hat{A}_{CHR}^T(s)\hat{\Omega}\left(\int_0^t \frac{\hat{\mu}_1(w)\hat{v}_1(w)K_1(w)d\hat{R}\big(w,\hat{\beta}\big)}{n\big(1+\hat{R}\big(w;\hat{\beta}\big)\big)}\right.$$
$$\left.+\int_0^t \frac{\hat{\mu}_2(w)\hat{v}_2(w)K_2(w)d\hat{R}\big(w,\hat{\beta}\big)}{n\big(e^{-\hat{\beta}_1}+e^{-\hat{\beta}_2}\hat{R}\big(w;\hat{\beta}\big)\big)}\right)$$
$$+\hat{B}_{CHR}(s)\hat{A}_{CHR}^T(t)\hat{\Omega}\left(\int_0^s \frac{\hat{\mu}_1(w)\hat{v}_1(w)K_1(w)d\hat{R}\big(w,\hat{\beta}\big)}{n\big(1+\hat{R}\big(w;\hat{\beta}\big)\big)}\right.$$
$$\left.+\int_0^s \frac{\hat{\mu}_2(w)\hat{v}_2(w)K_2(w)d\hat{R}\big(w,\hat{\beta}\big)}{n\big(e^{-\hat{\beta}_1}+e^{-\hat{\beta}_2}\hat{R}\big(w;\hat{\beta}\big)\big)}\right). \tag{21.9}$$

The estimators $\hat{\Omega}$, $\hat{A}_{RR}(t)$, $\hat{A}_{CHR}(t)$ involve the derivative vector $\partial\hat{R}\big(t;\beta\big)/\partial\beta$ and the derivative matrix in Ω. From various simulation studies, these derivatives can be approximated by numerical derivatives for easier calculation, and the results are fairly stable with respect to the choice of the jump size in the numerical derivatives.

For a fixed $t_0 \le \tau$, confidence intervals for $RR\big(t_0\big)$ can be obtained from the asymptotic normality of $\widehat{RR}(t_0)$ and the estimated variance $\hat{\sigma}_{RR}\big(t_0,t_0\big)$. For better small sample behavior and to ensure that the confidence intervals remain on the positive side of the axis as usual, we make a logarithm transformation resulting in the asymptotic $100(1-\alpha)\%$ confidence interval

$$\widehat{RR}(t_0)\exp\left(\mp z_{\alpha/2}\frac{\sqrt{\hat{\sigma}_{RR}(t_0,t_0)}}{\sqrt{n}\widehat{RR}(t_0)}\right), \tag{21.10}$$

where $z_{\alpha/2}$ is the $100(1-\alpha/2)\%$ percentile of the standard normal distribution.

Similarly, for $CHR(t_0)$, an asymptotic $100(1-\alpha)\%$ confidence interval is

$$\widehat{CHR}(t_0)\exp\left(\mp z_{\alpha/2}\frac{\sqrt{\hat{\sigma}_{CHR}(t_0,t_0)}}{\sqrt{n}\widehat{CHR}(t_0)}\right). \tag{21.11}$$

21.3 Confidence Bands

For simultaneous inference on $RR(t)$ over a time interval $I=[a,b]\subset[0,\tau]$, let $w_n(t)$ be a data-dependent function that converges in probability to a bounded function $w^*(t)>0$, uniformly in t over I. Then, it follows that U_n/w_n converges weakly U^*/w^*. Let c_α be the upper αth percentile of $\sup_{t\in I}|U^*/w^*|$, then an asymptotic $100(1-\alpha)\%$ simultaneous confidence band for $RR(t)$, $t\in I$, can be obtained as

$$\widehat{RR}(t)\exp\left(\mp c_\alpha\frac{w_n(t)}{\sqrt{n}\widehat{RR}(t)}\right). \tag{21.12}$$

The analytic form of c_α is quite intractable. The bootstrapping method provides a well established alternative approach. However, it is very time-consuming. More discussion on this is described further on the applications to clinical trial data in Sect. 21.5. Here we have used a normal resampling approximation similar to the approach used in Lin et al. (1993). This approach results in substantial savings in computing time, and has been used in many works, including Lin et al. (1994), Cheng et al. (1997), Tian et al. (2005), and Peng and Huang (2007).

For $t\le\tau$, let $N_i(t)=\delta_i I(X_i\le t)$, $i=1,\cdots,n$, and define the process

$$\begin{aligned}\hat{U}_n(t) &= \frac{\hat{A}_{RR}^T(t)\hat{\Omega}}{\sqrt{n}}\left(\sum_{i\le n_1}\int_0^\tau\hat{\mu}_1 d(\epsilon_i N_i)+\sum_{i>n_1}\int_0^\tau\hat{\mu}_2 d(\epsilon_i N_i)\right)\\ &\quad+\frac{\hat{B}_{RR}(t)}{\sqrt{n}}\left(\sum_{i\le n_1}\int_0^t\hat{\nu}_1 d(\epsilon_i N_i)+\sum_{i>n_1}\int_0^t\hat{\nu}_2 d(\epsilon_i N_i)\right)\\ &= \frac{\hat{A}_{RR}^T(t)\hat{\Omega}}{\sqrt{n}}\left(\sum_{i\le n_1}\epsilon_i\delta_i\hat{\mu}_1(X_i)I(X_i\le\tau)+\sum_{i>n_1}\epsilon_i\delta_i\hat{\mu}_2(X_i)I(X_i\le\tau)\right)\\ &\quad+\frac{\hat{B}_{RR}(t)}{\sqrt{n}}\left(\sum_{i\le n_1}\epsilon_i\delta_i\hat{\nu}_1(X_i)I(X_i\le t)+\sum_{i>n_1}\epsilon_i\delta_i\hat{\nu}_2(X_i)I(X_i\le t)\right),\end{aligned} \tag{21.13}$$

where ϵ_i, $i = 1, \ldots, n$, are independent standard normal variables that are also independent of the data. Conditional on (X_i, δ_i, Z_i), $i = 1, \ldots, n$, $\hat{U}_n$ is a sum of n independent variables at each time point. In Appendix B, it will be shown that $\hat{U}_n$ given the data converges weakly to U^*. It follows that $\sup_{t\in I} |\hat{U}_n(t)/w_n(t)|$ given the data converges in distribution to $\sup_{t\in I} |U^*(t)/w^*(t)|$. Therefore, c_α can be estimated empirically from a large number of realizations of the conditional distribution of $\sup_{t\in I} |\hat{U}/w|$ given the data.

Similarly, to considerations in Yang and Prentice (2011) for inference on the hazard ratio, we look at several choices of the weight w_n. For $w_n(t) = \sqrt{\hat{\sigma}_{RR}(t,t)}$ we obtain the equal precision bands (Nair 1984), which only differ from point-wise confidence intervals in using c_α instead of $z_{\alpha/2}$. For $w_n(t) = 1 + \hat{\sigma}_{RR}(t,t)$ we obtain the Hall–Wellner type bands recommended by Bie et al. (1987). The simplest case $w_n(t) \equiv 1$ does not require the computation of $\hat{\sigma}_{RR}(t,t)$, and hence is easier to implement.

To obtain simultaneous confidence bands for $CHR(t)$, let

$$\begin{aligned}
\hat{V}_n(t) &= \frac{\hat{A}^T_{CHR}(t)\hat{\Omega}}{\sqrt{n}} \left(\sum_{i\le n_1} \int_0^\tau \hat{\mu}_1 d(\epsilon_i N_i) + \sum_{i>n_1} \int_0^\tau \hat{\mu}_2 d(\epsilon_i N_i) \right) \\
&\quad + \frac{\hat{B}_{CHR}(t)}{\sqrt{n}} \left(\sum_{i\le n_1} \int_0^t \hat{\nu}_1 d(\epsilon_i N_i) + \sum_{i>n_1} \int_0^t \hat{\nu}_2 d(\epsilon_i N_i) \right) \\
&= \frac{\hat{A}^T_{CHR}(t)\hat{\Omega}}{\sqrt{n}} \left(\sum_{i\le n_1} \epsilon_i \delta_i \hat{\mu}_1(X_i) I(X_i \le \tau) + \sum_{i>n_1} \epsilon_i \delta_i \hat{\mu}_2(X_i) I(X_i \le \tau) \right) \\
&\quad + \frac{\hat{B}_{CHR}(t)}{\sqrt{n}} \left(\sum_{i\le n_1} \epsilon_i \delta_i \hat{\nu}_1(X_i) I(X_i \le t) + \sum_{i>n_1} \epsilon_i \delta_i \hat{\nu}_2(X_i) I(X_i \le t) \right),
\end{aligned} \tag{21.14}$$

where ϵ_i, $i = 1, \ldots, n$, are independent standard normal variables that are also independent of the data. Let $\tilde{w}_n(t)$ be a data-dependent function that converges in probability to a bounded function $\tilde{w}^*(t) > 0$, uniformly in t over I. Let $\tilde{c}_\alpha$ be upper αth percentile of $\sup_{t\in[a,b]} |V^*(t)/\tilde{w}^*|$. Similarly, to the argument above for $RR(t)$, an asymptotic $100(1-\alpha)\%$ simultaneous confidence band for $CHR(t)$, $t \in I$, can be obtained as

$$\widehat{CHR}(t) \exp\left(\mp \tilde{c}_\alpha \frac{\tilde{w}_n(t)}{\sqrt{n}\widehat{CHR}(t)} \right), \tag{21.15}$$

where $\tilde{c}_\alpha$ can be approximated empirically from a large number of realizations of the conditional distribution of $\sup_{t\in[a,b]} |\hat{V}(t)/\tilde{w}_n|$ given the data. For $\tilde{w}_n = \sqrt{\hat{\sigma}_{CHR}(t,t)}, 1 + \hat{\sigma}_{CHR}(t,t)$ and $\tilde{w}_n \equiv 1$ respectively, we obtain the equal precision, Hall–Wellner type, and unweighted confidence bands for $CHR(t)$.

21.4 Simulation Studies

For stable moderate sample behavior, we restrict the range of the confidence bands for both $RR(t)$ and $CHR(t)$. The range is between the 40th percentile of the uncensored data at the lower end and the 95th percentile of the uncensored data at the upper end. The lower end point of this range seems a little high compared to other situations such as the inference on the hazard ratio in Yang and Prentice (2011). This is to provide a range in which the nonparametric procedures and the proposed model-based procedures in (21.10–21.12) and (21.15) are to be compared. Toward the beginning of the data range, the nonparametric estimates can be very unstable and the confidence intervals can be quite wide, as will be illustrated in the data example to follow. Also, compared with the hazard ratio as a measure of the temporal pattern of the treatment effect, $RR(t)$ and $CHR(t)$ measure the cumulative treatment effect. Thus in biomedical research, there is little interest in their behaviors near the beginning of the data range. In various applications to clinical trial data, the specified range for the confidence bands is not nearly as restrictive as it seems and contains a meaningful interval of the data range. In the estimating procedures, the function $\hat{P}(t;\mathbf{b})$ is replaced by an asymptotically equivalent form

$$\exp\left(-\int_0^t \frac{1}{\sum_{i=1}^n I\left(X_i \geq s\right)} d\left\{\sum_{i=1}^n \delta_i e^{-b_2 Z_i} I\left(X_i \leq s\right)\right\}\right).$$

For simulation studies reported here and for the real data application in Sect. 21.5, τ was set to include all data in calculating $\hat{\beta}$. All numerical computations were done in *Matlab*. Some representative results of simulation studies are given in Table 21.1, where lifetime variables were generated with $R(t)$ chosen to yield the standard exponential distribution for the control group. The values of β were $(\log(.9), \log(1.2))$ and $(\log(1.2), \log(.8))$, representing 1/3 increase or decrease over time from the initial hazard ratio, respectively. The censoring variables were independent and identically distributed with the log-normal distribution, where the normal distribution had mean c and standard deviation 0.5, with c chosen to achieve various censoring rates. The data were split into the treatment and control groups by a 1:1 ratio. The empirical coverage probabilities were obtained from 1000 repetitions, and for each repetition, the critical values c_α and $\tilde{c}_\alpha$ were calculated empirically from 1,000 realizations of relevant conditional distributions. For both $RR(t)$ and $CHR(t)$, the equal precision bands, Hall–Wellner type bands and unweighted bands are denoted by EP, HW, and UW respectively.

Note that with 1,000 repetitions and $1.96\sqrt{.95 \cdot 0.05/1000} = 0.0135$, we expect the empirical coverage probabilities to be mostly greater than 0.9365. In Table 21.1, for *RR*, the empirical coverage probabilities are greater than 0.9365 for all but one case with the smallest sample size $n = 100$ and at 50 % censoring. For *CHR*, the confidence bands are mostly conservative, with all empirical coverage probabilities greater than 0.95. One plausible explanation for this conservative phenomenon could be that the estimate for $CHR(t)$ is more directly related to the martingales associated with censored data, resulting in better approximations.

Table 21.1 Empirical coverage probabilities of the three types of confidence bands HW, EP, and UW, for the failure probability ratio RR and the ratio of cumulative hazards CHR, under model (21.1), based on 1000 repetitions

Hazard ratio	Censoring	n	RR			CHR		
			HW	EP	UW	HW	EP	UW
0.9 ↑ 1.2	10 %	100	0.949	0.967	0.958	0.977	0.983	0.986
	30 %		0.967	0.966	0.962	0.976	0.984	0.989
	50 %		0.943	0.957	0.966	0.969	0.970	0.975
	10 %	200	0.950	0.973	0.965	0.972	0.978	0.985
	30 %		0.964	0.969	0.968	0.959	0.970	0.986
	50 %		0.959	0.968	0.974	0.969	0.975	0.980
	10 %	400	0.960	0.974	0.971	0.958	0.969	0.982
	30 %		0.961	0.970	0.971	0.959	0.968	0.981
	50 %		0.958	0.974	0.977	0.967	0.975	0.983
1.2 ↓ 0.8	10 %	100	0.946	0.963	0.962	0.974	0.980	0.980
	30 %		0.958	0.964	0.968	0.969	0.977	0.981
	50 %		0.926	0.940	0.966	0.962	0.969	0.974
	10 %	200	0.958	0.967	0.952	0.968	0.975	0.971
	30 %		0.958	0.959	0.958	0.972	0.974	0.968
	50 %		0.946	0.954	0.961	0.962	0.964	0.974
	10 %	400	0.960	0.957	0.954	0.969	0.972	0.973
	30 %		0.960	0.969	0.960	0.966	0.969	0.970
	50 %		0.949	0.962	0.959	0.967	0.974	0.972

21.5 Applications

For the Women's Health Initiative (WHI) randomized controlled trial of combined (estrogen plus progestin) postmenopausal hormone therapy, an elevated coronary heart disease risk was reported, with overall unfavorable health benefits versus risks over an average of 5.6-year study period (Writing Group 2002; Manson et al. 2003). After controlling for time from estrogen-plus-progestin initiation and confounding, hazard ratio estimates still indicate elevated risk of coronary heart disease and venous thromboembolism early on during the trial, under a piece-wise Cox model assuming constant hazard ratio separately on 0–2 years, 2–5 years, and 5+ years (Prentice et al. 2005). Let us first illustrate the methods developed in the previous sections with the venous thromboembolism (VTE) data from the WHI clinical trial. Among the 16,608 postmenopausal women ($n_1 = 8102$), there were 167 and 76 events observed in the treatment and control group respectively, implying about 98.5 % censoring, primarily by the trial stopping time. Fitting model (21.1) to this data set, we get $\hat{\beta} = (4.72, 0.014)^T$. Plots of the model based survival curves and the Kaplan–Meier curves for the two groups show that the model is reasonable. For $RR(t)$, the three 95 % simultaneous confidence bands (EP, HW, and UW) under model (21.1) are given in Fig. 21.1, together with the point estimates. The nonparametric point estimates are also included to compare with the model-based estimates. Furthermore, model-based 95 % point-wise confidence intervals are included as well, to indicate by how much the confidence intervals are widened to improve from point-wise to

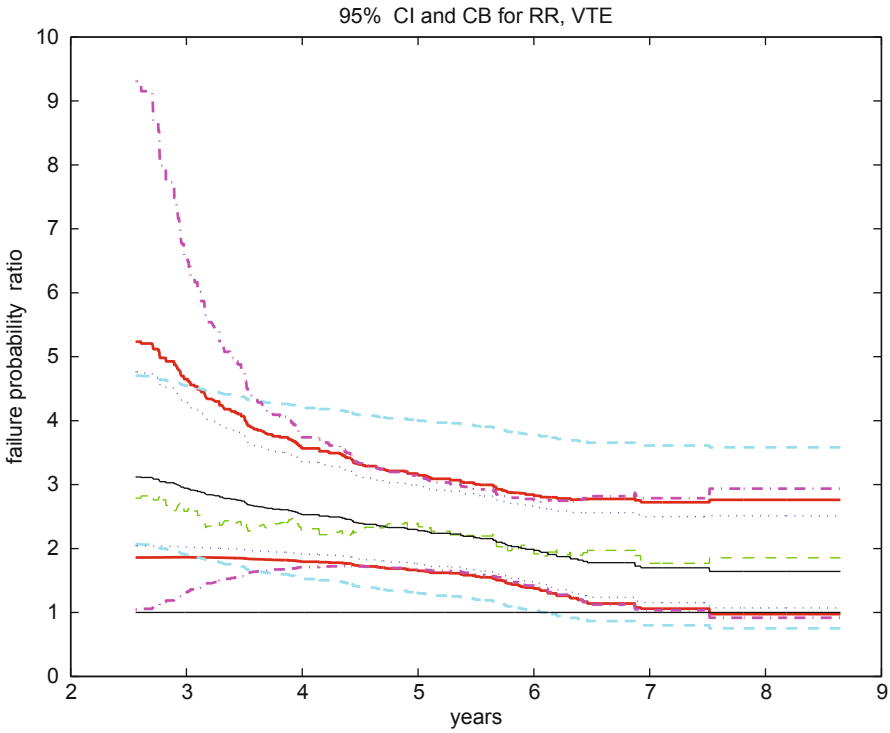


Fig. 21.1 95 % point-wise confidence intervals and simultaneous confidence bands of the failure probability ratio for the WHI VTE data: *Outside red solid lines*—equal precision confidence band, *magenta dash-dotted lines*—Hall–Wellner confidence band, *outside cyan dashed lines*—unweighted confidence band, *dotted lines*—95 % point-wise confidence intervals, *central black solid line*—the estimated failure probability ratio under the model, *central green dashed line*—the estimated failure probability ratio using Kaplan–Meier estimators

simultaneous coverage. From Fig. 21.1, it can be seen that the Hall–Wellner type band and the equal precision band are almost the same a little after the 4th year. However, the Hall–Wellner type band is noticeably wider toward the beginning of the date range. The unweighted band maintains a roughly constant width through the data range considered, which is roughly as wide as the equal precision band at the begining of the data range, but wider throughout the rest of the data range. Similar phenomena are often seen in additional applications not reported here. Based on various applications and simulation studies, we recommend that the equal precision band be used in making inference on $RR(t)$ under model (21.1).

For $CHR(t)$, the 95 % point-wise confidence intervals and confidence bands under model (21.1) are given in Fig. 21.2. Similarly to the case for $RR(t)$, the equal precision band is preferred in making inference on $CHR(t)$ under model (21.1). From Fig. 21.1 and 21.2, there is evidence that from 2.5 to 7.5 years, the event probability is higher in the treatment group than in the control group.

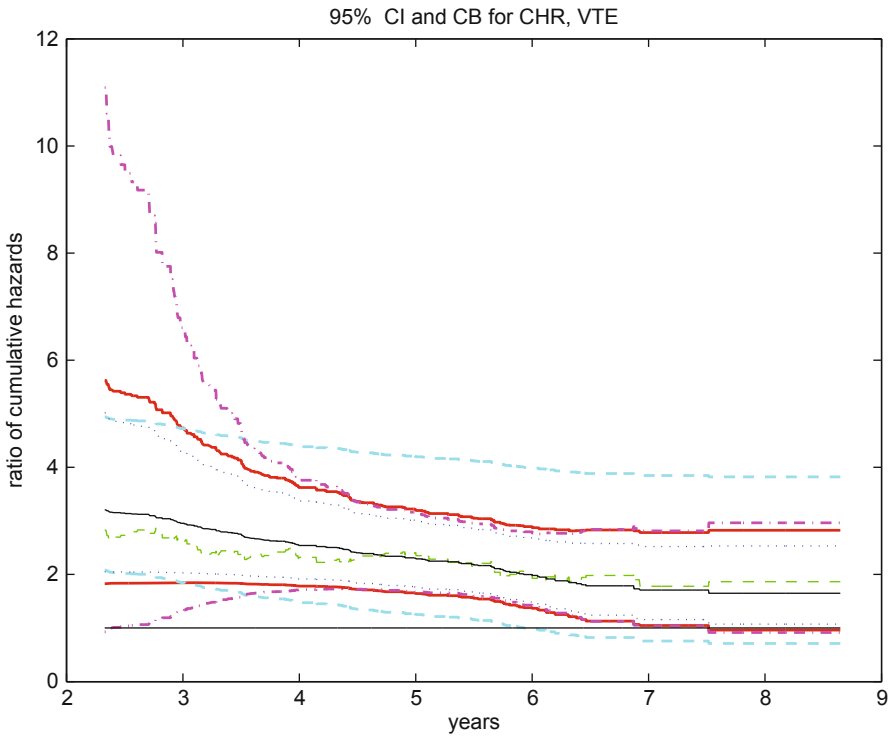


Fig. 21.2 95 % point-wise confidence intervals and simultaneous confidence bands of the ratio of cumulative hazards for the WHI VTE data: *Outside red solid lines*—equal precision confidence band, *magenta dash-dotted lines*—Hall–Wellner confidence band, em outside cyan dashed lines—unweighted confidence band, *dotted lines*—95 % point-wise confidence intervals, *central black solid line*—the estimated failure probability ratio under the model, *central green dashed line*—the estimated failure probability ratio using Kaplan–Meier estimators

For comparison, from Yang and Prentice (2011), the 95 % point-wise confidence intervals and equal precision confidence band are obtained for the hazard ratio under model (21.1), given in Fig. 21.3. The results are in good agreement with the results under the piece-wise Cox model used in Prentice et al. (2005). In an interval near the beginning of the data range, there is greater hazard of venous thromboembolism in the treatment group than in the control group. This interval has shorter length than the intervals in Fig. 21.1 and 21.2 where the treatment group has a higher event probability than in the control group.

Note that the simple bootstrap method for approximating c_α and $\tilde{c}_\alpha$, when $w_n \equiv 1$ and $\tilde{w}_n \equiv 1$ respectively, is already much more computationally intensive than the normal resampling approximation employed here. With $w_n(t) = \sqrt{\hat{\sigma}_{RR}(t)}$ and $\tilde{w}_n = \sqrt{\hat{\sigma}_{CHR}(t)}$, the bootstrap method would require one more level of bootstrapping samples to obtain the estimated variance functions, thus further increasing the computational burden. In comparison, once $\hat{\sigma}_{RR}(t)$ and $\hat{\sigma}_{CHR}(t)$ are obtained, the

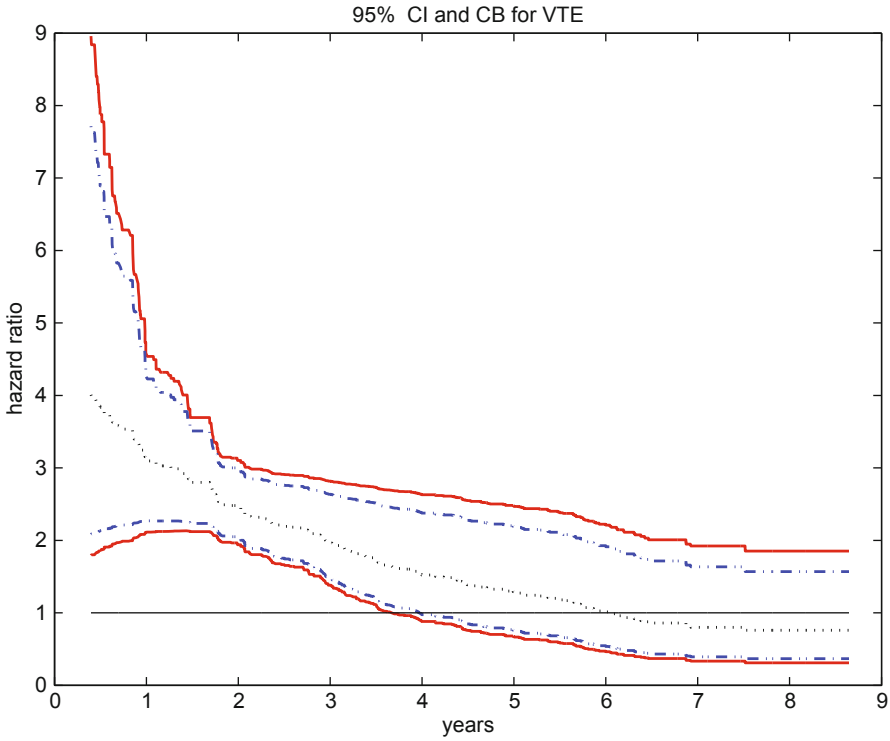


Fig. 21.3 95 % point-wise confidence intervals and simultaneous confidence bands of the hazard ratio function for the WHI VTE data: *Red solid lines*—equal precision confidence band, *blue dash-dotted lines*—95 % point-wise confidence intervals, *dotted line*—the estimated hazard ratio function

normal resampling approximation only needs a small additional computation and programming cost.

To see how the nonparametric procedures compare with the proposed model-based procedures, Fig. 21.4 presents 95 % point-wise confidence intervals, both model-based and nonparametric, together with the point estimates, of $CHR(t)$ for the VTE data from WHI. It can be seen that the nonparametric estimates and confidence intervals can be quite unstable near the beginning of the data range. As $t \downarrow 0$, the hazard ratio at t and $CHR(t)$ should both approach the same limit, which is e^{β_1} under the model. From Fig. 21.4, the model-based estimator of $CHR(t)$ near $t = 0$ takes values around 5, which is comparable to results in the literature, while the nonparametric estimator of $CHR(t)$ near $t = 0$ takes much more extreme values. Also, the model-based estimates and confidence intervals are smoother throughout, and the confidence intervals are often narrower than their nonparametric counterparts. Similar phenomena are also present for $RR(t)$ (omitted). This is a major reason that the nonparametric estimates for $RR(t)$ and $CHR(t)$ are rarely used in biomedical studies.

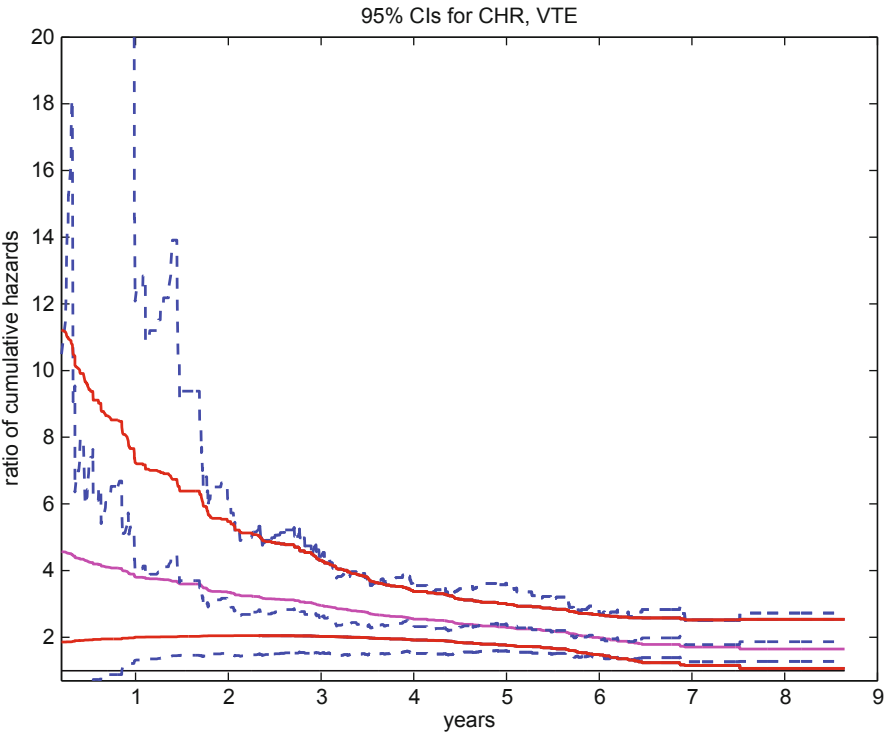


Fig. 21.4 Model-based and nonparametric 95 % point-wise confidence intervals of the ratio of cumulative hazards for the WHI VTE data: *Outside red solid lines*—model based 95 % point-wise confidence intervals, *outside blue dashed lines*—nonparametric 95 % point-wise confidence intervals, *central magenta solid line*—model based estimate of the ratio of cumulative hazards, *central blue dashed line*—nonparametric estimate of the ratio of cumulative hazards

Next, we look at an example with mild violation of the proportional hazards assumption. The Digoxin Intervention Group trial (The Digitalis investigation group 1997) was a randomized, double-blind clinical trial on the effect of digoxin on mortality and hospitalization. In the main trial, patients with left ventricular ejection fraction of 0.45 or less were randomized to digoxin (3397 patients) or placebo (3403 patients) in addition to diuretics and angiotensin-converting-enzyme inhibitors. We look at the data on death attributed to worsening heart failure. For testing the proportional hazards assumption, the acceleration test statistic of Breslow et al. (1984) gives a p- value of 0.098. This indicates some mild proportionality violation. For $RR(t)$,the 95 % point-wise confidence intervals and confidence bands under model (21.1) are given in Fig. 21.4. Possibly due to only a mild violation of the proportionality assumption, the Hall–Wellner type band, the equal precision band and the unweighted band are almost the same for the entire data range considered. From Fig. 21.4, there is evidence that for the range of 1.5–3 year, the treatment reduces the event probability.

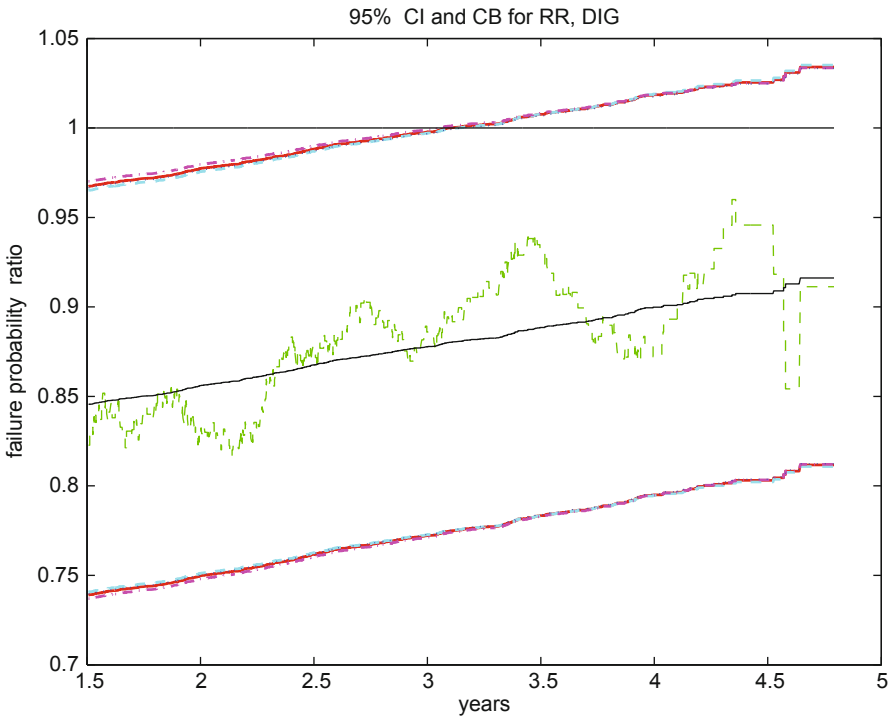


Fig. 21.5 95 % point-wise confidence intervals and simultaneous confidence bands of the failure probability ratio for the DIG data: *Outside red solid lines*—equal precision confidence band, *magenta dash-dotted lines*—Hall–Wellner confidence band, *outside cyan dashed lines*—unweighted confidence band, *Dotted lines*—95 % point-wise confidence intervals, *central black solid line*—the estimated failure probability ratio under the model, *central green dashed line*—the estimated failure probability ratio using Kaplan–Meier estimators

For $CHR(t)$, the 95 % point-wise confidence intervals and confidence bands under model (21.1) are given in Fig. 21.5. Again all three confidence bands are very close to each other. From Fig. 21.5, there is evidence of reduced event probability in the treatment group for the range of 1.3 year to 3 years.

Again for comparison, from Yang and Prentice (2011), the 95 % point-wise confidence intervals and equal precision confidence band are obtained for the hazard ratio under model (21.1), given in Fig. 21.6. From Fig. 21.6, there is evidence that from 0 to .75 year, in the treatment group there is reduced hazard of death attributed to worsening heart failure. Note that this range is much narrower than the range where there is evidence of reduced event probability in the treatment group seen from Fig. 21.4 and 21.5 (Fig. 21.7).

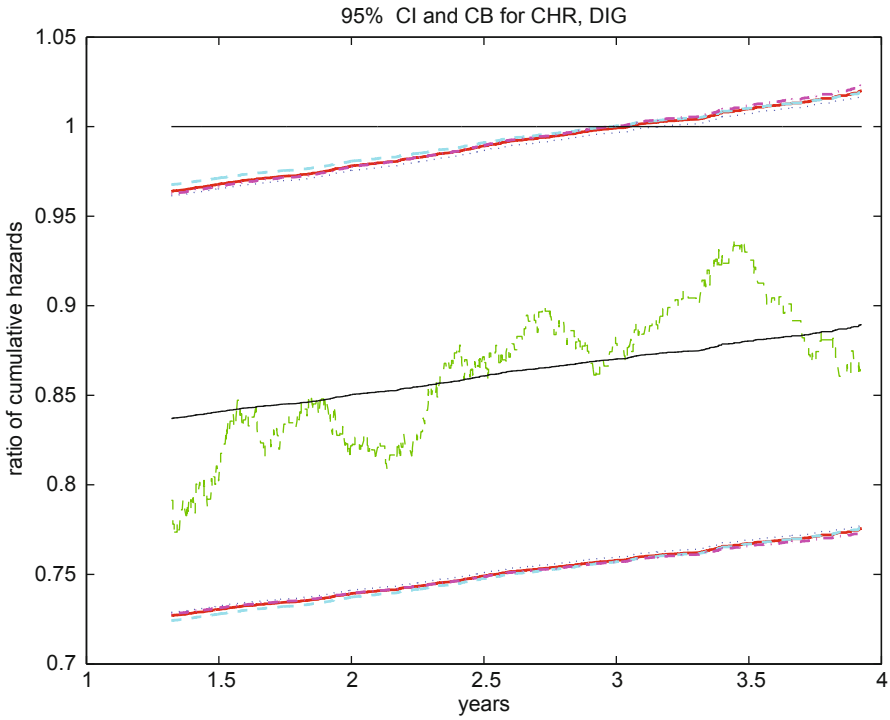


Fig. 21.6 95 % point-wise confidence intervals and simultaneous confidence bands of the ratio of cumulative hazards for the DIG data: *Outside red solid lines*—equal precision confidence band, *magenta dash-dotted lines*—Hall–Wellner confidence band, *outside cyan dashed lines*—unweighted confidence band, *dotted lines*—95 % point-wise confidence intervals, *central black solid line*—the estimated failure probability ratio under the model, *central green dashed line*—the estimated failure probability ratio using Kaplan–Meier estimators

21.6 Discussion

We have studied the asymptotic properties of the estimators for the failure probability ratio and the ratio of cumulative hazards under a semiparametric model applicable to a sufficiently wide range of applications. Point-wise confidence intervals and confidence bands are developed for the two ratios. In simulation studies, the confidence bands have good performance for moderate samples. Among the confidence bands with different weights, the equal precision confidence band is recommended based on various simulation studies and clinical trial data applications. Similarly, inference procedures can be developed for the odds ratio. The point-wise confidence intervals and confidence bands for the odds ratio are usually wider than the corresponding intervals and bands for the failure probability ratio and the ratio of cumulative hazards. Due to space limit those results are not presented here. When the censoring is heavy, there are very little differences among the confidence intervals and bands for the failure probability ratio, the ratio of cumulative hazards, and the odds ratio. The

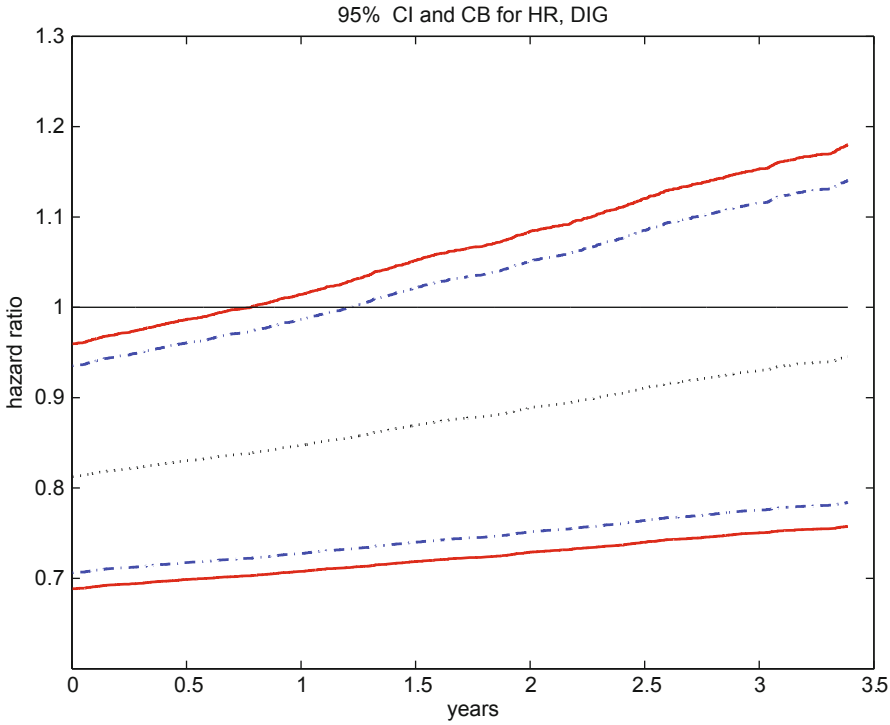


Fig. 21.7 95 % point-wise confidence intervals and simultaneous confidence bands of the hazard ratio function for the DIG data: *Red solid lines*—equal precision confidence band, *blue dash-dotted lines*—95 % point-wise confidence intervals, *dotted line*—the estimated hazard ratio function

confidence intervals and bands presented here provide good visual tools for assessing cumulative effect of the treatment. They can supplement the visual tools based on the hazard ratio which focuses the temporal pattern of the treatment effect. It is also of interest to extend the results here by considering adjustment for covariate via a regression analysis. These and other problems are worthy of further exploration.

Acknowledgements I would like to thank the reviewers and the editor for helpful comments and suggestions, which led to an improved version of the manuscript. This chapter is dedicated to my mentor Dr. Hira Koul. I am greatly indebted to Dr. Koul for his guidance, advice, and encouragement in the last 30 years.

Appendix A: Consistency

The following regularity conditions will be assumed throughout the Appendices:

Condition 1. $\lim \frac{n_1}{n} = \rho \in (0, 1)$.
Condition 2. The survivor function G_i of C_i given Z_i is continuous and satisfies

$$\frac{1}{n}\sum_{i \le n_1} G_i(t) \to \Gamma_1, \ \frac{1}{n}\sum_{i > n_1} G_i(t) \to \Gamma_2,$$

uniformly for $t \le \tau$, for some Γ_1, Γ_2, and $\tau < \tau_0$ such that $\Gamma_j(\tau) > 0$, $j = 1, 2$.

Condition 3. The survivor functions S_C and S_T are absolutely continuous and $S_C(\tau) > 0$.

Under these conditions, the strong law of large numbers implies that (21.3) is satisfied.

For $t \le \tau$, define

$$L(t) = \Gamma_1 S_C + \Gamma_2 S_T,$$

$$U_j(t;\mathbf{b}) = \int_0^t \Gamma_1 dF_C + \exp\big(-b_j\big)\int_0^t \Gamma_2 dF_T,\ j = 1, 2,$$

$$\Lambda_j(t;\mathbf{b}) = \int_0^t \frac{dU_j(s;\mathbf{b})}{L(s)},\ \ j = 1, 2,$$

$$P(t;\mathbf{b}) = \exp\{-\Lambda_2(t;\mathbf{b})\},\ R(t;\mathbf{b}) = \frac{1}{P(t;\mathbf{b})}\int_0^t P(s;\mathbf{b}) d\Lambda_1(s;\mathbf{b}),$$

$$f_j^0(t;\mathbf{b}) = \frac{\exp(-b_j)R^{j-1}(t;\mathbf{b})}{\exp(-b_1) + \exp(-b_2)R(t;\mathbf{b})},\ j = 1, 2,$$

$$m_j(\mathbf{b}) = \left\{\int_0^\tau f_j^0 \Gamma_2(t) dF_T(t) - \int_0^\tau \frac{f_j^0 \Gamma_2(t) S_T(t) dR(t;\mathbf{b})}{\exp\big(-b_1\big) + \exp\big(-b_2\big)R(t;\mathbf{b})}\right\},\ j=1, 2,$$

and $m(\mathbf{b}) = \big(m_1(\mathbf{b}), m_2(\mathbf{b})\big)'$. We will also assume

Condition 4. The function $m(\mathbf{b})$ is non-zero for $b \in \mathcal{B} - \{\beta\}$, where $\mathcal{B}$ is a compact neighborhood of β.

Theorem 1. *Suppose that Conditions 1 $\sim$ 4 hold. Then, (i) the zero $\hat{\beta}$ of $Q(\mathbf{b})$ in $\mathcal{B}$ is strongly consistent for β; (ii) $\widehat{RR}(t)$ is strongly consistent for $RR(t)$, uniformly for $t \in [0, \tau]$, and $\widehat{CHR}(t)$ is strongly consistent for $CHR(t)$, uniformly on $t \in [0, \tau]$; (iii) $\hat{\Omega}$ converges almost surely to a limiting matrix Ω^*.*

Proof. Under Conditions 1 $\sim$ 3, the limit of $\sum_{i=1}^n I\big(X_i \ge t\big)/n$ is bounded away from zero on $t \in [0, \tau]$. Thus, it can be shown that, with probability 1,

$$\frac{\sum_{i=1}^n \delta_i e^{-b_j Z_i} I\big(X_i = t\big)}{\sum_{i=1}^n \delta_i I\big(X_i \ge t\big)} \to 0,\ j = 1, 2,\ |\Delta\hat{P}(t;\mathbf{b})| \to 0,\ |\Delta\hat{R}(t;\mathbf{b})| \to 0, \tag{21.16}$$

uniformly for $t \in [0, \tau]$ and $b \in \mathcal{B}$, where Δ indicates the jump of the function in t.

Define the martingale residuals

$$\hat{M}_i(t;\mathbf{b}) = \delta_i I\big(X_i \le t\big) - \int_0^t I\big(X_i \ge s\big)\frac{\hat{R}(ds;\mathbf{b})}{e^{-b_1 Z_i} + e^{-b_2 Z_i}\hat{R}(s;\mathbf{b})},\ 1 \le i \le n.$$

From (21.16) and the fundamental theorem of calculus, it can be shown that, with probability 1,

$$Q(\mathbf{b}) = \sum_{i=1}^{n} \int_0^\tau \left\{ f_i(t;\mathbf{b}) + o(1) \right\} \hat{M}_i(dt;\mathbf{b}), \tag{21.17}$$

uniformly in $t \le \tau$, $b \in \mathcal{B}$ and $i \le n$, where $f_i = (f_{1i}, f_{2i})^T$, with

$$f_{1i}(t;\mathbf{b}) = \frac{Z_i e^{-b_1 Z_i}}{e^{-b_1 Z_i} + e^{-b_2 Z_i} \hat{R}(t;\mathbf{b})},\ f_{2i}(t;\mathbf{b}) = \frac{Z_i e^{-b_2 Z_i} \hat{R}(t;\mathbf{b})}{e^{-b_1 Z_i} + e^{-b_2 Z_i} \hat{R}(t;\mathbf{b})}.$$

From the strong law of large numbers (Pollard 1990, p. 41) and repeated use of Lemma A1 of Yang and Prentice (2005), one obtain, with probability 1,

$$\hat{P}(t;\mathbf{b}) \to \hat{P}(t;\mathbf{b}),\ \hat{R}(t;\mathbf{b}) \to R(t;\mathbf{b}),\ Q(\mathbf{b})/n \to m(\mathbf{b}), \tag{21.18}$$

uniformly in $t \le \tau$ and $\mathbf{b} \in \mathcal{B}$. From these results and Condition 4, one obtains the strong consistency of $\widehat{RR}(t)$ and $\widehat{CHR}(t)$, and almost sure convergence of $\hat{\Omega}$.

Appendix B: Weak Convergence

Let $\xi_0(t) = 1 + R(t), \xi(t) = e^{-\beta_1} + e^{-\beta_2} R(t), \hat{\xi}_0(t) = 1 + \hat{R}(t;\beta), \hat{\xi}(t) = e^{-\beta_1} + e^{-\beta_2} \hat{R}(t;\beta)$, and define

$$\begin{aligned} K_1(t) &= \sum_{i \le n_1} I(X_i \ge t),\ K_2(t) = \sum_{i > n_1} I(X_i \ge t), \\ H(t) &= \frac{1}{\hat{\xi}(t)} (e^{-\beta_1}, e^{-\beta_2} \hat{R}(t;\beta))^T, \\ J(t) &= \int_t^\tau \frac{H(s) K_1(s) K_2(s)}{\hat{\xi}(s) \hat{P}(s;\beta)} \left(\frac{e^{-\beta_2}}{\xi(s)} - \frac{1}{\xi_0(s)} \right) dR(s). \end{aligned}$$

Similarly, to the proof of Theorem 1, it can be shown that, with probability 1,

$$Q(\beta) = \sum_{i \le n_1} \int_0^\tau \{\mu_1(t) + o(1)\} dM_i(t) + \sum_{i > n_1} \int_0^\tau \left\{ \mu_2(t) + o(1) \right\} dM_i(t), \tag{21.19}$$

uniformly in $t \le \tau$ and $i \le n$, where

$$\begin{aligned} \mu_1(t) &= -\frac{\hat{\xi}_0(t) H(t) K_2(t)}{\hat{\xi}(t) K(t)} + \frac{\hat{\xi}_0(t) \hat{P}_-(t;\beta)}{K} J(t), \\ \mu_2(t) &= H(t) \frac{K_1(t)}{K(t)} + \frac{\hat{\xi}(t) \hat{P}_-(t;\beta)}{K(t)} J(t), \end{aligned} \tag{21.20}$$

$$M_i(t) = \delta_i I(X_i \le t) - \int_0^t I(X_i \ge s) \frac{dR(s)}{e^{-\beta_1 Z_i} + e^{-\beta_2 Z_i} R(s)},\ i = 1, \ldots, n.$$

By Lemma A3 of Yang and Prentice (2005),

$$\sqrt{n}(\hat{R}(t;\beta) - R(t)) = \frac{1}{\sqrt{n}\hat{P}(t;\beta)} \left(\sum_{i \le n_1} \int_0^t \nu_1 dM_i + \sum_{i > n_1} \int_0^t \nu_2 dM_i \right) \quad (21.21)$$

where

$$\nu_1(t) = \frac{n\xi_0(t)\hat{P}_-(t;\beta)}{K(t)},\ \nu_2(t) = \frac{n\xi(t)\hat{P}_-(t;\beta)}{K(t)}.$$

Define

$$A_{RR}(t) = \left(\frac{\hat{S}_T(t)}{\hat{F}_C(t)\hat{\xi}(t)} - \frac{\hat{F}_T(t)\hat{S}_C^2(t)}{\hat{F}_C^2(t)} \right) \frac{\partial \hat{R}(t;\beta)}{\partial \beta} + \frac{\hat{S}_T(t)}{\hat{F}_C(t)} \left(\frac{R(t)}{\xi(t)}, \Lambda_T(t) - \frac{R(t)}{\xi(t)} \right)^T,$$

$$B_{RR}(t) = \frac{1}{\hat{P}(t;\beta)} \left(\frac{\hat{S}_T(t)}{\hat{F}_C(t)\hat{\xi}(t)} - \frac{\hat{F}_T(t)\hat{S}_C^2(t)}{\hat{F}_C^2(t)} \right),$$

$$A_{CHR}(t) = \left(\frac{1}{\Lambda_C(t)\hat{\xi}(t)} - \frac{\Lambda_T(t)\hat{S}_C(t)}{\Lambda_C^2(t)} \right) \frac{\partial \hat{R}(t;\beta)}{\partial \beta} + \frac{1}{\Lambda_C(t)} \left(\frac{R(t)}{\xi(t)}, \Lambda_T(t) - \frac{R(t)}{\xi(t)} \right)^T,$$

$$B_{CHR}(t) = \frac{1}{\hat{P}(t;\beta)} \left(\frac{1}{\Lambda_C(t)\hat{\xi}(t)} - \frac{\Lambda_T(t)\hat{S}_C(t)}{\Lambda_C^2(t)} \right).$$

For $A_{RR}(t)$, $B_{RR}(t)$, $A_{CHR}(t)$, $B_{CHR}(t)$, $mu_1(t), \mu_2(t), \nu_1(t)$, $\nu_2(t)$, let $A^*_{RR}(t)$, $B^*_{RR}(t), \ldots$ etc. be their almost sure limit. In addition, let L_j be the almost sure limit of K_j/n, $j = 1, 2$. For $0 \le s, t < \tau$, let

$$\begin{aligned}
&\sigma_{RR}(s,t) \\
&= A^{*T}_{RR}(s)\Omega^* \left(\int_0^\tau \frac{\mu_1^* \mu_1^{*T}}{1+R} L_1 dR + \int_0^\tau \frac{\mu_2^* \mu_2^{*T}}{e^{-\beta_1} + e^{-\beta_2} R} L_2 dR \right) \Omega^{*T} D^*(t) \\
&\quad + B^*_{RR}(s) B^*_{RR}(t) \left(\int_0^s \frac{\nu_1^{*2}}{1+R} L_1 dR + \int_0^s \frac{\nu_2^{*2}}{e^{-\beta_1} + e^{-\beta_2} R} L_2 dR \right) \\
&\quad + B^*_{RR}(t) A^{*T}_{RR}(s) \Omega^* \left(\int_0^t \frac{\mu_1^* \nu_1^*}{1+R} L_1 dR + \int_0^t \frac{\mu_2^* \nu_2^*}{e^{-\beta_1} + e^{-\beta_2} R} L_2 dR \right) \\
&\quad + B^*_{RR}(s) A^{*T}_{RR}(t) \Omega^* \left(\int_0^s \frac{\mu_1^* \nu_1^*}{1+R} L_1 dR + \int_0^s \frac{\mu_2^* \nu_2^*}{e^{-\beta_1} + e^{-\beta_2} R} L_2 dR \right), \quad (21.22)
\end{aligned}$$

and

$$\sigma_{CHR}(s,t)$$

$$
\begin{aligned}
=&A^{*T}_{CHR}(s)\Omega^*\left(\int_0^\tau \frac{\mu_1^*\mu_1^{*T}}{1+R}L_1 dR + \int_0^\tau \frac{\mu_2^*\mu_2^{*T}}{e^{-\beta_1}+e^{-\beta_2}R}L_2 dR\right)\Omega^{*T} D^*(t)\\
&+ B^*_{CHR}(s)B^*_{CHR}(t)\left(\int_0^s \frac{\nu_1^{*2}}{1+R}L_1 dR + \int_0^s \frac{\nu_2^{*2}}{e^{-\beta_1}+e^{-\beta_2}R}L_2 dR\right)\\
&+ B^*_{CHR}(t)A^{*T}_{CHR}(s)\Omega^*\left(\int_0^t \frac{\mu_1^*\nu_1^*}{1+R}L_1 dR + \int_0^t \frac{\mu_2^*\nu_2^*}{e^{-\beta_1}+e^{-\beta_2}R}L_2 dR\right)\\
&+ B^*_{CHR}(s)A^{*T}_{CHR}(t)\Omega^*\left(\int_0^s \frac{\mu_1^*\nu_1^*}{1+R}L_1 dR + \int_0^s \frac{\mu_2^*\nu_2^*}{e^{-\beta_1}+e^{-\beta_2}R}L_2 dR\right). \quad (21.23)
\end{aligned}
$$

For $A_{RR}(t)$, $B_{RR}(t), \ldots$, etc. define corresponding estimator $\hat{B}_{RR}(t)$, $\hat{A}_{RR}(t)$, ... by replacing β with $\hat{\beta}$, $R(t)$ with $\hat{R}(t;\hat{\beta})$. Defince $\hat{\sigma}_{RR}(s,t)$ and $\hat{\sigma}_{CHR}(s,t)$ by repalcing $B_{RR}(t)$, $A_{RR}(t)$, $\mu_1(t)$, $\mu_2(t)$, $\nu_1(t)$, $\nu_2(t)$, ... in $\sigma_{RR}(s,t)$ and $\sigma_{CHR}(s,t)$ by $\hat{B}_{RR}(t)$, $\hat{A}_{RR}(t), \ldots$ etc.

Theorem 2. *Suppose that Conditions 1 ∼ 4 hold and that the matrix Ω^* is non-singular. Then, (i) U_n is asymptotically equivalent to the process $\tilde{U}_n$ in (21.6) which converges weakly to a zero-mean Gaussian process U^* on $[0,\tau]$, with covariance function $\sigma_{RR}(s,t)$ in (21.22). $\sigma_{RR}(s,t)$ can be consistently estimated by $\hat{\sigma}_{RR}(s,t)$. In addition, $\hat{U}_n(s)$ given the data converges weakly to the same limiting process U^*. (ii) $V_n(t)$ is asymptotically equivalent to the process $\tilde{V}_n$ in (21.7) which converges weakly to a zero-mean Gaussian process V^* on $[0,\tau]$, with covariance function $\sigma_{CHR}(s,t)$ in (21.23). $\sigma_{CHR}(s,t)$ can be consistently estimated by $\hat{\sigma}_{CHR}(s,t)$. In addition, $\hat{V}_n(s)$ given the data converges weakly to the same limiting process V^*.*

Proof. (i) As in the proof for Theorem A2 (ii) in Yang and Prentice (2005), by the strong embedding theorem and (21.19), $Q(\beta)/\sqrt{n}$ can be shown to be asymptotically normal. Now Taylor series expansion of $Q(\mathbf{b})$ around β and the non-singularity of Ω^* imply that $\sqrt{n}(\hat{\beta}-\beta)$ is asymptotically normal. From the $\sqrt{n}$- boundedness of $\hat{\beta}$,

$$
\sqrt{n}\big(\hat{R}(t;\hat{\beta}) - \hat{R}(t;\beta)\big) = \frac{\partial R(t;\beta)}{\partial\beta}\sqrt{n}(\hat{\beta}-\beta) + o_p(1),
$$

uniformly in $t \le \tau$. These results, some algebra and Taylor series expansion together show that U_n is asymptotically equivalent to $\tilde{U}_n$. Similarly, to the proof of the asymptotic normality of $Q(\beta)/\sqrt{n}$, one can show that $\tilde{U}_n$ converges weakly to a zero-mean Gaussian process. Denote the limiting process by U^*. From the martingale integral representation of $\tilde{U}_n$, it can be shown that the covariation process of U^* is given by $\sigma(s,t)$ in (21.22). The consistency of $\hat{\sigma}_{RR}(s,t)$ can be shown similarly to the proof of Theorem 1.

By checking the tightness condition and the convergence of the finite-dimensional distributions, it can be shown that $\hat{U}_n(s)$ given the data also converges weakly to U^*.

(ii) The assertions on V_n, $\tilde{V}_n$, etc. can be proved similarly to the case for U_n, $\tilde{U}_n$, etc. in (i).

References

Aalen OO (1975) Statistical inference for a family of counting processes. PhD thesis. University of California, Berkeley

Bie O, Borgan O, Liestøl lK (1987) Confidence intervals and confidence bands for the cumulative hazard rate function and their small-sample properties. Scand J Stat 14:221–233

Breslow N, Elder L, Berger L (1984). A two sample censored-data rank test for acceleration. *Biometrics* **40**, 1042–1069

Cheng SC, Wei LJ, Ying Z (1997) Predicting survival probabilities with semiparametric transformation models. J Am Stat Assoc 92:227–235

Cox DR (1972) Regression models and life-tables (with Discussion). J R Statist Soc B 34:187–220

Dabrowska DM, Doksum KA, Song J (1989) Graphical comparison of cumulative hazards for two populations. Biometrika 76:763–773

The Digitalis Investigation Group (1997) The effect of Digoxin on mortality and morbidity in patients with heart failure. N Engl J Med 336:525–533

Dong B, Matthews DE (2012) Empirical likelihood for cumulative hazard ratio estimation with covariate adjustment. Biometrics 68:408–418

Kalbfleisch JD, Prentice RL (2002) The statistical analysis of failure time data, 2nd ed. Wiley, New York

Kaplan E, Meier P (1958) Nonparametric estimation from incomplete observations. J Am Stat Assoc 53:457–481

Lin DY, Wei LJ, Ying Z (1993) Checking the Cox model with cumulative sums of martingale-based residuals. Biometrika 80:557–572

Lin DY, Fleming TR, Wei LJ (1994) Confidence bands for survival curves under the proportional hazards model. Biometrika 81:73–81

Manson JE, Hsia J, Johnson KC, Rossouw JE, Assaf AR, Lasser NL, Trevisan M, Black HR, Heckbert SR, Detrano R, Strickland OL, Wong ND, Crouse JR, Stein E, Cushman M, for the Women'S Health Initiative Investigators (2003) Estrogen plus progestin and the risk of coronary heart disease. New Eng J Med 349:523–534

McKeague IW, Zhao Y (2002) Simultaneous confidence bands for ratios of survival functions via empirical likelihood. Stat Probabil Lett 60:405–415

Nair VN (1984) Confidence bands for survival functions with censored data: a comparative study. Technometrics 26:265–275

Nelson W (1969) Hazard plotting for incomplete failure data. J Qual Technol 1:27–52

Parzen MI, Wei LJ, Ying Z (1997) Simultaneous confidence intervals for the difference of two survival functions. Scand J Stat 24:309–314

Peng L, Huang Y (2007) Survival analysis with temporal covariate effects. Biometrika 94:719–733

Pollard D (1990) Empirical processes: theory and applications. Institute of Mathematical Statistics, Hayward

Prentice RL, Langer R, Stefanick ML, Howard BV, Pettinger M, Anderson G, Barad D, Curb JD, Kotchen J, Kuller L, Limacher M, Wactawski-Wende J, for the Women'S Health Initiative Investigators (2005) Combined postmenopausal hormone therapy and cardiovascular disease: toward resolving the discrepancy between observational studies and the women's health initiative clinical trial. Amer J of Epi 162:404–414

Schaubel DE, Wei G (2011) Double inverse-weighted estimation of cumulative treatment effects under nonproportional hazards and dependent censoring. Biometrics 67:29–38

Tian L, Zucker D, Wei LJ (2005) On the Cox model with time-varying regression coefficients. J Am Statist Assoc 100:172–183

Tong X, Zhu C, Sun J (2007) Semiparametric regression analysis of two-sample current status data, with applications to tumorigenicity experiments. Canad J Statist 35:575–584

Writing Group for the Women's Health Initiative Investigators (2002) Risks and benefits of estrogen plus progestin in healthy postmenopausal women: principal results from the women's health initiative randomized controlled trial. J Amer Med Assoc 288:321–333

Yang S, Prentice RL (2005) Semiparametric analysis of short-term and long-term hazard ratios with two-sample survival data. Biometrika 92:1–17

Yang S, Prentice RL (2011) Estimation of the 2-sample hazard ratio function using a semiparametric model. Biostatistics 12:354–368

Chapter 22
Inference for the Standardized Median

Robert G. Staudte

22.1 Introduction

We carry out inference for the median, divided by a fixed interquantile range (IQR). It is a standardized effect or 'effect size' defined by 3 quantiles. Applied statisticians sometimes prefer effect sizes to raw effects, such as the median, because it is scale-free. Inference regarding the median itself has already been thoroughly investigated by a number of authors, including (McKean and Schrader (1984) and Sheather and McKean 1987). More generally, the problem of Studentizing a quantile estimator has been studied by (Bloch and Gastwirth 1968; Hall and Sheather 1988 and Siddiqui 1960), amongst others.

We are given a location-scale family $F_{\alpha,\beta}(x) = F((x-\alpha)/\beta)$ where $\alpha, \beta > 0$ are unknown. Assume $F = F_{0,1}$ has a continuous derivative f which is symmetric about 0 and is positive over a possibly infinite symmetric interval containing 0. Denote the quantile function of F by $G = F^{-1}$, its value at any $0 < r < 1$ by $x_r = G(r)$, and its derivative at r by $g(r) = \{f(x_r)\}^{-1}$. The rth quantile of $F_{\alpha,\beta}$ is related to x_r of F by $\alpha + \beta x_r$, and $g_{\alpha,\beta}(r) = \beta\, g(r)$.

For $0 < r < 0.5$, a value to be chosen later, let the rth *IQR* of $F_{\alpha,\beta}$ be $\beta\,\mathrm{IQR}_r$, where $\mathrm{IQR}_r = x_{1-r} - x_r$. Also define the rth *standardized median* of $F_{\alpha,\beta}$ by:

$$\delta_r = \delta_r(\alpha,\beta) = \frac{\alpha + \beta x_{0.5}}{\beta\,\mathrm{IQR}_r} = \frac{\alpha}{\beta\,\mathrm{IQR}_r}\,. \tag{22.1}$$

Let $X_{([nr])}$ denote the $[nr]$th order statistic of a sample of size n from F. Let $M = X_{([n/2])}$ be the sample median, which is consistent for $x_{0.5}$ and for a fixed $0 < r < 1/2$ define the rth sample IQR by $R_r = X_{([(1-r)n])} - X_{([nr])}$, which is a consistent estimator of $x_{1-r} - x_r$. We want to estimate the rth standardized median defined in (22.1) by $\hat{\delta}_r = M/R_r$.

In the next Sect. 22.2 we derive a variance stabilizing transformation (VST) of $\hat{\delta}_r$. This leads to confidence intervals for δ_r, whose effectiveness of coverage is evaluated

R. G. Staudte (✉)
La Trobe University, 3086 Melbourne, Australia
e-mail: r.staudte@latrobe.edu.au

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_22,

for several examples in Sect. 22.3. Then we consider the use of the VST-transformed standardized median for testing, and the choice of r, in Sect. 22.4, and in Sect. 22.5 we summarize the results and point to several extensions.

22.2 Deriving an Effective Variance Stabilizer

For $1 \le r \le s \le n$, ignoring terms of lower order, we have for large n, (see, for example, DasGupta 2006 p. 80)

$$\begin{aligned} \mathrm{E}[X_{([rn])}] &\doteq x_r \\ \mathrm{Cov}[X_{([rn])}, X_{([sn])}] &\doteq \frac{r(1-s)\, g(r)g(s)}{n} . \end{aligned} \tag{22.2}$$

When $F = F_{\alpha,\beta}$ the expectation of $\hat{\delta}$ is, up to terms of smaller order,

$$\mathrm{E}[\hat{\delta}_r] \doteq \frac{\mathrm{E}[M]}{\mathrm{E}[R_r]} = \frac{\alpha}{\beta \mathrm{IQR}_r} = \delta_r . \tag{22.3}$$

The first and second asymptotic moments of M and R_r are obtained from (22.2). The expression for $\mathrm{Var}[R_r]$ requires a little more care and is given by

$$n\mathrm{Var}[R_r] = h^2(r) = r(1-r)\{g^2(r) + g^2(1-r)\} - 2r^2\, g(r)g(1-r) . \tag{22.4}$$

When F is symmetric, $g(r) = g(1-r)$ and $h^2(r) = 2r(1-2r)g^2(r)$.

But the following expressions hold for asymmetric f as well. For the location-scale family, $F_{\alpha,\beta}$, $f_{\alpha,\beta}(\alpha + \beta x_r) = f(x_r)/\beta$. Without loss of generality, we take $\beta = 1$. Recall the standard formula for the variance of a ratio of random variables:

$$\mathrm{Var}[\hat{\delta}_r] \doteq \frac{1}{\mathrm{E}^2[R_r]} \left\{ \mathrm{Var}[M] - 2\,\mathrm{Cov}[M, R_r] \frac{\mathrm{E}[M]}{\mathrm{E}[R_r]} + \mathrm{Var}[R_r] \frac{\mathrm{E}^2[M]}{\mathrm{E}^2[R_r]} \right\} . \tag{22.5}$$

It follows from (22.2) that the required second moments of M, R_r under the distribution $F_{\alpha,1}$ are:

$$\begin{aligned} \mathrm{Var}[M] &\doteq \frac{g^2(0.5)}{4n} \\ \mathrm{Cov}[M, R_r] &\doteq \frac{r\, g(0.5)}{2n} \{g(1-r) - g(r)\} \\ \mathrm{Var}[R_r] &\doteq \frac{h^2(r)}{n} . \end{aligned} \tag{22.6}$$

Note that $\mathrm{Cov}[M, R_r] \doteq 0$ for symmetric F. Therefore, $V_r = \mathrm{Var}[\hat{\delta}_r]$ is from (22.2), (22.3), and (22.5), approximately equal to

$$V_r \doteq \frac{1}{n}\left\{\frac{g^2(0.5)}{4\mathrm{IQR}_r^2} + \frac{h^2(r)}{\mathrm{IQR}_r^2}\delta_r^2\right\} = \frac{b_r c_r^2}{n}\left(1 + \delta_r^2/c_r^2\right), \tag{22.7}$$

where

$$b_r = \frac{h^2(r)}{\mathrm{IQR}_r^2} \doteq \frac{n\mathrm{Var}[R_r]}{\mathrm{E}^2[R_r]}. \tag{22.8}$$

$$c_r = \frac{g(0.5)}{2\,h(r)} \doteq \frac{\mathrm{SD}[M]}{\mathrm{SD}[R_r]}. \tag{22.9}$$

We also introduce $a_r = \{b_r c_r^2\}^{-1/2} = 2\,\mathrm{IQR}_r/g(0.5) \doteq \mathrm{E}[R_r]/\mathrm{SD}[M]$. Note that the constants b_r, c_r are free of n, α, and β.

Now to first order $\mathrm{E}[\hat{\delta}_r] = \delta_r$, so (22.7) shows that the variance of $\hat{\delta}_r$ is a simple quadratic in its mean; solving in the usual way [Bickel and Doksum (1977) p.32] leads to the VST

$$T_n(x) = \sqrt{\frac{n}{b_r c_r^2}}\, c_r \log\left(\frac{x}{c_r} + \sqrt{1 + \frac{x^2}{c_r^2}}\right) = \sqrt{n}\, a_r\, c_r\, \sinh^{-1}\left(\frac{x}{c_r}\right). \tag{22.10}$$

We expect that, at least for large n, $T_n \sim N(\sqrt{n}\,\mathcal{K}(\delta_r), 1)$, where

$$\mathcal{K}(\delta_r) = a_r\, c_r\, \sinh^{-1}\left(\frac{\delta_r}{c_r}\right). \tag{22.11}$$

This function is called the *Key Inferential Function*, or simply the *Key*, by (Kulinskaya et al. 2008; Morgenthaler and Staudte 2012) because of its appearance in power functions in testing hypotheses, see Sect. 22.4 and in finding confidence intervals, as shown in Sect. 22.3. Further, it is quite generally an excellent approximation to the signed square root of the Kullback–Leibler symmetrized divergence between null and alternative distributions, for a large neighborhood of the null, see Morgenthaler and Staudte (2012).

22.3 Examples

Now for various choices of r we find the VST's for several examples and make plots of the empirical coverage probabilities of nominal 95 % confidence intervals for δ. Such intervals are found as follows: nominal 95 % confidence intervals for $\kappa = \mathcal{K}(\delta)$ are of the form $(T_n \pm z_{0.975})/\sqrt{n}$, and the intervals for δ, clearly having the same

Table 22.1 Normal model constants used for obtaining a variance stabilization of $\hat{\delta}_r$

r	IQR_r	$g(r)$	$h(r)$	a_r	b_r	c_r
0.20	1.683	3.572	1.750	1.343	1.081	0.716
0.25	1.349	3.147	1.573	1.076	1.360	0.797
0.30	1.049	2.876	1.409	0.837	1.805	0.890
0.35	0.771	2.700	1.237	0.615	2.577	1.013
0.40	0.507	2.588	1.035	0.404	4.175	1.211

coverage, are obtained by back-transformation; that is, applying $\mathcal{K}^{-1}$ to the interval for κ. Simulation studies were carried out using the software package R (2008), version 2.15.2, based on 40,000 replications at each δ in 0:6/0.25. The quantiles were estimated using the quantile function (Type 8) recommended by Hyndman and Fan (1996). It estimates quantiles by a mixture of adjacent order statistics and is median-unbiased.

22.3.1 Normal

For the normal density $f = \varphi$, the constant $g(1/2) = 1/\varphi(z_{0.5}) = 2.506628$. Some selected values of the other constants required for defining the VST for selected values of r are given in Table 22.1. In Fig. 22.1, the top four plots show the performance of the VST T_{15} defined by (22.10). The top left hand plot shows the *Key* together with the estimate of it $\hat{\kappa} = T_{15}/\sqrt{15}$. The graph is only plotted for $\delta \geq 0$ because the *Key* and its estimate are skew symmetric in δ. The top right hand plot shows the bias in T_{15} is approximately 0.1, while the bottom left plot shows that the VST has stabilized the variance to a value less than, but close to 1. In any case, the bias-squared of T_{15} is negligible compared to its variance (This is not the case in general, especially as r approaches 0.5). The bottom right plot shows the empirical coverage probability of nominal 95 % confidence intervals for δ based on samples of size $n = 15$, plotted as a function of δ.

In the bottom plots of Fig. 22.1, the empirical coverage probabilities of nominal 95 % confidence intervals for δ based on the VST's are shown for selected values of r and n. The solid line is for $n = 10$, the red-dashed line is for $n = 20$, the blue dotted line is for $n = 40$ and the magenta dot-dashed line is for $n = 100$. The coverage probabilities are close to the nominal value for n in the range 20–100 for $r = 0.2,\ 0.25$ and $r = 0.3$. Note that the coverage is not monotonically approaching the nominal value with increasing n, but is nevertheless satisfactory even for small n.

22.3.2 Uniform

When the distribution belongs to a location-scale family generated by the uniform distribution on $[-1, 1]$, the constants required for analysis are shown in Table 22.2. Note the simplicity of the constants for $r = 0.25$; in this case, the VST is $T_n = \sqrt{n}\ \sinh^{-1}(\hat{\delta})$. The empirical coverage probabilities of nominal 95 % confidence

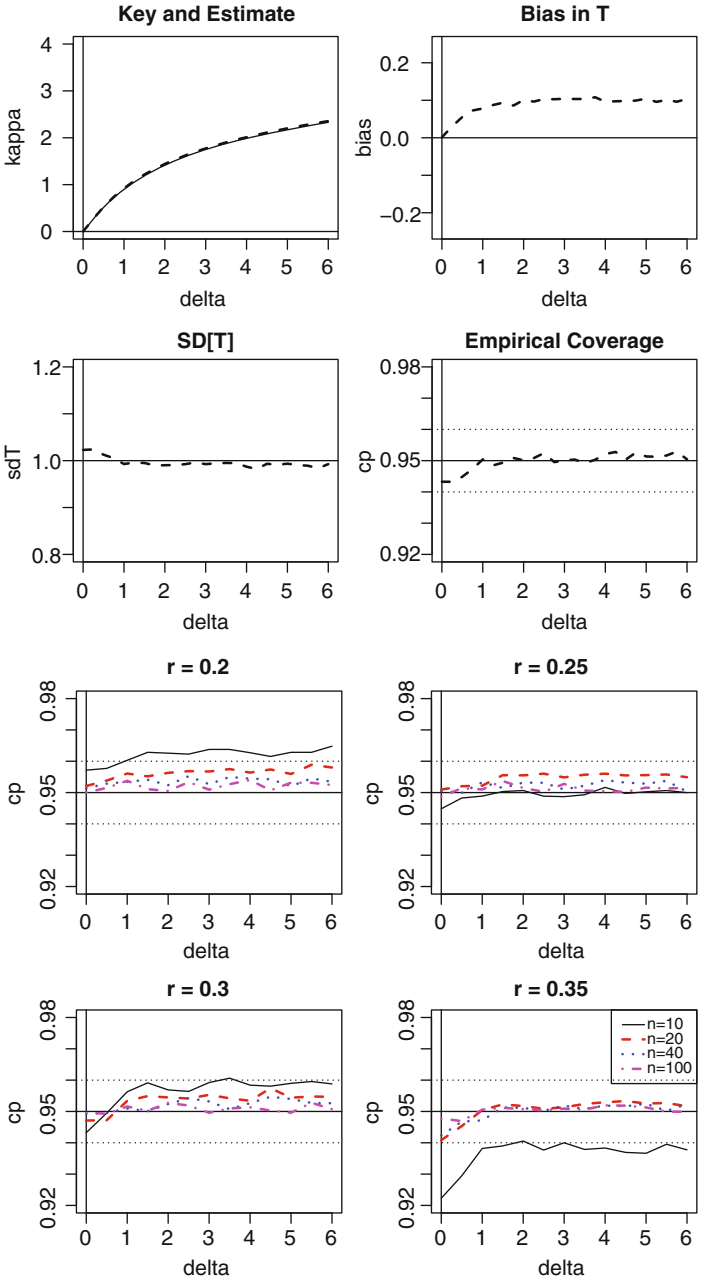


Fig. 22.1 Normal Model: The top left plot shows the graph of the Key (22.11) (*solid line*) together with the estimate $\hat{\kappa}$ (*dashed black line*) based on $n = 15$ observations when $r = 0.25$. The top right plot shows $\sqrt{15}$ times the bias of $\hat{\kappa}$. The bottom left plot depicts the standard deviations of T_{15}; and the bottom right plot shows the empirical coverages of nominal 95 % confidence intervals for δ based on T_{15}. Below are examples of empirical coverage probabilities based on the VST for $r = 0.2, 0.25, 0.3$, and 0.35 and sample sizes $n = 10, 20, 40$, and 100

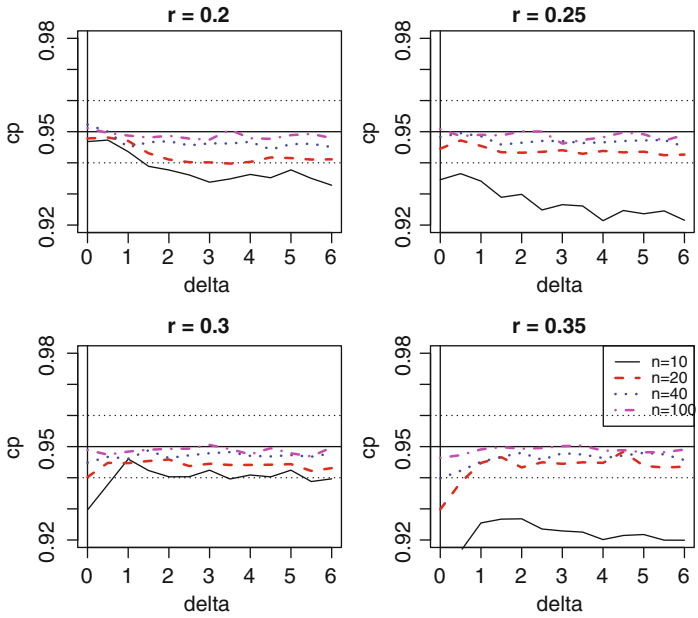


Fig. 22.2 Uniform Model: Plots of empirical coverage probabilities based on the variance stabilizing transformation (VST) defined by (22.10) for $r = 0.2, 0.25, 0.3$, and 0.35. The notation and sample sizes are the same as for the normal model, and indeed, for the examples to follow

Table 22.2 Uniform model constants for obtaining a variance stabilizer of $\hat{\delta}_r$

r	IQR_r	$g(r)$	$h(r)$	a_r	b_r	c_r
0.20	1.2	2	0.980	1.2	0.667	1.021
0.25	1.0	2	1.000	1.0	1.000	1.000
0.30	0.8	2	0.980	0.8	1.500	1.021
0.35	0.6	2	0.917	0.6	2.333	1.091
0.40	0.4	2	0.800	0.4	4.000	1.250

intervals are quite good over the entire range of δ for $n \geq 20$ and $r = 0.2$ to $r = 0.3$; see Fig. 22.2.

22.3.3 Logistic

For the location-scale family generated by the the logistic distribution function $F(x) = 1/(1 + e^{-x})$, the required constants for variance stabilization are given in Table 22.3. In Fig. 22.3 are shown examples of empirical coverages of nominal 95 % confidence intervals for δ. These plots demonstrate that sample sizes of 20 or more will yield accurate confidence intervals for δ for these values of r.

Table 22.3 Logistic model constants for obtaining a variance stabilizer of $\hat{\delta}_r$

r	IQR_r	$g(r)$	$h(r)$	a_r	b_r	c_r
0.20	2.773	6.250	3.062	1.386	1.220	0.653
0.25	2.197	5.333	2.667	1.099	1.473	0.750
0.30	1.695	4.762	2.333	0.847	1.895	0.857
0.35	1.238	4.396	2.014	0.619	2.647	0.993
0.40	0.811	4.167	1.667	0.405	4.224	1.200

22.3.4 Cauchy

For the location-scale family generated by the the Cauchy distribution with density $f(x) = 1/\{\pi(1 + x^2)\}$, the required constants for variance stabilization are given in Table 22.4. In Fig. 22.4 are shown examples of empirical coverages of nominal 95 % confidence intervals for δ. Much larger sample sizes are required to obtain adequate coverage than in the previous examples; this could be explained by the lack of moments for the Cauchy model.

22.3.5 Double Exponential

For the location-scale family generated by the the double exponential distribution with density $f(x) = e^{-|x|}/2$, the required constants for variance stabilization are

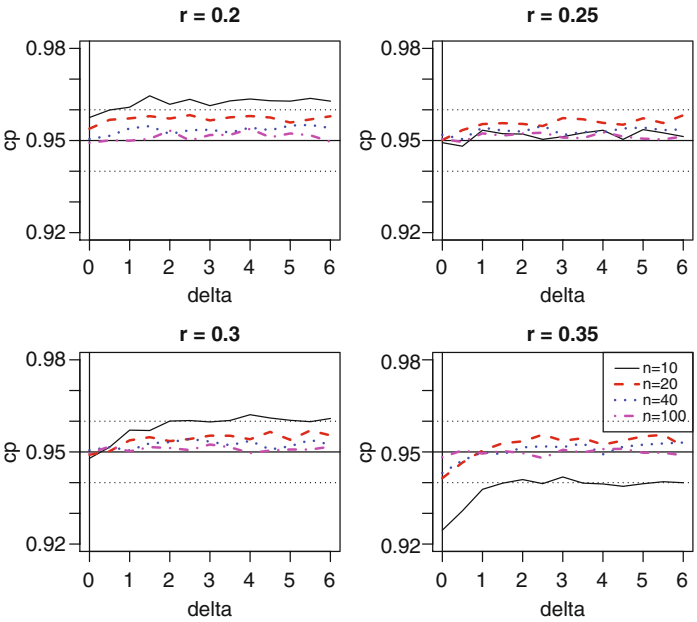


Fig. 22.3 Logistic Model: Examples of empirical coverage probabilities based on the variance stabilizing transformation (VST) defined by (22.10) for $r = 0.2, 0.25, 0.3$, and 0.35

Table 22.4 Cauchy model constants for obtaining a variance stabilizer of $\hat{\delta}_r$

r	IQR_r	$g(r)$	$h(r)$	a_r	b_r	c_r
0.20	2.753	9.093	4.455	1.752	2.619	0.353
0.25	2.000	6.283	3.142	1.273	2.467	0.500
0.30	1.453	4.800	2.351	0.925	2.619	0.668
0.35	1.019	3.957	1.813	0.649	3.167	0.866
0.40	0.650	3.473	1.389	0.414	4.571	1.131

given in Table 22.5. In Fig. 22.5 are shown examples of empirical coverages of nominal 95 % confidence intervals for δ. Sample sizes of 100 or more appear to be required to obtain satisfactory coverage, especially for δ near 0.

22.4 Testing for the Standardized Median

22.4.1 Choosing r

One of the reasons for estimating a parameter such as $\delta_r = \alpha/(\beta\,\text{IQR})$ for the location-scale family $F_{\alpha,\beta}$ is that its sample version $\hat{\delta}_r = M/R_r$ is then automatically robust to outliers in the sense that the breakdown point is $\min\{r, 1-2r\}$. The examples in the last section show that for short-tailed distributions, much smaller sample sizes

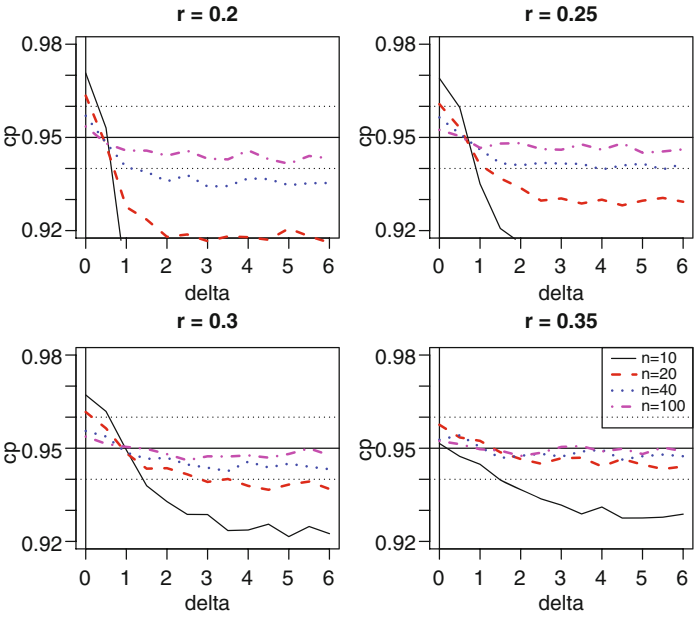


Fig. 22.4 Cauchy Model: Examples of empirical coverage probabilities of nominal 95 % confidence intervals for δ based on the variance stabilizing transformation (VST) given by (22.10) for $r = 0.2, 0.25, 0.3$, and 0.35

Table 22.5 Double exponential model constants for obtaining a variance stabilizer of $\hat{\delta}_r$

r	IQR_r	$g(r)$	$h(r)$	a_r	b_r	c_r
0.20	1.833	5.000	2.449	1.833	1.787	0.408
0.25	1.386	4.000	2.000	1.386	2.081	0.500
0.30	1.022	3.333	1.633	1.022	2.555	0.612
0.35	0.713	2.857	1.309	0.713	3.369	0.764
0.40	0.446	2.500	1.000	0.446	5.021	1.000

are required to obtain good coverage probability of δ_r, regardless of the choice of r ranging from 0.2 to 0.35.

Also, the *Key* appears to be a decreasing function of r as it increases from 0 to 0.5, even though δ_r is increasing in r. Why is a large *Key* important? Consider the Neyman–Pearson setting: a test of the null hypothesis $\delta_r = 0$ against the alternative $\delta_r > 0$ is to be based on the VST T_n defined by (22.10). Because it is asymptotically normal with mean $\sqrt{n}\,\mathcal{K}(\delta_r)$, variance 1, one can reject the hypothesis at level α if $T_n \geq z_{1-\alpha}$. Further, the power against alternative $\delta_r > 0$ at this level is readily shown to be

$$\Pi(\delta_r) \doteq \Phi(\sqrt{n}\,\mathcal{K}(\delta_r) - z_{1-\alpha})\,. \tag{22.12}$$

This shows that larger $\mathcal{K}$ means more power at any given level, so one would want to choose r small to maximize power. Of course the test is easily broken down by

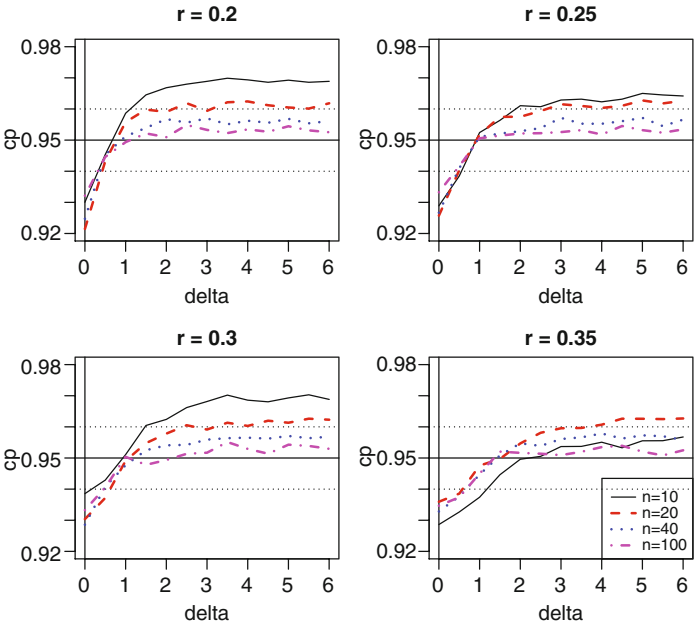


Fig. 22.5 Double Exponential Model: Plots of empirical coverage probabilities of confidence intervals for δ based on the VST for $r = 0.2$,0.25,0.3 and $r = 0.35$

outliers as r becomes small. Thus, a compromise value such as $r = 0.25$ to obtain good power and outlier resistance is recommended.

We remark that the VST T_n can be interpreted as the *evidence for the alternative hypothesis*. Its expectation factors into the square root of the sample size and the *Key*, so the latter can be interpreted as the expected evidence per $\sqrt{n}$ observation. Moreover, T_n has a natural calibration scale in terms of its standard error, which is unit normal. Thus, a value of $T_n = 4$ is interpreted as evidence 4 ± 1 for the alternative that δ_r is positive. Values near 1.645 are considered "weak" evidence, 3.3 "moderate" evidence, and 5 "strong" evidence, for the alternative, see (Kulinskaya et al. 2008). As seen earlier in this paper, the *Key* plays an important role in estimation by confidence intervals. Another advantage of variance stabilized statistics is that they can be readily combined in a meta-analysis of effects from multiple studies, as shown in (Kulinskaya et al. 2008; Malloy et al. 2013; and Morgenthaler and Staudte 2012).

22.4.2 Toward a Distribution-Free Version

The choice of constant a_r, b_r, and c_r requires knowledge of $g(r)$ and $g(0.5)$ where $g(p) = 1/f(x_p)$ for the f that generates the location-scale family. The Bloch and Gastwirth (1968) and Siddiqui (1960) estimator of $g(0.5)$, for example, divides the sample median by a normalized IQR $R_{r(n)}$, where $r = r(n) = 1/2 - n^{-\gamma}$, with $0 < \gamma < 1$ guaranteeing consistency of the estimator. Alternative estimates of $g(0.5)$ are also available in (McKean and Schrader 1984 and Sheather and McKean 1987). It is possible that the theory provided herein for a fixed range can be combined with such estimates of the variance of quantiles to provide a desirable data-driven choice of constants, and hence greatly reduce the assumptions on f.

22.5 Summary

We have derived a VST for the sample standardized median and demonstrated how to use it for inference in the form of confidence intervals and tests. Evaluation of examples shows that the sample sizes required for practical usage vary greatly, and are much larger depending on the fatness of the tails of the underlying location-scale family (Cauchy) or the peakedness at the median (double-expondential). It is possible to extend the results to the asymmetric case, and it would be insightful to relate the constants of the VST to measures of 'peakedness' based on quantiles as for example found in (Jones and Pewsey 2011 and Ruppert 1987), and references therein. Even better it would be useful to derive a distribution-free approach based on estimation of the constants, which now require specification of the location-scale family.

Acknowledgements The author thanks both referees and the editor for helpful comments which have greatly improved the manuscript. Also, many thanks to Prof. Hira Lal Koul for his friendship and introducing the author to robust statistics, as well as imparting his immense passion for research.

References

Bickel PJ, Doksum KA (1977) Mathematical Statistics: basic ideas and selected topics. Holden–Day, San Francisco

Bloch DA, Gastwirth JL. (1968) On a simple estimate of the reciprocal of the density function. Annals Math Stat 39(3):1083–1085

DasGupta A. (2006) Asymptotic Theory of Statistics and Probability. Springer. DOI: 10.1007/978-0-387-75971-5.

Hall P, Sheather SJ (1988) On the distribution of a studentized quantile. J Royal Stat Soci, Ser B, 50:381–391

Hyndman RJ, Fan Y (1996) Sample quantiles in statistical packages. Ame Stat 50:361–365

Rosco JF, Jones MC, Pewsey A (2011) Skewness invariant measures of kurtosis. Am Stat 65(2): 89–95

McKean JW, Schrader RM (1984) A comparison of methods for studentizing the median. Commun Stat–Simul Compu 13(6):751–773

Kulinskaya E, Morgenthaler S, Staudte RG (2008) Meta Analysis: a Guide to Calibrating and Combining Statistical Evidence. Wiley Series in Probability and Statistics. Wiley, Chichester. ISBN 978-0-470-02864-3.

Malloy M, Prendergast L, Staudte RG (2013) Transforming the model-T: random effects meta-analysis with stable weights. Stat Medi 32:1842–1864. DOI: 10.1002/sim.5666.

Morgenthaler S, Staudte RG (2012) Advantages of variance stabilization. Scandinav J Stat 39: 714–728. DOI: 10.1111/j.1467-9469.2011.00768.x.

Ruppert D (1987) What is kurtosis? an influence function approach. Am Stat 41(1):1–5

Sheather SJ, McKean JW (1987) A comparison of testing and confidence intervals for the median. Stat Probab Lett 6:31–36

Siddiqui MM (1960) Distribution of quantiles in samples from a bivariate population. J Rese Nation Bureau Stand Seri B 64:1960

R Development Core Team (2008) R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna. ISBN 3-900051-07-0

Chapter 23
Efficient Quantile Regression with Auxiliary Information

Ursula U. Müller and Ingrid Van Keilegom

23.1 Introduction

Completely observed data. The quantile regression model (Koenker and Bassett 1978; Koenker 2005) for a random sample (X_i, Y_i), $i = 1, \ldots, n$, assumes that the conditional quantile of a response variable Y given a covariate vector X can be modeled parametrically, i.e. it can be written as a parametric quantile regression function $q_\theta(X)$, $\theta \in \mathbb{R}^d$. In this chapter, we consider, more generally, a class of regression models that can be written in the form

$$E\{a_\vartheta(X, Y)|X\} = 0, \quad a_\theta = (a_{1\theta}, \tilde{a}_\theta^\top)^\top, \tag{23.1}$$

where the true parameter ϑ belongs to the interior of some compact parameter space $\Theta \subset \mathbb{R}^d$. The first component of the k-dimensional vector a_θ is

$$a_{1\theta}(X, Y) = p - 1\{Y - q_\theta(X) < 0\}, \quad p \in (0, 1).$$

This specifies the familiar quantile regression model since

$$\begin{aligned} 0 &= E\{a_{1\vartheta}(X, Y)|X\} = E[p - 1\{Y - q_\vartheta(X) < 0\}|X] \\ &\iff P\{Y < q_\vartheta(X)|X\} = p. \end{aligned} \tag{23.2}$$

The vector $\tilde{a}_\theta$ represents auxiliary information in the form of $k - 1$ conditional parametric constraints. This is the case, for example, if there are reliable parametric models for certain moments of the conditional distribution of Y given X, including the conditional mean and the conditional variance.

I. Van Keilegom (✉)
Institut de statistique, Université catholique de Louvain,
Voie du Roman Pays 20, B-1348 Louvain-la-Neuve, Belgium
e-mail: ingrid.vankeilegom@uclouvain.be

U. U. Müller
Department of Statistics, Texas A&M University,
College Station, TX 77843-3143, USA
e-mail: uschi@stat.tamu.edu

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_23,

Note that the number of parameters, d, and the number of equations, k, are unrelated. Later on we will transform the equations so as to obtain as many equations as parameters. Also, note that the vector θ contains the parameters that determine the model for the p-th quantile, as well as the parameters on which the auxiliary information depends. Usually when we have auxiliary information the latter set is part of (or equal to) the former, but this is not necessarily the case.

We are interested in finding an efficient estimator for ϑ. Efficient estimation of ϑ in model (23.1), with an arbitrary vector a_θ of constraints, has been addressed by Müller and Van Keilegom 2012. There we also briefly discuss the quantile regression model (23.2) as an example of a one-dimensional constraint, without assuming the presence of the additional vector $\tilde{a}_\theta$. Let us discuss this model first. The usual estimator under model (23.2) is based on the check function approach, and solves the quantile regression estimating equation

$$\sum_{i=1}^{n} \dot{q}_\theta(X_i)[p - 1\{Y_i - q_\theta(X_i) < 0\}] = 0$$

with respect to θ (see e.g., Koenker 2005) or more precisely, it minimizes

$$\Big\| \sum_{i=1}^{n} \dot{q}_\theta(X_i)[p - 1\{Y_i - q_\theta(X_i) < 0\}] \Big\| \tag{23.3}$$

with respect to θ, where $\|\cdot\|$ denotes the Euclidean norm, since an exact solution of the earlier equation might not exist. Here, $\dot{q}_\theta(X)$ denotes the $(d \times 1)$-vector of partial derivatives $\partial/(\partial\theta_j)\, q_\theta(X)$, $j = 1, \ldots, d$. This is indeed an unbiased estimating equation since

$$E(\dot{q}_\vartheta(X)[p - 1\{Y - q_\vartheta(X) < 0\}]) = E(\dot{q}_\vartheta(X) E[p - 1\{Y - q_\vartheta(X) < 0\}|X]) = 0.$$

This calculation shows that one could, more generally, obtain a consistent estimator $\hat{\vartheta}$ by minimizing the norm of a weighted sum

$$\Big\| \sum_{i=1}^{n} W_\theta(X_i)[p - 1\{Y_i - q_\theta(X_i) < 0\}] \Big\|,$$

where W_θ is a d-dimensional vector of weights. Müller and Van Keilegom 2012 proved that an asymptotically efficient estimator of ϑ is obtained for the weight vector

$$W_\theta(X) = -\frac{f_{Y|X}\{q_\theta(X)\}\dot{q}_\theta(X)}{p^2 + (1-2p)F_{Y|X}\{q_\theta(X)\}}, \tag{23.4}$$

with $f_{Y|X}(y) = \frac{d}{dy} F_{Y|X}(y)$ (provided it exists) and $F_{Y|X}(y) = P\{Y \le y|X\}$. A simpler (but asymptotically equivalent) version of this estimator is based on weights

$$f_{Y|X}\{q_\theta(X)\}\dot{q}_\theta(X), \tag{23.5}$$

since the denominator in (23.4) equals $p - p^2$ for $\theta = \vartheta$ and hence it does not need to be estimated. The weight vector is undetermined in both cases: it involves

the unknown conditional density $f_{Y|X}\{q_\theta(X)\}$ (and $F_{Y|X}$ in the first case) and must therefore be replaced by a suitable consistent estimator. Using these estimated weight vectors in the estimating previous equation will yield two asymptotically efficient estimators of ϑ.

Note that if we use the simpler weights (23.5), then an asymptotically efficient estimator of ϑ is obtained by minimizing

$$\Big\| \sum_{i=1}^{n} f_{Y|X}\{q_\theta(X_i)\} \dot{q}_\theta(X_i)[p - 1\{Y_i - q_\theta(X_i) < 0\}] \Big\|, \tag{23.6}$$

which is different from the widely used and commonly accepted estimator given in (23.3), that corresponds to the check function approach, and that is in fact not efficient.

In this chapter, we consider model (23.1) which, aside from (23.2), assumes that auxiliary information in the form of a constraint $E\{\tilde{a}_\vartheta(X, Y)|X\} = 0$ is available. This is related to Tang and Leng 2012 who consider the *linear* quantile regression model with $q_\beta(X) = X^\top \beta$, where β is a parameter vector. They assume additional information in the form of an *unconditional* constraint, $E\{\tilde{a}_\vartheta(X, Y)\} = 0$, and suggest an empirical likelihood approach. Such a constraint applies if, for example, there is knowledge about unconditional moments of the joint distribution of (X, Y). This is conceptually different from our model, since models for moments of the conditional distribution are not included, e.g., models for the conditional mean $E(Y|X)$ or the variance function mentioned earlier. Another related paper that does consider a conditional constraint is by Qin and Wu (2001), who estimate conditional quantiles. However, neither the quantiles nor the constraint are modeled parametrically. There is more literature dating back several years on estimating unconditional quantiles when auxiliary information is available; see, e.g., (Kuk and Mak 1989; Rao et al. 1990; Zhang 1997).

Missing data. As in Müller and Van Keilegom (2012) we now assume further that some responses Y are allowed to be *missing at random* (MAR). This means that one has independent identically distributed (i.i.d.) observations $(X_i, \delta_i Y_i, \delta_i)$, $i = 1, \ldots, n$, having the same distribution as $(X, \delta Y, \delta)$, with indicator $\delta = 1$ if Y is observed and $\delta = 0$ if Y is missing. In particular, one assumes that the missingness mechanism depends only on X,

$$P(\delta = 1|X, Y) = P(\delta = 1|X) = \pi(X),$$

where $\pi(\cdot)$ is the propensity score. This implies that Y and δ are conditionally independent given X. One reason for considering the MAR model is that it contains the "full model", where no data are missing, as a special case with $\pi(\cdot) = 1$ (and all indicators $\delta = 1$), so both models can be treated together. Of course, this does not always apply since the construction of reasonable estimators can be quite different, depending on the model. Here we are specifically interested in estimating the *parameter* ϑ. In this case, we can work with a simple complete case estimator (an estimator

for the full model that ignores observations that are only partially observed). One possibility is to estimate ϑ by a minimizer of

$$\Big\| \sum_{i=1}^{n} \delta_i W_\theta(X_i) a_\theta(X_i, Y_i) \Big\|$$

with respect to θ, where W_θ is a $d \times k$ weight matrix. In this way, we obtain a system of d equations with d unknown parameters, although we started off with k constraints. That the weighted previous sum leads to an unbiased estimator is easy to see using the fact that δ and Y are conditionally independent given X under the MAR assumption. Beyond that, one can show that a complete case version of *any* consistent estimator of ϑ is again consistent. This can be seen by applying the *transfer principle for complete case statistics*, introduced by Koul et al. 2012, which makes it possible to adapt results for the full model to the MAR model. The transfer principle provides the limiting distribution of a complete case version of a statistic as the limiting distribution of that statistic conditional on $\delta = 1$. To verify consistency, one only has to show that the functional of interest, i.e., in our case the parameter vector ϑ, is the same in the unconditional and in the conditional model. This is indeed true, since ϑ is in both models defined as a solution of the same conditional constraint:

$$\begin{aligned} 0 = E\{a_\vartheta(X, Y)|X\} &= \frac{E(\delta|X) E\{a_\vartheta(X, Y)|X\}}{E(\delta|X)} = \frac{E\{\delta a_\vartheta(X, Y)|X\}}{E(\delta|X)} \\ &= E\{a_\vartheta(X, Y)|X, \delta = 1\}. \end{aligned}$$

So far we know that an efficient estimator of ϑ in the (unextended) quantile regression model with MAR responses is given as a minimizer of

$$\Big\| \sum_{i=1}^{n} \delta_i \widehat{W}_\theta(X_i)[p - 1\{Y_i - q_\theta(X_i) < 0\}] \Big\|$$

with respect to θ, where $\widehat{W}_\theta$ is a suitable estimator of the weight vector W_θ given in (23.4) or (23.5) (see Sect. 3.4 in Müller and Van Keilegom 2012).

In the next section, we will provide the formulas for an efficient estimator of ϑ in the general quantile regression model (23.1) with auxiliary information in the form of a general conditional constraint. In Sect. 23.3 we discuss three examples of auxiliary information, namely when we have a parametric model for the mean and for the median, respectively, and when we have two responses that share the same p-th quantile. Section 23.4 shows the results of a small simulation study, and we end this chapter in Sect. 23.5 with some general conclusions.

23.2 The Estimator

As in Müller and Van Keilegom (2012) we write

$$\ell_\theta(X,Y) = -W_\theta(X)a_\theta(X,Y), \qquad I = E\{\delta\ell_\vartheta(X,Y)\ell_\vartheta(X,Y)^\top\},$$

$$W_\theta(X) = [\frac{\partial}{\partial\theta}E\{a_\theta(X,Y)|X\}]^\top A_\theta(X)^{-1}, \tag{23.7}$$

now with

$$a_\theta(X,Y) = \begin{pmatrix} a_{1\theta}(X,Y) \\ \tilde{a}_\theta(X,Y) \end{pmatrix} = \begin{pmatrix} p - 1\{Y - q_\theta(X) < 0\} \\ \tilde{a}_\theta(X,Y) \end{pmatrix}, \tag{23.8}$$

where for $\theta = \vartheta$ the $k \times k$ matrix $A_\vartheta(X)$ is given by

$$A_\vartheta(X) = E\{a_\vartheta(X,Y)a_\vartheta(X,Y)^\top|X\},$$

and where for $\theta \neq \vartheta$, the matrix $A_\theta(X)$ is obtained by replacing ϑ by θ in the expression of $A_\vartheta(X)$. Note that, in general $A_\theta(X)$ and $E\{a_\theta(X,Y)a_\theta(X,Y)^\top|X\}$ are different, since in certain entries of the matrix $A_\vartheta(X)$, the parameter ϑ will disappear when using the underlying model assumptions. For example, the first entry is $E([p - 1\{Y - q_\vartheta(X) < 0\}]^2|X) = p^2 + (1-2p)F_{Y|X}\{q_\vartheta(X)\} = p - p^2$, which is independent of ϑ.

The estimator in model (23.1) can then be written $\hat{\vartheta} = \text{argmin}_\theta \|\sum_{i=1}^n \delta_i \ell_\theta(X_i, Y_i)\|$. In the full model, we simply set $\delta_i = 1$ for $i = 1, \ldots, n$, i.e., the indicators δ can be ignored. Under the assumptions stated in Müller and Van Keilegom (2012), $\hat{\vartheta}$ is asymptotically linear,

$$n^{1/2}(\hat{\vartheta} - \vartheta) = I^{-1}n^{-1/2}\sum_{i=1}^n \delta_i \ell_\vartheta(X_i, Y_i) + o_p(1),$$

and efficient in the sense of Hájek and Le Cam.

Let us take a closer look at the formula of the weight matrix. The estimating equation for model (23.1) involves W_θ and a_θ given in equations (23.7) and (23.8). Using the specific form of a_θ, the matrix $W_\theta(X)$ computes to

$$W_\theta(X) = \left(-f_{Y|X}\{q_\theta(X)\}\dot{q}_\theta(X) \quad \frac{\partial}{\partial\theta}E\{\tilde{a}_\theta(X,Y)^\top|X\}\right) A_\theta(X)^{-1}, \tag{23.9}$$

where $A_\vartheta(X)$ is the matrix

$$\begin{pmatrix} p - p^2 & E([p - 1\{Y < q_\vartheta(X)\}]\tilde{a}_\vartheta(X,Y)^\top|X) \\ E([p - 1\{Y < q_\vartheta(X)\}]\tilde{a}_\vartheta(X,Y)|X) & E\{\tilde{a}_\vartheta(X,Y)\tilde{a}_\vartheta(X,Y)^\top|X\} \end{pmatrix},$$

and where the matrix $A_\theta(X)$ is obtained by replacing in the formula of $A_\vartheta(X)$ every ϑ that does not disappear after using the model assumptions, by θ.

In Sect. 3.1 of Müller and Van Keilegom (2012), it is shown that if we replace the weight matrix $W_\theta(X)$ given in (23.9) by an estimator $\widehat{W}_\theta(X)$ that is uniformly consistent in θ and x, i.e., $\sup_{\theta\in\Theta}\sup_x \|\widehat{W}_\theta(x) - W_\theta(x)\| = o_p(1)$, then the resulting estimator (that depends now on $\widehat{W}_\theta$ instead of W_θ) remains asymptotically efficient.

Note that the weight matrix $W_\theta(X)$ involves, among others, the conditional density $f_{Y|X}$. The density can be estimated by using e.g., a kernel smoother of the form

$$\hat{f}_{Y|X=x}(y) = \sum_{i=1}^{n} \frac{\delta_i k_b(x - X_i) k_h(y - Y_i)}{\sum_{i=1}^n \delta_i k_b(x - X_i)},$$

with kernel k and smoothing parameters b and h, and where $k_b(\cdot) = k(\cdot/b)/b$ for any bandwidth b. The estimation of the other components of the weight matrix $W_\theta(X)$ depends on the specific form of the auxiliary information. We will consider three examples in the next section.

23.3 Examples

Example 1. We start with a situation in which we have some auxiliary information concerning the conditional mean $r(X) = E(Y|X)$. Suppose that $r(X)$ can be modeled parametrically $r(X) = r_\vartheta(X)$. The function $\tilde{a}_\theta(X, Y)$ is given by

$$\tilde{a}_\theta(X, Y) = Y - r_\theta(X),$$

i.e., $k = 2$ and, for example, $r_\theta(X) = \theta^\top X$. Some straightforward algebra shows that the optimal weight matrix is then given by

$$W_\theta(X) = \big(-f_{Y|X}\{q_\theta(X)\}\dot{q}_\theta(X) \quad -\dot{r}_\theta(X)\big) A_\theta(X)^{-1},$$

where $A_\theta(X)$ is the 2×2 matrix

$$\begin{pmatrix} p - p^2 & p\, r_\theta(X) - E(1\{Y < q_\theta(X)\}]Y|X) \\ p\, r_\theta(X) - E(1\{Y < q_\theta(X)\}]Y|X) & \mathrm{Var}(Y|X) \end{pmatrix}.$$

The conditional variance $\mathrm{Var}(Y|X)$ can be estimated by standard kernel smoothers, whereas a consistent estimator of the term $E(1\{Y < q_\theta(X)\}]Y|X)$ in the off-diagonal element of the matrix $A_\theta(X)$ is given by

$$\sum_{i=1}^{n} \frac{\delta_i k_b(x - X_i) 1\{Y_i < q_\theta(X_i)\} Y_i}{\sum_{i=1}^n \delta_i k_b(x - X_i)},$$

with kernel k and smoothing parameter b.

Example 2. Let us now consider the case when $p \neq 1/2$, i.e., we want to estimate quantiles other than the median, and we have some auxiliary information regarding

the median. For instance, we know that the p-th quantile and the median are parallel functions (of X). Let us denote the parametric model for the median by v_θ, and so

$$\tilde{a}_\theta(X, Y) = 1/2 - 1\{Y - v_\theta(X) < 0\}.$$

In this case, it is easily seen that

$$W_\theta(X) = \big(-f_{Y|X}\{q_\theta(X)\}\dot{q}_\theta(X) \quad - f_{Y|X}\{v_\theta(X)\}\dot{v}_\theta(X)\big)A_\theta(X)^{-1},$$

where

$$A_\theta(X) \quad = \begin{pmatrix} p - p^2 & p \wedge (1/2) - p/2 \\ p \wedge (1/2) - p/2 & 1/4 \end{pmatrix},$$

since for $\theta = \vartheta$ the off-diagonal element is given by

$$\begin{aligned} &p/2 - pF_{Y|X}\{v_\vartheta(X)\} - (1/2)F_{Y|X}\{q_\vartheta(X)\} + F_{Y|X}\{q_\vartheta(X) \wedge v_\vartheta(X)\} \\ &\quad = -p/2 + \{p \wedge (1/2)\}. \end{aligned}$$

The estimation of this weight matrix only involves the estimation of the conditional density $f_{Y|X}$, which was discussed in the previous section.

Example 3. The model considered in this chapter can be extended to the case where we have a multivariate response $Y = (Y_1, \ldots, Y_{d_Y})^\top$. For simplicity, we consider the bivariate case. Let $(X_i, \delta_i Y_i, Y_i)$, $i = 1, \ldots, n$, be an i.i.d. sample, where $Y_i = (Y_{1i}, Y_{2i})^\top$, $X_i = (X_{1i}, \ldots, X_{d_X i})^\top$ and $\delta_i = (\delta_{1i}, \delta_{2i})^\top$. We could then consider the case where the two responses have the same conditional p-th quantile, which means that we should take

$$a_{1\theta}(X, Y) = p - I\{Y_1 - q_\theta(X) < 0\},$$

and

$$\tilde{a}_\theta(X, Y) = p - I\{Y_2 - q_\theta(X) < 0\},$$

and the true parameter vector ϑ satisfies $F_{Y_1|X}\{q_\vartheta(X)\} = p = F_{Y_2|X}\{q_\vartheta(X)\}$. As in the previous two examples, it can be seen that the weight matrix that leads to an asymptotically efficient estimator of ϑ is given by

$$W_\theta(X) = \big(-f_{Y_1|X}\{q_\theta(X)\} \quad - f_{Y_2|X}\{q_\theta(X)\}\big)\dot{q}_\theta(X)A_\theta(X)^{-1}$$

with

$$A_\theta(X) \quad = \begin{pmatrix} p - p^2 & F_{Y_1,Y_2|X}\{q_\theta(X), q_\theta(X)\} - p^2 \\ F_{Y_1,Y_2|X}\{q_\theta(X), q_\theta(X)\} - p^2) & p - p^2 \end{pmatrix},$$

where $F_{Y_1,Y_2|X}(y_1, y_2) = P(Y_1 \le y_1, Y_2 \le y_2|X)$. To verify the formula for the off-diagonal element of $A_\theta(X)$ consider the corresponding entry in $A_\vartheta(X)$, which computes to

$$\begin{aligned} &p^2 - p[F_{Y_1|X}\{q_\vartheta(X)\} + F_{Y_2|X}\{q_\vartheta(X)\}] + F_{Y_1,Y_2|X}\{q_\vartheta(X), q_\vartheta(X)\} \\ &\quad = -p^2 + F_{Y_1,Y_2|X}\{q_\vartheta(X), q_\vartheta(X)\}. \end{aligned}$$

Table 23.1 Simulated MSEs of parameter estimators of $q_\vartheta(X) = \vartheta_1 + \vartheta_2 X$

		Estimators of ϑ_1			Estimators of ϑ_2		
$\sigma(x)$	p	QR1	QR2	EFF	QR1	QR2	EFF
$0.6 - 0.5x$	0.25	0.00632	0.00637	0.00143	0.01083	0.00805	0.00271
	0.5	0.00546	0.00513	0.00095	0.00902	0.00601	0.00190
	0.75	0.00523	0.00502	0.00177	0.00610	0.00438	0.00358
$0.6 + 0.5x$	0.25	0.00248	0.00151	0.00111	0.00294	0.00158	0.00169
	0.5	0.00287	0.00177	0.00088	0.00495	0.00250	0.00110
	0.75	0.00311	0.00209	0.00169	0.00579	0.00365	0.00149

The table entries give the simulated mean squared errors (MSE) for three estimators of ϑ_1 and ϑ_2. The estimator "QR1" is based on the check function approach (23.3), the estimator "QR2" is the minimizer of (23.6), and "EFF" is the efficient estimator that uses the auxiliary information that r_ϑ is linear, $r_\vartheta(X) = \vartheta_3 X$.

23.4 Simulations

In order to gain some insight into the performance of our proposed method if n is finite, we conducted a small simulation study based on 50,000 simulated samples of size $n = 100$. In this study, we consider only the case where all responses are observed. Since our estimator for missing data is a complete case statistic, this essentially means that we use all $n = 100$ data pairs (X, Y), and not just a proportion. The comparisons are equally meaningful.

We considered the scenario from Example 1 in the previous section, with a linear quantile regression function $q_\vartheta(X) = \vartheta_1 + \vartheta_2 X$, and with the auxiliary information that the mean regression function is linear as well, $E(Y|X) = r_\vartheta(X) = \vartheta_3 X$. The parameters of interest are ϑ_1 and ϑ_2, whereas ϑ_3 can be regarded a nuisance parameter. In order to create this scenario, we generated responses Y given $X = x$ from a normal distribution with mean $r_\vartheta(x) = \vartheta_3 x$ (with $\vartheta_3 = 1$) and standard deviation $\sigma(x) = a + bx$. Modeling r_ϑ and σ linearly suffices to ensure that the quantile function is also linear: we have $p = \Phi[\{q_\vartheta(x) - r_\vartheta(x)\}/\sigma(x)]$ (see (23.2)), with Φ the distribution function of the standard normal distribution. Solving this with respect to $q_\vartheta(x)$ gives $q_\vartheta(x) = \vartheta_1 + \vartheta_2 x$ with $\vartheta_1 = \vartheta_1(p) = a\Phi^{-1}(p)$, and $\vartheta_2 = \vartheta_2(p) = \vartheta_3 + b\Phi^{-1}(p)$. The covariates X were generated from a uniform$(-1, 1)$ distribution.

The results are in Table 23.1. For simplicity, we used the true (3×2) weight matrix W_θ to implement our efficient estimator. We compared it with the two estimators discussed in the introduction that use only the quantile regression structure, namely the check function approach (23.3) and the estimator that minimizes (23.6) (based on weights (23.5)). To compute the latter estimator we also used the true weights. Comparing the estimators that employ the true weights with the check function approach (which does not require estimation of weights) may not be quite fair, we nevertheless find it interesting since the results make us feel confident that our estimator will outperform the usual approach even if an estimated weight matrix $\widehat{W}_\theta$ is used.

Let us briefly discuss these results. We considered two different slopes for $\sigma(x) = 0.6 + bx$, namely $b = -0.5$ and $b = 0.5$. The first case yields a variance reduction and the second case a variance gain as x increases from -1 to 1. In most cases, the

efficient estimator (EFF) is clearly better than the two approaches that do not exploit the auxiliary information. This is remarkable, in particular when one considers that the optimization algorithm involves an additional parameter. The efficient estimator performs best in the case of a conditional median ($p = 0.5$), which is not surprising since we use a normal density $f_{Y|X}$ in our simulations. The conditional median q_ϑ and the conditional mean r_ϑ are the same in this setting.

The proposed estimator lacks performance only in the case $p = 0.25$ with an increasing variance ($b = 0.5$). Here, the estimator QR2 is slightly better for ϑ_2. Simulations with larger sample sizes confirm, however, that our estimator indeed outperforms QR2 *asymptotically*. (For example, for $n = 500$ our simulated MSEs for QR2 and EFF were 0.00097 and 0.00048, respectively.)

Comparing the two estimators QR1 and QR2 that use only the quantile regression structure, we notice that both estimate the intercept similarly well in the case of a decreasing variance function. In all other cases, the weighted estimator QR2 is better than the check function approach QR1. Since QR2 is efficient in the original quantile regression model that does not assume auxiliary information, this corresponds to the theoretical findings.

23.5 Concluding Remarks

In this chapter, we studied a parametric quantile regression model in which the responses are allowed to be missing at random (but do not have to be), and in which, the covariates are always observed. We were interested in the estimation of a particular conditional quantile when auxiliary information regarding that quantile is available. We constructed an asymptotically efficient estimator of the model parameters based on weighted estimating equations, and studied three examples in more detail. One of these examples was further examined via a small simulation study, which confirmed the effectiveness of the proposed estimation procedure.

There are numerous other situations where auxiliary information is available. We could, for example, have information regarding the variance, the interquartile range, or the quantile of order $1 - p$. It would also be interesting to study a model where responses are subject to censoring, or the case with missing covariates. Such extensions definitely seem feasible, but will be somewhat more challenging from a technical point of view. Finally, an interesting project for future work would be to develop an *efficient* empirical likelihood-based method to estimate conditional quantiles, in a similar spirit as Qin and Lawless (1994) or Tang and Leng (2012). This would provide an alternative (asymptotically equivalent) approach to exploit information in the form of (conditional) constraints. Although our estimator cannot be outperformed *asymptotically*, it is nevertheless possible that there are situations where the empirical likelihood approach performs better for moderate sample sizes, or where it has computational advantages.

Acknowledgements Sincere thanks to Hira L. Koul, an inspiration to us and to others who have guided, challenged, and inspired us. I. Van Keilegom acknowledges financial support from the European Research Council under the European Community's Seventh Framework Programme (FP7/2007-2013)/ERC Grant agreement No. 203650, from IAP research network P7/06 of the Belgian Government (Belgian Science Policy), and from the contract 'Projet d'Actions de Recherche Concertées' (ARC) 11/16-039 of the 'Communauté française de Belgique', granted by the 'Académie universitaire Louvain'. The authors would also like to thank two referees for their helpful comments.

References

Koenker R (2005) Quantile Regression. Cambridge

Koenker R, Bassett G (1978) Regression quantiles. Econometrica 46:33–50

Koul HL, Müller UU, Schick A (2012) The transfer principle: a tool for complete case analysis. Ann Statist 40:3031–3049

Kuk AYC, Mak TK (1989) Median estimation in the presence of auxiliary information. J Roy Statist Soc Ser B 51:261–269

Müller UU, Van Keilegom I (2012) Efficient parameter estimation in regression with missing responses. Electron J Stat 1200–1219

Qin J, Lawless J (1994) Empirical likelihood and general estimating equations. Ann Statist 22:300–325

Qin YS, Wu Y (2001) An estimator of a conditional quantile in the presence of auxiliary information. J Statist Plann Inference 99:59–70

Rao JNK, Kovar JG, Mantel HJ (1990) On estimating distribution functions and quantiles from survey data using auxiliary information. Biometrika 77:365–375

Tang CY, Leng C (2012) An empirical likelihood approach to quantile regression with auxiliary information. Statist Probab Lett 82:29–36

Zhang B (1997) Quantile processes in the presence of auxiliary information. Ann Inst Statist Math 49:35–55

Chapter 24
Nonuniform Approximations for Sums of Discrete m-Dependent Random Variables

P. Vellaisamy and V. Čekanavičius

24.1 Introduction

Nonuniform estimates for normal approximation are well known, see the classical results in Chap. 5 of Petrov (1995) and the references (Chen and Shao 2001, Chen and Shao 2004 and Nefedova and Shevtsova 2012) for some recent developments. On the other hand, nonuniform estimates for discrete approximations are only a few. For example, the Poisson approximation to Poisson binomial distribution has been considered in (Neammanee 2003) and translated Poisson approximation for independent lattice summands via the Stein method has been discussed in Barbour and Choi (2004). Some general estimates for independent summands under assumption of matching of pseudomoments were obtained in Čekanavičius (1993). For possibly dependent Bernoulli variables, nonuniform estimates for Poisson approximation problems were discussed in Teerapabolarn and Santiwipanont (2007). However, the estimates obtained had a better accuracy than estimates in total variation only for x larger than exponent of the sum's mean. In Čekanavičius and Petrauskienė (2011), 2-runs statistic was approximated by compound Poisson distribution. In this paper, we obtain nonuniform estimates for Poisson, compound Poisson, translated Poisson, negative binomial and binomial approximations, under a quite general set of assumptions.

We recall that the sequence of random variables $\{X_k\}_{k\geq 1}$ is called m-dependent if, for $1 < s < t < \infty$, $t - s > m$, the sigma algebras generated by $X_1, \ldots, X_s$ and $X_t, X_{t+1} \ldots$ are independent. Without loss of generality,

P. Vellaisamy (✉)
Department of Mathematics, Indian Institute of Technology Bombay,
Powai, 400076 Mumbai, India
e-mail: pv@math.iitb.ac.in

V. Čekanavičius
Department of Mathematics and Informatics,
Vilnius University, Naugarduko 24, 03225 Vilnius, Lithuania
e-mail: vydas.cekanavicius@mif.vu.lt

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0_24,

we can reduce the sum of m-dependent variables to the sum of 1-dependent ones, by grouping consecutive m summands. Therefore, we consider henceforth, without loss of generality, the sum $S_n = X_1 + X_2 + \cdots + X_n$ of non-identically distributed 1-dependent random variables concentrated on nonnegative integers.

We denote the distribution function and the characteristic function of S_n by $F_n(x)$ and $\widehat{F}_n(t)$, respectively. Similarly, for a signed measure M concentrated on the set $\mathbb{N}$ of nonnegative integers, we denote by $M(x) = \sum_{k=0}^{x} M\{k\}$ and $\widehat{M}(t) = \sum_{k=0}^{\infty} \mathrm{e}^{\mathrm{i}tk} M\{k\}$, the analogues of distribution function and Fourier-Stieltjes transform, respectively. Though our aim is to obtain the nonuniform estimates, we obtain also estimates for Wasserstein norm defined as

$$\|M\|_W = \sum_{j=0}^{\infty} |M(j)|.$$

Note that Wasserstein norm is stronger than total variation norm defined by $\|M\| = \sum_{j=0}^{\infty} |M\{j\}|$.

Next we introduce the approximations considered in this paper. Let

$$\lambda = \mathrm{E}S_n, \quad \Gamma_2 = \frac{1}{2}(\mathrm{Var}S_n - \mathrm{E}S_n).$$

For brevity, let $z(t) = \mathrm{e}^{\mathrm{i}t} - 1$. Also, let Π and Π_1 respectively denote the Poisson distribution with parameter λ and its second order difference multiplied by Γ_2. More precisely,

$$\widehat{\Pi}(t) = \exp\{\lambda z\}, \quad \widehat{\Pi}_1(t) = \widehat{\Pi}(t)\Gamma_2 z^2.$$

It is clear that $\Pi + \Pi_1$ is second-order (and, consequently, two-parametric) Poisson approximation. As an alternative to the Poisson based two-parametric approximation, we choose compound Poisson measure G with the following Fourier-Stieltjes transform

$$\widehat{\mathrm{G}}(t) = \exp\{\lambda z + \Gamma_2 z^2\}.$$

The approximation G was used in many papers, see Barbour and Čekanavičius (2002), Barbour and Xia (1999), Roos (2003) and the references therein. If $\Gamma_2 < 0$, then G becomes signed measure, which is not always convenient and natural for approximation to nonnegative S_n. Therefore, we define next three distributional approximations. Translated Poisson (TP) approximation has the following characteristic function:

$$\widehat{\mathrm{TP}}(t) = \exp\{\lfloor -2\Gamma_2 \rfloor \mathrm{i}t + (\lambda + 2\Gamma_2 + \tilde{\delta})z\} = \exp\{\lambda z + (2\Gamma_2 + \tilde{\delta})(z - \mathrm{i}t)\}.$$

Here $\lfloor -2\Gamma_2 \rfloor$ and $\tilde{\delta}$ are respectively the integer part and the fractional part of $-2\Gamma_2$, so that $-2\Gamma_2 = \lfloor -2\Gamma_2 \rfloor + \tilde{\delta}, \quad 0 \leqslant \tilde{\delta} < 1$. The TP approximation was investigated

in numerous papers, see, for example, Barbour and Čekanavičius (2002), Barbour and Choi (2004), Röllin (2005) and Röllin (2007). If $\mathrm{E}S_n < \mathrm{Var}S_n$, then one can apply the negative binomial approximation, which is defined in the following way:

$$\mathrm{NB}\{j\} = \frac{\Gamma(r+j)}{j!\Gamma(r)}\overline{q}^r(1-\overline{q})^j, \quad (j \in \mathbb{Z}_+), \quad \frac{r(1-\overline{q})}{\overline{q}} = \lambda, \quad r\Big(\frac{1-\overline{q}}{\overline{q}}\Big)^2 = 2\Gamma_2.$$

Note that

$$\widehat{\mathrm{NB}}(t) = \Big(\frac{\overline{q}}{1-(1-\overline{q})\mathrm{e}^{\mathrm{i}t}}\Big)^r = \Big(1 - \frac{(1-\overline{q})z}{\overline{q}}\Big)^{-r}.$$

If $\mathrm{Var}S_n < \mathrm{E}S_n$, the more natural approximation is the binomial one defined as follows:

$$\widehat{\mathrm{Bi}}(t) = (1+\overline{p}z)^N, \quad N = \lfloor \tilde{N} \rfloor, \quad \tilde{N} = \frac{\lambda^2}{2|\Gamma_2|}, \quad \bar{p} = \frac{\lambda}{N}\varepsilon = \tilde{N} - N.$$

Note that symbols $\overline{q}$ and $\overline{p}$ are not related and, in general, $\overline{q} + \overline{p} \neq 1$.

Finally, we introduce some technical notations, related to the method of proof. Let $\{Y_k\}_{k\geq 1}$ be a sequence of arbitrary real or complex-valued random variables. We assume that $\widehat{\mathrm{E}}(Y_1) = \mathrm{E}Y_1$ and, for $k \geqslant 2$, define $\widehat{\mathrm{E}}(Y_1, Y_2, \cdots, Y_k)$ by

$$\widehat{\mathrm{E}}(Y_1, Y_2, \cdots, Y_k) = \mathrm{E}Y_1Y_2\cdots Y_k - \sum_{j=1}^{k-1}\widehat{\mathrm{E}}(Y_1, \cdots, Y_j)\mathrm{E}Y_{j+1}\cdots Y_k.$$

Let

$$\begin{aligned}
\widehat{\mathrm{E}}^+(X_1) &= \mathrm{E}X_1, \qquad \widehat{\mathrm{E}}^+(X_1, X_2) = \mathrm{E}X_1X_2 + \mathrm{E}X_1\mathrm{E}X_2,\\
\widehat{\mathrm{E}}^+(X_1, \ldots, X_k) &= \mathrm{E}X_1\ldots X_k + \sum_{j=1}^{k-1}\widehat{\mathrm{E}}^+(X_1, \ldots, X_j)\mathrm{E}X_{j+1}X_{j+2}\cdots X_k,\\
\widehat{\mathrm{E}}_2^+(X_{k-1}, X_k) &= \widehat{\mathrm{E}}^+(X_{k-1}(X_{k-1}-1), X_k) + \widehat{\mathrm{E}}^+(X_{k-1}, X_k(X_k-1)),\\
\widehat{\mathrm{E}}_2^+(X_{k-2}, X_{k-1}, X_k) &= \widehat{\mathrm{E}}^+(X_{k-2}(X_{k-2}-1), X_{k-1}, X_k)\\
&\quad + \widehat{\mathrm{E}}^+(X_{k-2}, X_{k-1}(X_{k-1}-1), X_k).
\end{aligned}$$

We define j-th factorial moment of X_k by $\nu_j(k) = \mathrm{E}X_k(X_k-1)\cdots(X_k-j+1)$, $(k = 1, 2, \ldots, n,\ j = 1, 2, \ldots)$. For the sake of convenience, we assume that $X_k \equiv 0$ and $\nu_j(k) = 0$ if $k \leqslant 0$ and $\sum_k^n = 0$ if $k > n$. Next, we define remainder terms R_0 and R_1, which appear in the main results, as

$$R_0 = \sum_{k=1}^{n}\big\{\nu_2(k) + \nu_1^2(k) + \mathrm{E}X_{k-1}X_k\big\},$$

$$R_1 = \sum_{k=1}^{n} \Big\{ \nu_1^3(k) + \nu_1(k)\nu_2(k) + \nu_3(k) + [\nu_1(k-2) + \nu_1(k-1) + \nu_1(k)]\,\mathrm{E}X_{k-1}X_k + \widehat{\mathrm{E}}_2^{+}(X_{k-1}, X_k) + \widehat{\mathrm{E}}^{+}(X_{k-2}, X_{k-1}, X_k) \Big\}.$$

We use symbol C to denote (in general different) positive absolute constants.

24.2 The Main Results

All the results are obtained under the following conditions:

$$\nu_1(k) \leqslant 1/100, \quad \nu_2(k) \leqslant \nu_1(k), \quad |X_k| \leqslant C_0, \quad (k = 1, 2, \ldots, n), \tag{24.1}$$

$$\lambda \geqslant 1, \qquad \sum_{k=1}^{n} \nu_2(k) \leqslant \frac{\lambda}{20}, \qquad \sum_{k=2}^{n} |Cov(X_{k-1}, X_k)| \leqslant \frac{\lambda}{20}. \tag{24.2}$$

Assumptions (24.1) and (24.2) are rather restrictive. However, they (a) allow to include independent random variables as partial case of general results and (b) are satisfied for many cases of k-runs and (k_1, k_2)-events. The method of proof does not allow to get small constants. Therefore, we have concentrated our efforts on the order of the accuracy of approximation. Next, we state the main results of this paper.

Theorem 2.1 *Let conditions (24.1) and (24.2) be satisfied. Then, for any* $x \in \mathbb{N}$,

$$\left(1 + \frac{(x-\lambda)^2}{\lambda}\right) |F_n(x) - \Pi(x)| \leqslant C_1 \frac{R_0}{\lambda}, \tag{24.3}$$

$$\left(1 + \frac{(x-\lambda)^2}{\lambda}\right) |F_n(x) - \Pi(x) - \Pi_1(x)| \leqslant C_2 \left(\frac{R_0^2}{\lambda^2} + \frac{R_1}{\lambda\sqrt{\lambda}}\right), \tag{24.4}$$

$$\left(1 + \frac{(x-\lambda)^2}{\lambda}\right) |F_n(x) - \mathrm{G}(x)| \leqslant C_3 \frac{R_1}{\lambda\sqrt{\lambda}}, \tag{24.5}$$

$$\left(1 + \frac{(x-\lambda)^2}{\lambda}\right) |F_n(x) - \mathrm{TP}(x)| \leqslant C_4 \left(\frac{R_1 + |\Gamma_2|}{\lambda\sqrt{\lambda}} + \frac{\tilde{\delta}}{\lambda}\right). \tag{24.6}$$

If in addition $\Gamma_2 > 0$, *then*

$$\left(1 + \frac{(x-\lambda)^2}{\lambda}\right) |F_n(x) - \mathrm{NB}(x)| \leqslant C_5 \left(\frac{R_1}{\lambda\sqrt{\lambda}} + \frac{\Gamma_2^2}{\lambda^2\sqrt{\lambda}}\right). \tag{24.7}$$

If instead $\Gamma_2 < 0$, *then*

$$\left(1 + \frac{(x-\lambda)^2}{\lambda}\right) |F_n(x) - \mathrm{Bi}(x)| \leqslant C_6 \left(\frac{R_1}{\lambda\sqrt{\lambda}} + \frac{\Gamma_2^2}{\lambda^2\sqrt{\lambda}}\right). \tag{24.8}$$

Remark 2.1 Nonuniform normal estimates usually match estimates in Kolmogorov metric. Similarly, the bounds in (24.3)-(24.8) match estimates in total variation:

$$\|F_n - \Pi\| \leqslant C_7 \frac{R_0}{\lambda}, \ \|F_n - \Pi - \Pi_1\| \leqslant C_8 \left(\frac{R_0^2}{\lambda^2} + \frac{R_1}{\lambda\sqrt{\lambda}} \right), \ \|F_n - \mathrm{G}\| \leqslant C_9 \frac{R_1}{\lambda\sqrt{\lambda}},$$

and etc., see Čekanavičius and Vellaisamy (2013).

Estimates for Wasserstein metric easily follow by summing up nonuniform estimates.

Theorem 2.2 *Let conditions (24.1) and (24.2) be satisfied. Then,*

$$\|F_n - \Pi\|_W \leqslant C_{10} \frac{R_0}{\sqrt{\lambda}}, \tag{24.9}$$

$$\|F_n - \Pi - \Pi_1\|_W \leqslant C_{11} \left(\frac{R_0^2}{\lambda\sqrt{\lambda}} + \frac{R_1}{\lambda} \right), \tag{24.10}$$

$$\|F_n - \mathrm{G}\|_W \leqslant C_{12} \frac{R_1}{\lambda}, \tag{24.11}$$

$$\|F_n - \mathrm{TP}\|_W \leqslant C_{13} \left(\frac{R_1 + |\Gamma_2|}{\lambda} + \frac{\tilde{\delta}}{\sqrt{\lambda}} \right). \tag{24.12}$$

When in addition $\Gamma_2 > 0$*, we have*

$$\|F_n - \mathrm{NB}\|_W \leqslant C_{14} \left(\frac{R_1}{\lambda} + \frac{\Gamma_2^2}{\lambda^2} \right), \tag{24.13}$$

and when $\Gamma_2 < 0$*, we have*

$$\|F_n - \mathrm{Bi}\|_W \leqslant C_{15} \left(\frac{R_1}{\lambda} + \frac{\Gamma_2^2}{\lambda^2} \right). \tag{24.14}$$

Observe that the local nonuniform estimates have better order of accuracy.

Theorem 2.3 *Let conditions* (24.1) *and* (24.2) *hold. Then, for any* $x \in \mathbb{N}$,

$$\left(1 + \frac{(x-\lambda)^2}{\lambda}\right) |F_n\{x\} - \Pi\{x\}| \leqslant C_{16} \frac{R_0}{\lambda\sqrt{\lambda}}, \tag{24.15}$$

$$\left(1 + \frac{(x-\lambda)^2}{\lambda}\right) |F_n\{x\} - \Pi\{x\} - \Pi_1\{x\}| \leqslant C_{17} \left(\frac{R_0^2}{\lambda^2\sqrt{\lambda}} + \frac{R_1}{\lambda^2} \right), \tag{24.16}$$

$$\left(1 + \frac{(x-\lambda)^2}{\lambda}\right) |F_n\{x\} - \mathrm{G}\{x\}| \leqslant C_{18} \frac{R_1}{\lambda^2}, \tag{24.17}$$

$$\left(1 + \frac{(x-\lambda)^2}{\lambda}\right) |F_n\{x\} - \mathrm{TP}\{x\}| \leqslant C_{19} \left(\frac{R_1 + |\Gamma_2|}{\lambda^2} + \frac{\tilde{\delta}}{\lambda\sqrt{\lambda}} \right). \tag{24.18}$$

If in addition $\Gamma_2 > 0$, then

$$\Big(1 + \frac{(x-\lambda)^2}{\lambda}\Big)|F_n\{x\} - \mathrm{NB}\{x\}| \leqslant C_{20}\Big(\frac{R_1}{\lambda^2} + \frac{\Gamma_2^2}{\lambda^3}\Big). \tag{24.19}$$

If instead $\Gamma_2 < 0$, then

$$\Big(1 + \frac{(x-\lambda)^2}{\lambda}\Big)|F_n\{x\} - \mathrm{Bi}\{x\}| \leqslant C_{21}\Big(\frac{R_1}{\lambda^2} + \frac{\Gamma_2^2}{\lambda^3}\Big). \tag{24.20}$$

Remark 2.2 (i) Estimates in (24.15)-(24.20) match estimates in local metric, see Čekanavičius and Vellaisamy (2013).
(ii) Consider the case of independent Bernoulli variables with $p_j \leqslant 1/20$ and $\lambda \geqslant 1$. Then, for all integers x, Poisson approximation is of the order

$$\frac{C\sum_{j=1}^n p_j^2}{(1 + (x-\lambda)^2/\lambda)\lambda\sqrt{\lambda}},$$

which is usually much better than

$$\min\Big(\mathrm{x}^{-1}, \lambda^{-1}\Big)\sum_{\mathrm{j}=1}^{\mathrm{n}} p_j^2$$

from Neammanee (2003).

24.3 Some Applications

(i) Asymptotically sharp constant for Poisson approximation to Poisson binomial distribution. Formally, independent random variables make a subset of 1-dependent variables. Therefore, one can rightly expect that results of the previous section apply to independent summands as well. We exemplify this fact by considering one of the best known cases in Poisson approximation theory. Let $W = \xi_1 + \xi_2 + \cdots + \xi_n$, where ξ_i are independent Bernoulli variables with $P(\xi_i = 1) = 1 - P(\xi_i = 0) = p_i$. Let $\lambda = \sum_1^n p_i$, $\lambda_2 = \sum_1^n p_i^2$. As shown in Barbour and Xia (2006) (see Eq. (1.8)),

$$\|\mathcal{L}(W) - \Pi\|_W \leqslant \frac{1.1437\lambda_2}{\sqrt{\lambda}}. \tag{24.21}$$

Though absolute constant in (24.21) is small, we shall show that asymptotically sharp constant is much smaller. Let $\max_i p_i \to 0$ and $\lambda \to \infty$, as $n \to \infty$. Then

$$\lim_{n\to\infty} \frac{\sqrt{\lambda}}{\lambda_2}\|\mathcal{L}(W) - \Pi\|_W = \frac{1}{\sqrt{2\pi}} \leqslant 0.399. \tag{24.22}$$

Indeed, we have

$$\left| \|\mathcal{L}(W) - \Pi\|_W - \frac{\lambda_2}{\sqrt{2\pi\lambda}}| \leqslant \|\mathcal{L}(W) - \Pi - \Pi_1\|_W + |\|\Pi_1\|_W - \frac{\lambda_2}{\sqrt{2\pi\lambda}} \right|.$$

If $\max\limits_i p_i \leqslant 1/20$ and $\lambda \geqslant 1$, then it follows from (24.10) that

$$\|\mathcal{L}(W) - \Pi - \Pi_1\|_W \leqslant \frac{C\lambda_2}{\sqrt{\lambda}} \left(\max_j p_j + \frac{1}{\sqrt{\lambda}} \right).$$

For the estimation of the second difference, we require some notations for measures. Let Z be a measure, corresponding to Fourier-Stieltjes transform $z(t) = (e^{it} - 1)$. Let product and powers of measures be understood in the convolution sense. Then, by the properties of norms and Proposition 4 from Roos (1999) (see also Lemma 6.2 in Čekanavičius and Vellaisamy (2013)), we get

$$\left| \|\Pi_1\|_W - \frac{\lambda_2}{\sqrt{2\pi\lambda}} \right| = \left| \frac{\lambda_2}{2} \|\Pi Z^2\|_W - \frac{\lambda_2}{\sqrt{2\pi\lambda}} \right| = \frac{\lambda_2}{2} \left| \|\Pi Z^2\|_W - \frac{\sqrt{2/\pi}}{\sqrt{\lambda}} \right|$$
$$= \frac{\lambda_2}{2} \left| \|\Pi Z\| - \frac{\sqrt{2/\pi}}{\sqrt{\lambda}} \right| \leqslant \frac{C\lambda_2}{2\lambda} = \frac{\lambda_2}{\sqrt{\lambda}} \frac{C}{2\sqrt{\lambda}}.$$

Thus, for $\max\limits_i p_i \leqslant 1/20$ and $\lambda \geqslant 1$, we obtain asymptotically sharp norm estimate

$$\left| \|\mathcal{L}(W) - \Pi\|_W - \frac{\lambda_2}{\sqrt{2\pi\lambda}} \right| \leqslant \frac{C\lambda_2}{\sqrt{\lambda}} \left(\max_j p_j + \frac{1}{\sqrt{\lambda}} \right),$$

which is even more general than (24.22).

(ii) Negative binomial approximation to 2-runs The k-runs (and especially 2-runs) statistic is one of the best investigated cases of sums of dependent discrete random variables, see Wang and Xia (2008) and the references therein. Let $S_n = X_1 + X_2 + \cdots + X_n$, where $X_i = \eta_i \eta_{i+1}$ and $\eta_j \sim Be(p)$, $(j = 1, 2, \ldots, n+1)$ are independent Bernoulli variables. Then S_n is called 2-runs statistic. It is known that then

$$\lambda = np^2, \quad \Gamma_2 = \frac{np^3(2-3p) - 2p^3(1-p)}{2}.$$

Let $p \leqslant 1/20$ and $np^2 \geqslant 1$. Then, from (24.7) it follows for any $x \in \mathbb{N}$,

$$\left(1 + \frac{(x-\lambda)^2}{\lambda}\right) |F_n(x) - \mathrm{NB}(x)| \leqslant C \frac{p}{\sqrt{n}}.$$

This estimate has the same order as the estimate in total variation, see and Brown and Xia (2001) and Čekanavičius and Vellaisamy (2013).

(iii) Binomial approximation to (k_1, k_2)-events Let η_i be independent Bernoulli $Be(p)$ $(0 < p < 1)$ variables and let $Y_j = (1 - \eta_{j-m+1}) \cdots (1 - \eta_{j-k_2}) \eta_{j-k_2+1} \cdots \eta_{j-1} \eta_j$, $j = m, m+1, \ldots, n$, $k_1 + k_2 = m$. Further, we assume

that $k_1 > 0$ and $k_2 > 0$. Then $N(n; k_1, k_2) = Y_m + Y_{m+1} + \cdots + Y_n$ denote the number of (k_1, k_2)-events and we denote its distribution by H. The Poisson approximation to H has been considered in Vellaisamy (2004). Let $a(p) = (1-p)^{k_1} p^{k_2}$.

Note that $Y_1, Y_2, \ldots$ are m-dependent. However, one can group summands in the following natural way:

$$\begin{aligned} N(n; k_1, k_2) &= (Y_m + Y_{m+1} + \cdots + Y_{2m-1}) + (Y_{2m} + Y_{2m+1} + \cdots + Y_{3m-1}) + \ldots \\ &= X_1 + X_2 + \ldots. \end{aligned}$$

Each X_j, with probable exception of the last one, contains m summands. It is not difficult to check that $X_1, X_2, \ldots$ are 1-dependent Bernoulli variables. Then all parameters can be written explicitly. Set $N = \lfloor \tilde{N} \rfloor$ be the integer part of $\tilde{N}$ defined by

$$\tilde{N} = \frac{(n-m+1)^2}{(n-m+1)(2m-1) - m(m-1)}, \quad \bar{p} = \frac{(n-m+1)a(p)}{N}.$$

It is known (see Čekanavičius and Vellaisamy (2013)) that

$$\begin{aligned} \lambda =& (n-m+1)a(p), \quad \Gamma_2 = -\frac{a^2(p)}{2}[(n-m+1)(2\,m-1) - m(m-1)], \\ & R_1 \leqslant C(n-m+1)m^2 a^3(p). \end{aligned}$$

Let now $\lambda \geq 1$ and $ma(p) \leqslant 0.01$. Then it follows from (24.8) that, for any $x \in \mathbb{N}$,

$$\left(1 + \frac{(x-\lambda)^2}{\lambda}\right) |\mathrm{H}(x) - \mathrm{Bi}(x)| \leqslant C \frac{a^{3/2}(p)m^2}{\sqrt{n-m+1}}.$$

24.4 Auxiliary results

Let θ denote a real or complex quantity satisfying $|\theta| \leqslant 1$. Moreover, let $Z_j = \exp\{\mathrm{i}tX_j\} - 1$, $\Psi_{j,k} = \widehat{\mathrm{E}}(Z_j, Z_{j+1}, \ldots, Z_k)$. As before, we assume that $\nu_j(k) = 0$ and $X_k = 0$ for $k \leqslant 0$ and $z(t) = \mathrm{e}^{\mathrm{i}t} - 1$. Also, we omit the argument t, wherever possible and, for example, write z instead of $z(t)$. Hereafter, the primes denote the derivatives with respect to t.

Lemma 4.1 *Let X be concentrated on nonnegative integers and $\nu_3 < \infty$. Then, for all $t \in \mathbb{R}$,*

$$\begin{aligned} \mathrm{E}\exp\{\mathrm{i}tX\} &= 1 + \nu_1 z + \nu_2 \frac{z^2}{2} + \theta \frac{\nu_3 |z|^3}{6}, \\ \mathrm{E}(\exp\{\mathrm{i}tX\})' &= \nu_1 z' + \nu_2 \frac{(z^2)'}{2} + \theta \frac{\nu_3 |z|^2}{2}, \\ \mathrm{E}(\exp\{\mathrm{i}tX\})'' &= \nu_1 z'' + \nu_2 \frac{(z^2)''}{2} + 2\theta \nu_3 |z|. \end{aligned}$$

Proof First equality is well known expansion of characteristic function in factorial moments. The other two equalities also easily follow from expansions in powers of z. For example,

$$\begin{aligned}(\mathrm{e}^{\mathrm{i}tX})'' &= \mathrm{i}^2X^2\mathrm{e}^{\mathrm{i}tX} = \mathrm{i}^2X(X-1)(\mathrm{e}^{\mathrm{i}t})^2\mathrm{e}^{\mathrm{i}t(X-2)} + \mathrm{i}^2\mathrm{e}^{\mathrm{i}t}X\mathrm{e}^{\mathrm{i}t(X-1)}\\ &= \mathrm{i}^2(\mathrm{e}^{\mathrm{i}t})^2X(X-1)[1+\theta(X-2)|z|]\\ &+ \mathrm{i}^2\mathrm{e}^{\mathrm{i}t}X[1+(X-1)z+\theta(X-1)(X-2)|z|^2/2]\\ &= Xz'' + \frac{X(X-1)}{2}(z^2)'' + 2\theta X(X-1)(X-2)|z|. \quad \square \end{aligned} \tag{24.23}$$

Lemma 4.2 (*Heinrich* 1982) *Let $Y_1, Y_2, \ldots, Y_k$ be 1-dependent complex-valued random variables with $\mathrm{E}|Y_m|^2 < \infty$, $1 \leq m \leq k$. Then*

$$|\widehat{\mathrm{E}}(Y_1, Y_2, \cdots, Y_k)| \leqslant 2^{k-1}\prod_{m=1}^{k}(\mathrm{E}|Y_m|^2)^{1/2}.$$

Lemma 4.3 *Let conditions* (24.1) *be satisfied and $j < k-1$. Then, for all real t,*

$$|\Psi_{j,k}| \leqslant 4^{k-j}|z|\prod_{l=j}^{k}\sqrt{\nu_1(l)}, \tag{24.24}$$

$$|\Psi'_{j,k}| \leqslant 4^{k-j}|z|(k-j+1)\prod_{l=j}^{k}\sqrt{\nu_1(l)}, \tag{24.25}$$

$$|\Psi''_{j,k}| \leqslant \sqrt{2}C_0 4^{k-j}|z|(k-j+1)(k-j)\prod_{l=j}^{k}\sqrt{\nu_1(l)}. \tag{24.26}$$

Proof First two estimates follow from more general estimates in (47) and Lemma 7.5 in Čekanavičius and Vellaisamy (2013). Note also the following inequalities:

$$|z| \leqslant 2, \quad |Z_k| \leqslant 2, \quad |Z_k| \leqslant X_k|z|, \quad \mathrm{E}X_i^2 = \nu_2(i) + \nu_1(i) \leqslant 2\nu_1(i). \tag{24.27}$$

Therefore, by Lemma 4.2 and for $m \leq k$,

$$|\widehat{\mathrm{E}}(Z_j, \ldots, Z'_m, \ldots, Z'_i, \ldots Z_k)| \leqslant 2^{k-j}\sqrt{\mathrm{E}|Z'_m|^2\mathrm{E}|Z'_i|^2}\prod_{l=j,l\neq m,i}^{k}\sqrt{2|z|\nu_1(l)}$$

$$\leqslant 2^{k-j}\sqrt{2\nu_1(m)2\nu_1(i)}2^{(k-j-1)/2}|z|^{(k-j-1)/2}\prod_{l=j,l\neq m,i}^{k}\sqrt{\nu_1(l)} \leqslant 4^{k-j}2^{-1}|z|\prod_{l=j}^{k}\sqrt{\nu_1(l)}.$$

Similarly,

$$|\widehat{\mathrm{E}}(Z_j, \ldots, Z''_i, \ldots, Z_k)| \leqslant 2^{k-j}\sqrt{\mathrm{E}|Z''_i|^2}\prod_{l=j,l\neq i}^{k}\sqrt{2|z|\nu_1(l)}$$

$$\leqslant 2^{k-j}\sqrt{\mathrm{E}X_i^4}\,2^{(k-j)/2}|z|^{(k-j)/2}\prod_{l=j,l\neq i}^{k}\sqrt{\nu_1(l)}$$

$$\leqslant 4^{k-j}2^{-1}|z|C_0\sqrt{\mathrm{E}X_i^2}\prod_{l=j,l\neq i}^{k}\sqrt{\nu_1(l)} \leqslant 4^{k-j}2^{-1/2}C_0\prod_{l=j}^{k}\sqrt{\nu_1(l)}.$$

Thus,

$$|\Psi''_{j,k}| \leqslant \sum_{i=j}^{k}|\widehat{\mathrm{E}}(Z_j,\dots,Z''_i,\dots,Z_k)| + \sum_{i=j}^{k}$$

$$\sum_{m=j,m\neq i}^{k}|\widehat{\mathrm{E}}(Z_j,\dots,Z'_m,\dots,Z'_i,\dots,Z_k)|$$

$$\leqslant (k-j+1)4^{k-j}C_0 2^{-1/2}|z|\prod_{l=j}^{k}\sqrt{\nu_1(l)} + (k-j+1)$$

$$(k-j)4^{k-j}2^{-1}|z|\prod_{l=j}^{k}\sqrt{\nu_1(l)}$$

$$\leqslant \sqrt{2}C_0 4^{k-j}(k-j+1)(k-j)|z|\prod_{l=j}^{k}\sqrt{\nu_1(l)}. \qquad \square$$

In the following Lemmas 4.4–4.5, we present some facts about characteristic function $\widehat{F}_n(t)$ from Čekanavičius and Vellaisamy (2013). Here again we assume (24.1), though many relations hold also under weaker assumptions, see Čekanavičius and Vellaisamy (2013). We begin from Heinrich's representation of $\widehat{F}_n$ as product of functions.

Lemma 4.4 *Let* (24.1) *hold. Then* $\widehat{F}_n(t) = \varphi_1(t)\varphi_2(t)\dots\varphi_n(t)$, where $\varphi_1(t) = \mathrm{E}e^{\mathrm{i}tX_1}$ *and, for* $k = 2,\dots,n$,

$$\varphi_k = 1 + \mathrm{E}Z_k + \sum_{j=1}^{k-1}\frac{\Psi_{j,k}}{\varphi_j\varphi_{j+1}\dots\varphi_{k-1}}. \tag{24.28}$$

Let

$$g_j(t) = \exp\left\{\nu_1(j)z + \Big(\frac{\nu_2(j)-\nu_1^2(j)}{2} + \widehat{\mathrm{E}}(X_{j-1},X_j)\Big)z^2\right\},$$

$$\lambda_k = 1.6\nu_1(k) - 0.3\nu_1(k-1) - 2\nu_2(k) - 0.1\mathrm{E}X_{k-2}X_{k-1} - 15.58\mathrm{E}X_{k-1}X_k,$$

$$\gamma_2(k) = \frac{\nu_2(k)}{2} + \widehat{\mathrm{E}}(X_{k-1},X_k),$$

$$r_1(k) = \nu_3(k) + \sum_{l=0}^{5}\nu_1^3(k-l) + [\nu_1(k-1)+\nu_1(k-2)]\mathrm{E}X_{k-1}X_k + \widehat{\mathrm{E}}_2^+(X_{k-1},X_k)$$

$$+\widehat{\mathrm{E}}^{+}(X_{k-2}, X_{k-1}, X_k),$$

Lemma 4.5 *Let the conditions in* (24.1) *hold. Then*

$$\frac{1}{|\varphi_k|} \leqslant \frac{10}{9}, \tag{24.29}$$

$$|\varphi_k| \leqslant \exp\{-\lambda_k \sin^2(t/2)\}, \quad |g_k| \leqslant \exp\{-\lambda_k \sin^2(t/2)\}, \tag{24.30}$$

$$\frac{1}{\varphi_{k-1}} = 1 + C\theta|z|\{\nu_1(k-2) + \nu_1(k-1)\}, \tag{24.31}$$

$$\varphi_k' = 33\theta[\nu_1(k) + \nu_1(k-1)], \tag{24.32}$$

$$\sum_{k=1}^{n} |\varphi_k - g_k| \leqslant CR_1|z|^3, \quad \sum_{k=1}^{n} |\varphi_k' - g_k'| \leqslant CR_1|z|^2. \tag{24.33}$$

Similar estimates hold for the second derivative, as seen in the next lemma.

Lemma 4.6 Let (24.1) hold. Then, for $k = 1, 2, \ldots, n$,

$$\varphi_k'' = \theta C_{22}[\nu_1(k) + \nu_1(k-1)], \tag{24.34}$$

$$\varphi_k'' = \nu_1(k)z'' + \gamma_2(k)(z^2)'' + \theta C|z|r_1(k). \tag{24.35}$$

Proof From Lemma 4.4, it follows that

$$\begin{aligned}\varphi_k'' = (\mathrm{E}Z_k)'' &+ \sum_{j=1}^{k-1} \frac{\Psi_{j,k}''}{\varphi_j \cdots \varphi_{k-1}} - 2\sum_{j=1}^{k-1} \frac{\Psi_{j,k}'}{\varphi_j \cdots \varphi_{k-1}} \sum_{i=j}^{k-1} \frac{\varphi_i'}{\varphi_i} \\ &+ \sum_{j=1}^{k-1} \frac{\Psi_{j,k}}{\varphi_j \cdots \varphi_{k-1}} \Big(\sum_{i=j}^{k-1} \frac{\varphi_i'}{\varphi_i}\Big)^2 + \sum_{j=1}^{k-1} \frac{\Psi_{j,k}}{\varphi_j \cdots \varphi_{k-1}} \sum_{i=j}^{k-1} \Big(\frac{\varphi_i'}{\varphi_i}\Big)^2 \\ &- \sum_{j=1}^{k-1} \frac{\Psi_{j,k}}{\varphi_j \cdots \varphi_{k-1}} \sum_{i=j}^{k-1} \frac{\varphi_i''}{\varphi_i}.\end{aligned} \tag{24.36}$$

We prove (24.34) by mathematical induction. Note that by Lemma 4.1, $(\mathrm{E}Z_k)'' = C\theta\nu_1(k)$. Moreover, for $j \leqslant k-2$,

$$\prod_{l=j}^{k} \sqrt{\nu_1(l)} = \sqrt{\nu_1(k)\nu_1(k-1)} \prod_{l=j}^{k-2} \sqrt{\nu_1(l)} \leqslant \frac{\nu_1(k) + \nu_1(k-1)}{2} 10^{-(k-j-1)}. \tag{24.37}$$

Applying (24.37) to (24.24), for all $j \leqslant k-2$, we prove

$$|\Psi_{j,k}| \leqslant 10\Big(\frac{4}{10}\Big)^{k-j}[\nu_1(k)+\nu_1(k-1)]. \tag{24.38}$$

Taking into account (24.27) and (24.1), it is easy to check that

$$\begin{aligned}|\widehat{\mathrm{E}}(Z_{k-1},Z_k)| &\leqslant \mathrm{E}|Z_{k-1}Z_k| + \mathrm{E}|Z_{k-1}|\mathrm{E}|Z_k| = \mathrm{E}|Z_{k-1}Z_k|/2\\ &\quad + \mathrm{E}|Z_{k-1}Z_k|/2 + \mathrm{E}|Z_{k-1}|\mathrm{E}|Z_k|/2 + \mathrm{E}|Z_{k-1}|\mathrm{E}|Z_k|/2\\ &\leqslant \mathrm{E}|Z_{k-1}| + \mathrm{E}|Z_k| + 0.01\mathrm{E}|Z_{k-1}| + 0.01\mathrm{E}|Z_k|\\ &\leqslant 2.02[\nu_1(k-1)+\nu_1(k)].\end{aligned}$$

Therefore, we see that (24.38) holds also for $j=k-1$. From inductional assumption, (24.29), (24.32) and (24.1), it follows

$$\frac{|\varphi_i''|}{|\varphi_i|} \leqslant C_{22}[\nu_1(i-1)+\nu_1(i)]\frac{10}{9} \leqslant \frac{2C_{22}}{90}.$$

Using (24.29) and the previous estimate, we obtain

$$\begin{aligned}&\Big|\sum_{j=1}^{k-1}\frac{\Psi_{j,k}}{\varphi_j\cdots\varphi_{k-1}}\sum_{i=j}^{k-1}\frac{\varphi_i''}{\varphi_i}\Big| \leqslant \sum_{j=1}^{k-1}\Big(\frac{10}{9}\Big)^{k-j}|\Psi_{j,k}|\sum_{i=j}^{k-1}\frac{|\varphi_i''|}{|\varphi_i|}\\ &\leqslant \sum_{j=1}^{k-1}10\Big(\frac{4}{9}\Big)^{k-j}[\nu_1(k)+\nu_1(k-1)](k-j)\frac{2C_{22}}{90} \leqslant \frac{8C_{22}}{25}[\nu_1(k)+\nu_1(k-1)].\end{aligned}$$

Estimating all other sums (without using induction arguments) in a similar manner, we finally arrive at the estimate

$$|\varphi_k''| \leqslant C_{23}[\nu_1(k-1)+\nu_1(k)] + \frac{8C_{22}}{25}[\nu_1(k)+\nu_1(k-1)].$$

It remains to choose $C_{22} = 25C_{23}/17$ to complete the proof of (24.34).
Since the proof of (24.35) is quite similar, we give only a general outline of it. First, we assume that $k \geqslant 6$. Then in (24.36) split all sums into $\sum_{j=1}^{k-5} + \sum_{j=k-4}^{k-1}$. Next, note that

$$\prod_{l=j}^{k}\sqrt{\nu_1(l)} \leqslant \prod_{l=k-5}^{k}\sqrt{\nu_1(l)}\prod_{l=j}^{k-6}\Big(\frac{1}{10}\Big) \leqslant \sum_{l=k-5}^{k}\nu_1^3(l)10^{-(k-j-5)} \leqslant r_1(k)10^{-(k-j-5)}.$$

Therefore, applying (24.24)–(24.26) and using (24.29), (24.32) and (24.34), we easily prove that all sums $\sum_{j=1}^{k-5}$ are by absolute value less than $C|z|r_1(k)$. The cases $j=k-4, k-3, k-2$ all contain at least three Z_i and can be estimated directly by $C|z|r_1(k)$. For example,

$$|\widehat{\mathrm{E}}(Z_{k-3},Z_{k-2},Z_{k-1},Z_k)| \leqslant 4\widehat{\mathrm{E}}^{+}(|Z_{k-2}|,|Z_{k-1}|,|Z_k|)$$

$$\leqslant C|z|^3\widehat{\mathrm{E}}^+(X_{k-2}, X_{k-1}, X_k) \leqslant C|z|r_1(k).$$

Easily verifiable estimates $|(\widehat{\mathrm{E}}(Z_{k-1}, Z_k))'||\varphi'_{k-1}| \leqslant C|z|r_1(k)$, $|\widehat{\mathrm{E}}(Z_{k-1}, Z_k)||\varphi'_{k-1}|^2 \leqslant C|z|r_1(k)$, and $|\widehat{\mathrm{E}}(Z_{k-1}, Z_k)||\varphi''_{k-1}| \leqslant C|z|r_1(k)$ and Lemma 4.1 allow us to obtain the expression

$$\varphi''_k = \nu_1(k)z'' + \frac{\nu_2(k)}{2}(z^2)'' + \frac{(\widehat{\mathrm{E}}(Z_{k-1}, Z_k))''}{\varphi_{k-1}} + C|(z)|r_1(k). \tag{24.39}$$

It follows, from (24.31), that

$$\frac{(\widehat{\mathrm{E}}(Z_{k-1}, Z_k))''}{\varphi_{k-1}} = (\widehat{\mathrm{E}}(Z_{k-1}, Z_k))'' + C|z|r_1(k). \tag{24.40}$$

Now $(\widehat{\mathrm{E}}(Z_{k-1}, Z_k))'' = \widehat{\mathrm{E}}(Z''_{k-1}, Z_k) + 2\widehat{\mathrm{E}}(Z'_{k-1}, Z'_k) + \widehat{\mathrm{E}}(Z_{k-1}, Z''_k)$.
Due to

$$\begin{aligned} Z'_{k-1} &= \mathrm{i}X_{k-1}\mathrm{e}^{\mathrm{i}tX_{k-1}} = z'X_{k-1}(1 + \theta(X_{k-1} - 1)|z|/2) \\ &= z'X_{k-1} + \theta X_{k-1}(X_{k-1} - 1), \end{aligned}$$

we obtain

$$\begin{aligned} 2\widehat{\mathrm{E}}(Z'_{k-1}, Z'_k) &= 2z'\widehat{\mathrm{E}}(X_{k-1}, Z'_k) + \theta\widehat{\mathrm{E}}_2^+(X_{k-1}, X_k)|z| \\ &= 2(z')^2\widehat{\mathrm{E}}(X_{k-1}, X_k) + \theta C\widehat{\mathrm{E}}_2^+(X_{k-1}, X_k)|z|. \end{aligned}$$

Similarly, $Z_k = X_k z + \theta X_k(X_k - 1)|z|^2/2$ and

$$\begin{aligned} \widehat{\mathrm{E}}(Z''_{k-1}, Z_k) + \widehat{\mathrm{E}}(Z_{k-1}, Z''_k) &= z(\widehat{\mathrm{E}}(Z''_{k-1}, X_k) + \widehat{\mathrm{E}}(X_{k-1}, Z''_k)) \\ &\quad + \theta C|z|\widehat{\mathrm{E}}_2^+(X_{k-1}, X_k). \end{aligned}$$

Applying (24.23), we prove $\widehat{\mathrm{E}}(Z''_{k-1}, X_k) = z''\widehat{\mathrm{E}}(X_{k-1}, X_k) + \theta C\widehat{\mathrm{E}}_2^+(X_{k-1}, X_k)$. Consequently,

$$(\widehat{\mathrm{E}}(Z_{k-1}, Z_k))'' = (z^2)''\widehat{\mathrm{E}}(X_{k-1}, X_k) + \theta C|z|\widehat{\mathrm{E}}_2^+(X_{k-1}, X_k).$$

Combining the last estimate with (24.40) and (24.39), we complete the proof of (24.35). The case $k < 6$ is proved exactly by the same arguments. □
Let $\tilde{\varphi}_k = \varphi_k\exp\{-\mathrm{i}t\nu_1(k)\}$, $\tilde{g}_k = g_k\exp\{-\mathrm{i}t\nu_1(k)\}$, $\psi = \exp\{-0.1\lambda\sin^2(t/2)\}$.

Lemma 4.7 *Let* (24.1) *hold. Then*

$$\sum_{l=1}^{n}|\tilde{\varphi}'_l| \leqslant C\lambda|z|, \quad \sum_{l=1}^{n}|\tilde{g}'_l| \leqslant C\lambda|z|, \quad \sum_{l=1}^{n}|\tilde{\varphi}''_l| \leqslant C\lambda,$$

$$\sum_{l=1}^{n}|\tilde{g}''_l| \leqslant C\lambda, \quad |\prod_{l=1}^{n}\tilde{\varphi}_l - \prod_{l=1}^{n}\tilde{g}_l| \leqslant CR_1|z|^3\psi,$$

$$\Big|\Big(\prod_{l=1}^{n}\tilde{\varphi}_l-\prod_{l=1}^{n}\tilde{g}_l\Big)'\Big|\leqslant CR_1|z|^2\psi,\quad \Big|\Big(\prod_{l=1}^{n}\tilde{\varphi}_l-\prod_{l=1}^{n}\tilde{g}_l\Big)''\Big|\leqslant CR_1|z|\psi.$$

Proof The first four estimates follow from Lemmas 4.5 and 4.6 and trivial estimate $\mathrm{E}X_{k-1}X_k\leqslant C_0\nu_1(k)$. Also, using (24.1) and (24.30), we get

$$\prod_{l=1,l\neq k}^{n}\exp\{-\lambda_l\sin^2(t/2)\}\leqslant C\prod_{l=1}^{n}\exp\{-\lambda_l\sin^2(t/2)\}\leqslant C\psi^2.$$

Therefore, by (24.30) and (24.33),

$$\begin{aligned}\Big|\prod_{l=1}^{n}\tilde{\varphi}_l-\prod_{l=1}^{n}\tilde{g}_l\Big|=\Big|\prod_{l=1}^{n}\varphi_l-\prod_{l=1}^{n}g_l\Big|&\leqslant\sum_{j=1}^{n}|\varphi_j-g_j|\prod_{l=1}^{j-1}|g_l|\prod_{l=j+1}^{n}|\varphi_l|\\&\leqslant C\psi^2\sum_{j=1}^{n}|\varphi_j-g_j|\leqslant CR_1|z|^3\psi^2.\end{aligned}$$

From (24.1) and trivial estimate $z\mathrm{e}^{-x}\leqslant 1$, for $x>0$, we get

$$|\Gamma_2|\leqslant 0.08\lambda,\quad \lambda|z|^2\psi\leqslant C.$$

Therefore,

$$\begin{aligned}\Big|\Big(\prod_{l=1}^{n}\tilde{\varphi}_l-\prod_{l=1}^{n}\tilde{g}_l\Big)'\Big|&\leqslant\sum_{l=1}^{n}|\tilde{\varphi}_l'-\tilde{g}_l'|\prod_{k\neq l}|\tilde{\varphi}_k|+\sum_{l=1}^{n}|\tilde{g}_l'|\Big|\prod_{k\neq l}\tilde{\varphi}_k-\prod_{k\neq l}\tilde{g}_k\Big|\\&\leqslant C\psi^2[R_1|z|^2+\lambda|z|R_1|z|^3]\leqslant C\psi R_1|z|^2.\end{aligned}$$

The proof of last estimate is very similar and therefore omitted. □

24.5 Proof of Theorems

Proof of Theorem 2.1 Hereafter, $x\in\mathbb{N}$, the set of nonnegative integers. The beginning of the proof is almost identical to the proof of Tsaregradsky's inequality. Let M be concentrated on integers. Then summing up the formula of inversion

$$M\{k\}=\frac{1}{2\pi}\int_{-\pi}^{\pi}\widehat{M}(t)\mathrm{e}^{-\mathrm{i}tk}\mathrm{d}t,\tag{24.41}$$

we get

$$\sum_{k=m}^{x} M\{k\} = \frac{1}{2\pi}\int_{-\pi}^{\pi} \widehat{M}(t)\frac{\mathrm{e}^{-\mathrm{i}t(m-1)} - \mathrm{e}^{-\mathrm{i}tx}}{z}\,\mathrm{d}t.$$

If $|\widehat{M}(t)/z|$ is bounded, then as $m \to -\infty$ and by Riemann-Lebesgue theorem, we get

$$M(x) = -\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\widehat{M}(t)\mathrm{e}^{-\mathrm{i}tx}}{z}\,\mathrm{d}t = -\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\widehat{M}(t)\mathrm{e}^{-\mathrm{i}t/2}\mathrm{e}^{-\mathrm{i}tx}}{2\mathrm{i}\sin(t/2)}\,\mathrm{d}t. \tag{24.42}$$

The Tsaregradsky's inequality

$$|M(x)| \leqslant \frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{|\widehat{M}(t)|}{|z|}\,\mathrm{d}t \tag{24.43}$$

now follows easily. Let next $M = F_n - \mathrm{G}$. Then expressing $\widehat{M}(t)$ in powers of z, we get $\widehat{M}(t) = \sum_{k=2}^{\infty} a_k z^k$, for some coefficients a_k which depend on factorial moments of S_n. Therefore, $\widehat{M}(\pi)/z(\pi) = \widehat{M}(-\pi)/z(-\pi)$. Consequently, integrating (24.42) by parts, we obtain, for $x \neq \lambda$,

$$M(x) = -\frac{1}{2\pi}\int_{-\pi}^{\pi} \frac{\widehat{M}(t)\mathrm{e}^{-\mathrm{i}t(\lambda+1/2)}}{2\mathrm{i}\sin(t/2)}\,\mathrm{e}^{-\mathrm{i}t(x-\lambda)}\,\mathrm{d}t = \frac{1}{2\pi(x-\lambda)^2}\int_{-\pi}^{\pi} u''(t)\mathrm{e}^{-\mathrm{i}t(x-\lambda)}\mathrm{d}t,$$

where

$$u(t) = \mathrm{e}^{-(\lambda+1/2)\mathrm{i}t}\frac{\widehat{M}(t)}{2\mathrm{i}\sin(t/2)} = \frac{\prod_{j=1}^{n}\tilde{\varphi}_j - \prod_{j=1}^{n}\tilde{g}_j}{z}.$$

Thus, for all $x \in \mathbb{N}$,

$$(x-\lambda)^2 M(x) \leqslant \frac{1}{2\pi}\int_{-\pi}^{\pi} |u''(t)|\mathrm{d}t. \tag{24.44}$$

Using Lemma 24.7, Eqs. (24.43), (24.44) and the trivial estimate

$$\int_{-\pi}^{\pi} |z|^k \psi(t)\mathrm{d}t \leqslant \frac{C(k)}{\lambda^{(k+1)/2}} \tag{24.45}$$

the proof of (24.5) follows.

All other approximations are compared to compound Poisson measure G and then the triangle inequality is applied. We begin from the negative binomial distribution.

Due to the assumptions,

$$\Gamma_2 \leqslant \frac{3}{40}\lambda, \quad \frac{1-\overline{q}}{\overline{q}} = \frac{2\Gamma_2}{\lambda} \leqslant 0.15,$$

see Čekanavičius and Vellaisamy (2013). Therefore, $\widehat{NB}(t)\exp\{-\lambda \mathrm{i}t\} = \exp\{A\}$,where

$$A = \lambda z - \mathrm{i}t + \Gamma_2 z^2 + \sum_{j=3}^{\infty} \frac{r}{j}\Big(\frac{1-\overline{q}}{\overline{q}}\Big)^j z^j = \lambda(z-\mathrm{i}t) + \Gamma_2 z^2 + \theta C \Gamma_2^2 \lambda^{-1}|z|^3.$$

Moreover,

$$|A'| \leqslant C\lambda|z|, \quad |A''| \leqslant C\lambda, \quad |\mathrm{e}^A| \leqslant \psi^2.$$

Let $B = \lambda(z-\mathrm{i}t)+\Gamma_2 z^2$ so that $\widehat{\mathrm{G}}(t)\exp\{-\lambda \mathrm{i}t\} = \exp\{B\}$ and $u_1(t) = (\mathrm{e}^A - \mathrm{e}^B)/z$. Then

$$|u_1| \leqslant \frac{|\mathrm{e}^A - \mathrm{e}^B|}{|z|} \leqslant \psi^2 \frac{|A-B|}{|z|} \leqslant C\psi^2 \frac{\Gamma_2^2|z|^2}{\lambda}, \quad \int_{-\pi}^{\pi} |u_1|\mathrm{d}t \leqslant C\frac{\Gamma_2^2}{\lambda^2\sqrt{\lambda}}. \tag{24.46}$$

Also,

$$\begin{aligned}|(\mathrm{e}^A - \mathrm{e}^B)''| &\leqslant |A''||\mathrm{e}^A - \mathrm{e}^B| + |(A')^2||\mathrm{e}^A - \mathrm{e}^B| + |A'' - B''||\mathrm{e}^B| \\ &\quad + |(A')^2 - (B')^2||\mathrm{e}^B| \\ &\leqslant C\psi^2\Big\{\lambda\frac{\Gamma_2^2}{\lambda}|z|^3 + \lambda^2|z|^2\frac{\Gamma_2^2}{\lambda}|z|^3 + \frac{\Gamma_2^2}{\lambda}|z| + \lambda|z|\frac{\Gamma_2^2}{\lambda}|z|^2\Big\} \leqslant C\psi|z|\frac{\Gamma_2^2}{\lambda}.\end{aligned}$$

Similarly,

$$|(\mathrm{e}^A - \mathrm{e}^B)'| \leqslant |A'||\mathrm{e}^A - \mathrm{e}^B| + |\mathrm{e}^B||A' - B'| \leqslant C\psi|z|^2\frac{\Gamma_2^2}{\lambda}$$

and we obtain finally

$$|u_1''| \leqslant C\psi\frac{\Gamma_2^2}{\lambda}, \quad \int_{-\pi}^{\pi} |u_1''|\mathrm{d}t \leqslant C\frac{\Gamma_2^2}{\lambda\sqrt{\lambda}}. \tag{24.47}$$

Estimates in (24.46) and (24.47) allow us to write

$$\Big(1 + \frac{(x-\lambda)^2}{\lambda}\Big)|\mathrm{G}(x) - \mathrm{NB}(x)| \leqslant C\frac{\Gamma_2^2}{\lambda^2\sqrt{\lambda}},$$

which combined with (24.5) proves (24.7).

For the proof of translated Poisson approximation, let B be defined as in above,

$$T = \lambda(z - \mathrm{i}t) + (2\Gamma_2 + \tilde{\delta})(z - \mathrm{i}t), \quad D = \lambda(z - \mathrm{i}t) + (\Gamma_2 + \tilde{\delta}/2)z^2,$$

and

$$u_2 = (\mathrm{e}^D - \mathrm{e}^T)/z, \quad u_3 = (\mathrm{e}^B - \mathrm{e}^D)/z.$$

Note that, for $|t| \leqslant \pi$, we have $|t|/\pi \leqslant |\sin(t/2)| \leqslant |t|/2$. Therefore, arguing similarly as in above, we obtain

$$\int_{-\pi}^{\pi} |u_2|\mathrm{d}t \leqslant \frac{C(|\Gamma_2| + \tilde{\delta})}{\lambda\sqrt{\lambda}}, \quad \int_{-\pi}^{\pi} |u_2''|\mathrm{d}t \leqslant \frac{C(|\Gamma_2| + \tilde{\delta})}{\sqrt{\lambda}}. \tag{24.48}$$

Observe next that

$$u_3 = \frac{\mathrm{e}^B}{z}(\mathrm{e}^{\tilde{\delta}z^2/z} - 1) = \frac{\mathrm{e}^B}{z}\int_0^1 (\tilde{\delta}z^2/2)\mathrm{e}^{\tau\tilde{\delta}z^2/2}\mathrm{d}\tau = \int_0^1 \frac{\tilde{\delta}z}{2}\mathrm{e}^{B+\tau\tilde{\delta}z^2/2}\mathrm{d}\tau.$$

Consequently,

$$\int_{-\pi}^{\pi} |u_3|\mathrm{d}t \leqslant C\int_{-\pi}^{\pi} \psi^2\tilde{\delta}|z|\mathrm{d}t \leqslant \frac{C\tilde{\delta}}{\lambda}. \tag{24.49}$$

Similarly,

$$u_3'' = \frac{\tilde{\delta}}{2}\int_0^1 \mathrm{e}^{B+\tau\tilde{\delta}t}[z'' + 2z'(B' + \tau\tilde{\delta}zz') + z(B'' + \tau\tilde{\delta}(zz')') + z(B' + \tau\tilde{\delta}zz')^2]\mathrm{d}\tau$$

and using $\tilde{\delta} \leqslant 1 \leqslant \lambda$, we get

$$|u_3''| \leqslant C\psi^2\tilde{\delta}(1 + \lambda|z| + \tilde{\delta}|z| + |z|(\lambda|z| + \tilde{\delta}|z|)^2) \leqslant C\tilde{\delta}\psi\sqrt{\lambda}.$$

Consequently,

$$\int_{-\pi}^{\pi} |u_3''|\mathrm{d}t \leqslant C\tilde{\delta}.$$

Combining the last estimate, the inequalities in (24.48), (24.49) and the estimate for $\widehat{G} = \mathrm{e}^B$, the result in (24.6) is proved.

For binomial approximation, note first that

$$\mathrm{e}^{-\lambda\mathrm{i}t}\widehat{\mathrm{Bi}} = \mathrm{e}^E,\ \varepsilon = \tilde{\mathrm{N}} - \mathrm{N}. \quad E = \lambda(z - \mathrm{i}t) + \Gamma_2 z^2 + z^2\theta\frac{50\Gamma_2^2}{21\lambda^2}\varepsilon + \theta\frac{5\,N\overline{p}^3|z|^3}{9},$$

$$\overline{p} \leqslant \frac{50|\Gamma_2|}{21\lambda} < \frac{1}{5}, \quad |\Gamma_2| \leqslant 0.08\lambda, \quad |N_{\tilde{p}}^3| \leqslant C\frac{\Gamma_2^2}{\lambda},$$

see Čekanavičius and Vellaisamy (2013). Let

$$L = \lambda(z - \mathrm{i}t) + \Gamma_2 z^2 + z^2\theta\frac{50\Gamma_2^2}{21\lambda^2}\varepsilon, \quad u_4 = (\mathrm{e}^L - \mathrm{e}^E)/z, \quad u_5 = (\mathrm{e}^B - \mathrm{e}^L)/z.$$

Next,

$$u_5 = \int_0^1 \mathrm{e}^B z \exp\left\{\tau z^2\theta\frac{50\Gamma_2^2}{21\lambda^2}\varepsilon\right\}\theta\frac{50\Gamma_2^2}{21\lambda^2}\varepsilon \mathrm{d}\tau.$$

Now the proof is practically identical to that of (24.6) and is, therefore, omitted.

The proofs of (24.3) and (24.4) are also very similar and use the facts

$$\frac{\mathrm{e}^B - \mathrm{e}^{-\lambda \mathrm{i}t}(\widehat{\Pi} + \widehat{\Pi}_1)}{z} = \int_0^1 (1-\tau)\Gamma_2^2 z^3 \exp\{\lambda(z - \mathrm{i}t) + \tau\Gamma_2 z^2\}\mathrm{d}\tau,$$

$$\frac{\mathrm{e}^B - \mathrm{e}^{-\lambda \mathrm{i}t}\widehat{\Pi}}{z} = \int_0^1 \Gamma_2 z \exp\{\lambda(z - \mathrm{i}t) + \tau\Gamma_2 z^2\}\mathrm{d}\tau.$$

□

Proof of Theorem 2.3 Let M be a measure concentrated on integers and $\widehat{M}(t) = \sum_{k=1}^{\infty} M\{k\}\mathrm{e}^{\mathrm{i}tk}$. Then from formula (24.41) of inversion, we get

$$|M\{x\}|\frac{1}{2\pi} \leqslant \int_{-\pi}^{\pi} |\widehat{M}(t)|\mathrm{d}t.$$

Moreover, integrating (24.41) by parts, we obtain

$$(x - \lambda)^2|M\{x\}| \leqslant \frac{1}{2\pi}\int_{-\pi}^{\pi} |(\widehat{M}(t))\exp\{-\lambda \mathrm{i}t\})''|\mathrm{d}t.$$

The rest of the proof is a simplified version of the proof of Theorem 2.1 and hence omitted. □

Acknowledgements The authors are grateful to Dr. Sriram for inviting us to contribute this article for Dr. Koul's Festschrift and to the referee for several helpful comments. A special note from the first author, P. Vellaisamy: I have known Dr. Koul for more than a decade and came to know him very closely when I first visited MSU in 2007. His immense hospitality and friendliness made my stay comfortable. On the professional side, I owe him a lot for teaching me not to compromise on quality and how to present my ideas more precisely. Another important quality I learnt form him is to always stay brisk and energetic. I am grateful forever for all his support, guidance, and encouragement. I wish him a long, healthy, and productive life.

References

Barbour AD, Čekanavičius V (2002) Total variation asymptotics for sums of independent integer random variables. Ann Probab 30:509–545

Barbour AD, Choi KP (2004) A non-uniform bound for translated Poisson approximation. Electron J Probab 9:18–36

Barbour AD, Xia A (1999) Poisson perturbations. ESAIM Probab Statist 3:131–150

Barbour AD, Xia A (2006) On Stein's factors for Poisson approximation in Wasserstein distance. Bernoulli 12(6):943–954

Brown TC, Xia A (2001) Stein's method and birth-death processes. Ann Probab 29:1373–1403

Čekanavičius V (1993) Non-uniform theorems for discrete measures. Lith Math J33:114–126

Čekanavičius V, Petrauskienė J (2011) Note on nonuniform estimate for compound Poisson approximation to 2-runs. Lith Math J 51(2):162–170

Čekanavičius V, Vellaisamy P (2013) Discrete approximations for sums of m-dependent random variables. (Submitted for publication, preprint version is at arXiv:1301.7196)

Chen LHY, Shao QM (2001) A non-uniform Berry–Esseen bound via Stein's method. Probab Theory Relat Fields 120:236–254

Chen LHY, Shao QM (2004) Normal approximation under local dependence. Ann Probab 32:1985–2028

Heinrich L (1982) A method for the derivation of limit theorems for sums of m-dependent random variables. Z Wahrscheinlichkeitstheorie verw Gebiete 60:501–515

Neammanee K (2003a) Pointwise approximation of Poisson binomial by Poisson distribution. Stoch Model Appl 6:20–26

Neammanee K (2003b) A nonuniform bound for the approximation of Poisson binomial by Poisson distribution. Int J Math Sci 48:3041–3046

Nefedova YS, Shevtsova IG (2012) Nonuniform estimates of convergence rate in the central limit theorem (in Russian). Teor Veroyatnost i Primenen 57(1):62–97

Petrov VV (1995) Limit theorems of probability theory: sequences of independent random variables. Oxford Studies in Probability 4. Clarendon Press, Oxford

Röllin A (2005) Approximation of sums of conditionally independent variables by the translated Poisson distribution. Bernoulli 11:1115–1128

Röllin A (2007) Translated Poisson approximation using exchangeable pair couplings. Ann Appl Probab 17:1596–1614

Roos B (1999) Asymptotics and sharp bounds in the Poisson approximation to the Poisson-binomial distributions. Bernoulli 5:1021–1034

Roos B (2003) Poisson approximation via the convolution with Kornya-Presman signed measures. Theory Probab Appl 48:555–560

Teerapabolarn K, Santiwipanont T (2007) Two non-uniform bounds in the Poisson approximation of sums of dependent indicators. Thai J Math 5(1):15–39

Vellaisamy P (2004) Poisson approximation for (k_1, k_2)-events via Stein–Chen method. J Appl Probab 41:1081–1092

Wang X, Xia A (2008) On negative binomial approximation to k-runs. J Appl Probab 45:456–471

About the Editors

Soumendra Lahiri is a Professor of Statistics at the North Carolina State University. His research interests include Nonparametric Statistics, Time Series, Spatial Statistics, and Statistical inference for high dimensional data. He served as an Editor of Sankhya, Series A (2007–2009) and currently, he is on the editorial boards of the Annals of Statistics and the Journal of Statistical Planning and Inference. He is a fellow of the ASA and the IMS, and an elected member of the International Statistical Institute.

Anton Schick is Professor and Chair of the Department of Mathematical Sciences at Binghamton University. His research interests include semiparametric efficiency, U-statistics, empirical likelihood, residual-based inference, incomplete data, curve estimation, and inference for stochastic processes. He currently serves on the editorial boards of Statistics and Probability Letters, Statistical Inference for Stochastic Processes, and the Journal of the Indian Statistical Association.

Ashis SenGupta is Professor (HAG), at Indian Statistical Institute, Kolkata. His research interests are in Multivariate Analysis and Inference, Probability distributions on manifolds, Directional Statistics, Reliability and Environmental Statistics. He has been a frequent vistor to several universities in USA as a visiting faculty including Stanford University; University of Wisconsin – Madison; University of California – Santa Barabara and Riverside; and Michigan State University-East Lansing. He is an Editor-in-Chief of Environmental & Ecological Statistics, Springer, USA, and an Editor of Scientiae Mathematicae Japonicae, Japan. He has been President of International Indian Statistical Association (India Chapter) for several successive terms. He is a Member of the Project Advisory Committee – Math. Sciences, DST, Government of India; a Fellow of National Academy of Sciences, India and a Fellow of American Statistical Association, USA.

T. N. Sriram is currently a Professor at the Department of Statistics, University of Georgia (UGA). He is a Fellow of the American Statistical Association and the winner of Special Sandy Beaver Teaching Award at UGA. His research interests focus on theory and applications for a variety of statistical scenarios, such as *sequential*

S. Lahiri et al. (eds.), *Contemporary Developments in Statistical Theory,*
Springer Proceedings in Mathematics & Statistics 68, DOI 10.1007/978-3-319-02651-0,

analysis for independent & dependent data; *bootstrap* methods for linear & non-linear time series, and branching processes; *robust estimation* methods for mixture models; *dimension reduction* methods and its *robust* modifications in time series and in association studies; and *sample size determination for classifiers* arising in biological studies. His research has been funded by federal agencies such as National Science Foundation, National Security Agency, and the Bureau of Labor Statistics. He has directed eleven doctoral dissertations, served as an editorial board member of three statistics journals, organized three international conferences, and served on NSF panels.