Jianqing Fan
Jianxin Pan *Editors*

Contemporary Experimental Design, Multivariate Analysis and Data Mining

Festschrift in Honour of Professor Kai-Tai Fang

Springer

Contemporary Experimental Design, Multivariate Analysis and Data Mining

Jianqing Fan · Jianxin Pan
Editors

Contemporary Experimental Design, Multivariate Analysis and Data Mining

Festschrift in Honour of Professor Kai-Tai Fang

Springer

Editors
Jianqing Fan
Department of Financial Engineering
Princeton University
Princeton, NJ, USA

Jianxin Pan
Department of Mathematics
The University of Manchester
Manchester, UK

ISBN 978-3-030-46160-7 ISBN 978-3-030-46161-4 (eBook)
https://doi.org/10.1007/978-3-030-46161-4

Mathematics Subject Classification: 62F, 62H, 62G, 62K, 62J, 62N, 62P

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

Collection and analysis of data play an important role in many fields of science and technology, such as computational biology, quantitative finance, information engineering, machine learning, neuroscience, medicine, and social sciences. Especially in the era of big data, researchers can easily collect data with huge size and complexity. While, it also often occurs that the cost for collection of each item of data is high, such as the data for missile intercept experiments. In this case, the use of design of experiments will be very crucial. Design of experiments requires researchers to collect data at some carefully designed points to enhance their statistical efficiency. Moreover, analysis of such collected multivariate data is equally important. At the occasion of his 80th birthday, we present four review papers on the contributions of Prof. Kai-Tai Fang to the fields of design of experiments, multivariate analysis, data mining, and education. Moreover, this monograph also includes twenty research articles in various fields of statistics such as experimental design, multivariate data analysis, data mining, and biostatistics.

Professor Kai-Tai Fang was elected as Fellow of the Institute of Mathematical Statistics in 1992 and Fellow of the American Statistical in 2001 as well as elective member of International Statistical Institute in 1985. He is an international expert on experimental design, multivariate analysis, and data mining. He is a distinguished scholar and prolific researcher. He has published 27 books including 6 monographs in English, and edited 11 lecture notes and proceedings on a wide range of subjects, including multivariate analysis, design of experiments, and quasi Monte Carlo methods, in addition to more than 330 referred papers. He is the co-inventor of the uniform experimental design, which nowadays has been widely used by engineers to expedite product developments. He has also developed novel statistical methods for inference in generalized multivariate analysis.

Professor Fang has received 10 Chinese nationwide awards, including The State Natural Science Award (second class) with Prof. Yuan Wang in 2008. He also received the 2014 Distinguished Achievement Award by the International Chinese Statistical Association (ICSA). In addition, Prof. Fang has actively participated in a large array of consulting projects, including the designs of chemical and biological experiments and standardization of Chinese garments. As a leading figure in Hong

Kong and Mainland China, he has greatly popularized the use of statistics in academic research and industry, enthusiastically participated in organizing various professional meetings, and provided conscientious professional and editorial services. He is a strong professional leader and a dedicated educator, who has fostered many generations of fertile statisticians worldwide.

In this monograph, four review papers on Prof. Kai-Tai Fang's contributions to four different areas are presented. Min-Qian Liu, Dennis Kon-Jin Lin, and Yongdao Zhou introduced Prof. Kai-Tai Fang's contributions to design of experiments especially the fields of uniform design. Jianxin Pan, Jiajuan Liang, and Guoliang Tian reviewed Prof. Kai-Tai Fang's contributions to multivariate statistics. Ping He, Xiaoling Peng, and Qingsong Xu gave an overview of the contributions of Prof. Fang to data mining. In addition, Gang Li and Xiaoling Peng presented an overview of Prof. Kai-Tai Fang's contributions to the education, promotion, and advancement of statistics in China.

Besides the four review papers on Prof. Kai-Tai Fang's numerous contributions, we also collect twenty invited research articles on a wide range of topics that are grouped into three parts. They are independent of each other. Each is dedicated to a specific issue on multivariate analysis, design of experiments, biostatistics, and other statistical issues. This book is targeted to a broad readership. We hope that regardless of their background, readers will find some parts that are of their interests and suit their needs.

The second part of the monograph includes seven articles on design of experiments. It begins with the topic of low discrepancy design by Yiou Li, Lulu Kang, and Fred J. Hickernell, followed by Peter Winker, Jianbin Chen, and Dennis Kon-Jin Lin. Then Mei Zhang, Aijun Zhang, and Yongdao Zhou introduced the tool of inverse Rosenblatt transformation for the construction of uniform designs on arbitrary domains. Ming T. Tan and Hong-Bin Fang applied uniform experimental design to drug combination studies. Hongyan Jiang and Rongxian Yue proposed a modified robust design criterion for Poisson mixed effects models. Si Qiu, Minyu Xie, Hong Qin, and Jianhui Ning enriched the theory of orthogonal array composite design. Moreover, Yu Tang proposed certain construction methods of the uniform design on manifold.

The third part of the monograph includes four articles on multivariate analysis. It begins with an application of the theory of spherical distributions to multiple mean comparisons by Jiajuan Liang, Man-Lai Tang, Jing Yang, and Xuejing Zhao, followed by Jian-Lun Xu's investigation on estimating the location vector for spherically symmetric distributions. Milan Stehlík, Mirtha Pari Ruiz, Silvia Stehlíková, and Ying Lu discussed equidistant designs, symmetries, and their violations in multivariate models. Defei Zhang, Xiangzhao Cui, Chun Li, and Jianxin Pan proposed a novel method to estimate the high-dimensional covariance matrix with autoregressive moving average (ARMA) structure through quadratic loss function.

The fourth part is about recent developments in data mining with three articles. Victoria Chen and Heping Zhang proposed a novel implementation of a depth variable importance score in a classification tree designed for precision medicine.

Then, Elaheh Oftadeh and Jian Zhang investigated Bayesian mixture models with weight-dependent component priors for Bayesian clustering. Moreover, Tianming Zhu and Jin-Ting Zhang proposed the cosine similarity-based classifiers for functional data.

The last part of the monograph includes two articles on statistical hypothesis test and four articles on statistical modeling and analysis. It begins with projection test for high-dimensional one sample mean problem by Wanjun Liu and Runze Li, followed by goodness-of-fit tests for correlated bilateral data from multiple groups investigated by Xiaobin Liu and Chang-Xing Ma. In the development of statistical models, Chengcheng Hao, Feng Li, and Dietrich von Rosen introduced a bilinear reduced rank model; Xiaoying Sun and Yuehua Wu proposed a new method for the estimation of simultaneous multiple change points in generalized linear models; Mingyao Ai, Yimin Huang, and Jun Yu considered two data-based algorithms for proper priors in Bayesian model averaging; Moreover, Baobin Wang, Ting Hu, and Hong Yin discussed the quantile regression with Gaussian kernels.

We are most grateful to the enthusiastic supports from colleagues and friends who helped to make this volume possible. We owe many thanks to Yongdao Zhou for his assistance in turning collective contributions and editing the draft of such a wonderful monograph. Each article was reviewed critically by referees. We are especially grateful to Mingyao Ai, Gang Li, Runze Li, Jiajuan Liang, Min-Qian Liu, Chang-Xing Ma, Jianhui Ning, Xiaoling Peng, Yu Tang, Guoliang Tian, Yuehua Wu, Jian-Lun Xu, Rongxian Yue, Aijun Zhang, Jian Zhang, Jin-Ting Zhang, and Yongdao Zhou for their invaluable and conscientious refereeing services. We give special thanks to Simo Puntanen of University of Tampere, Finland and Eva Hiripi of Springer for advice and encouragement. Most importantly, as the former students of Professor Kai-Tai Fang, we would like to wholeheartedly thank him for bringing us into the world of statistics, sharing with us his scientific creativity and fertile imagination, teaching us philosophy of sciences, and showing us how to mentor and foster younger generations. Many of our achievements reflect his scientific vision and dedication. We are very proud of him, as a teacher and a friend. We wish him all the best for his future life.

Princeton, USA — Jianqing Fan
Manchester, UK — Jianxin Pan
January 2020

Contents

Part I Review of Kai-Tai Fang's Contribution

1 Walking on the Road to the Statistical Pyramid 3
Jianxin Pan, Jiajuan Liang, and Guoliang Tian
1.1 Statistics in China Before 1980's 3
1.2 Multivariate Analysis and Generalized Multivariate Statistics 5
1.2.1 Development of the Theory of Elliptically Contoured Distributions 5
1.2.2 Application of the Theory of Spherical Matrix Distributions 7
1.3 General Multivariate Symmetric and Related Distributions 8
1.3.1 From Spherical Distributions to the l_1-norm Symmetric Distributions 9
1.3.2 Other Related Multivariate Distributions 10
1.4 Directional Data Analysis, Occupancy Problem, Growth Curve Model, and Miscellaneous Directions 12
1.4.1 Directional Data Analysis and Occupancy Problem 12
1.4.2 Growth Curve Model and Miscellaneous Directions 13
References 14

2 The Contribution to Experimental Designs by Kai-Tai Fang 21
Min-Qian Liu, Dennis K. J. Lin, and Yongdao Zhou
2.1 Introduction 21
2.2 The Contribution to Uniform Designs 22
2.2.1 Uniformity Measures 23
2.2.2 Construction Methods of Uniform Designs 25

2.3 More About Uniform Designs 27
2.3.1 Connection Between Uniform Designs and Other Types of Designs 27
2.3.2 Uniform Designs for Experiments with Mixture 28
2.3.3 Application of Uniform Designs 29
2.4 The Contribution to Orthogonal Designs 30
2.5 The Contribution to Supersaturated Designs 31
2.6 Conclusion 32
References 32

3 From "Clothing Standard" to "Chemometrics" 37
Ping He, Xiaoling Peng, and Qingsong Xu
3.1 Introduction 38
3.2 Establishment of the First Chinese Adult Clothing Standard . . . 38
3.3 Revision of the National Standard for Alloy Structural Steel . . . 40
3.4 Contributions to Chemometrics 42
3.5 Research Group's Further Contributions to Chemometrics 44
3.6 Summary 46
References 46

4 A Review of Prof. Kai-Tai Fang's Contribution to the Education, Promotion, and Advancement of Statistics in China 49
Gang Li and Xiaoling Peng
4.1 Background 49
4.2 Development and Popularization of Statistics Through Applications 50
4.2.1 The Early Days of Statistical Popularization and Education in China 50
4.2.2 Determination and Examination of National Standards 51
4.2.3 Number Theory Methods in Statistics 53
4.3 Contributions to Statistical Education 54
4.3.1 Cultivating Outstanding Statistical Talents as a Pioneer of Statistics in China 54
4.3.2 Creating Undergraduate Statistics Major for Liberal Arts Education 56
4.3.3 Improving and Writing Statistical Textbooks 57
4.4 Academic Services 58
4.4.1 Reforms in the Institute of Applied Mathematics at Chinese Academy 59
4.4.2 Organize Academic Conferences and Promote Research Communications 59

Appendix A.1: Academic Conferences Organized by Prof. Kai-Tai Fang . . . 61
Appendix A.2: Prof. Kai-tai Fang's Academic Services . . . 62
References . . . 64

Part II Design of Experiments

5 Is a Transformed Low Discrepancy Design Also Low Discrepancy? . . . 69
Yiou Li, Lulu Kang, and Fred J. Hickernell
5.1 Introduction . . . 70
5.2 The Discrepancy . . . 71
5.2.1 Definition in Terms of a Norm on a Hilbert Space of Measures . . . 72
5.2.2 Definition in Terms of a Deterministic Cubature Error Bound . . . 74
5.2.3 Definition in Terms of the Root Mean Squared Cubature Error . . . 76
5.3 When a Transformed Low Discrepancy Design Also Has Low Discrepancy . . . 78
5.4 Do Transformed Low Discrepancy Points Have Low Discrepancy More Generally . . . 80
5.5 Improvement by the Coordinate-Exchange Method . . . 83
5.6 Simulation . . . 87
5.7 Discussion . . . 88
Appendix . . . 90
References . . . 91

6 The Construction of Optimal Design for Order-of-Addition Experiment via Threshold Accepting . . . 93
Peter Winker, Jianbin Chen, and Dennis K. J. Lin
6.1 Introduction . . . 94
6.2 Preliminary . . . 95
6.2.1 PWO Model . . . 95
6.2.2 Tapered PWO Model . . . 96
6.2.3 Some Optimality Criteria . . . 97
6.3 The Threshold Accepting Algorithm . . . 98
6.4 Main Results . . . 100
6.5 Example: Scheduling Problem . . . 103
6.6 Conclusion . . . 107
References . . . 108

7 Construction of Uniform Designs on Arbitrary Domains by Inverse Rosenblatt Transformation 111
Mei Zhang, Aijun Zhang, and Yongdao Zhou
7.1 Introduction 111
7.2 Inverse Rosenblatt Transformation Method 113
7.3 Construction Results 116
7.3.1 Flexible Regions 116
7.3.2 Constrained Domain 120
7.3.3 Manifold Domain 120
7.3.4 Geographical Domain 122
7.4 Conclusion 123
Appendix: Good Lattice Point Method 124
References 126

8 Drug Combination Studies, Uniform Experimental Design and Extensions 127
Ming T. Tan and Hong-Bin Fang
8.1 Introduction 127
8.2 Statistical Modeling for Drug Combinations 129
8.3 Experimental Design Based on Uniform Measures 132
8.3.1 Design for Log-Linear Dose-Responses 132
8.3.2 Design for Linear Dose-Responses 134
8.3.3 Design for Two Linear Dose-Responses and One Log-Linear Dose-Response 135
8.3.4 Design for One Linear Dose-Response and Two Log-Linear Dose-Responses 136
8.4 Experimental Design for Multi-drug Combinations 138
8.5 Discussion and Further Research 142
References 143

9 Modified Robust Design Criteria for Poisson Mixed Models 145
Hongyan Jiang and Rongxian Yue
9.1 Introduction 145
9.2 The Poisson Mixed Model 146
9.2.1 Poisson Mixed Models 147
9.2.2 Fisher Information Matrix of the Model 147
9.3 Robust Optimal Designs 148
9.3.1 Locally D-Optimal Designs 148
9.3.2 RPD-and RPMMD-Optimalities 150
9.4 Numerical Studies 151
9.4.1 Designs for the First-Order Poisson Mixed Model 152
9.4.2 Designs for the Second-Order Poisson Mixed Model 156

9.5 Concluding Remarks . . . 159
References . . . 160

10 Study of Central Composite Design and Orthogonal Array Composite Design . . . 163
Si Qiu, Minyu Xie, Hong Qin, and Jianhui Ning
10.1 Introduction . . . 163
10.2 Preliminaries . . . 165
10.3 *D*-efficiencies of CCDs and OACDs . . . 166
10.4 The Determination of the α Value . . . 170
10.5 Conclusion Remarks . . . 171
Appendix . . . 172
References . . . 174

11 Uniform Design on Manifold . . . 177
Yu Tang
11.1 Background . . . 177
11.2 General Discrepancy on Manifold . . . 179
11.3 Uniform Design on Semi-spherical Surface . . . 181
11.4 Uniform Design on Spherical Surface . . . 183
11.5 Conclusion and Discussion . . . 185
References . . . 186

Part III Multivariate Analysis

12 An Application of the Theory of Spherical Distributions in Multiple Mean Comparison . . . 189
Jiajuan Liang, Man-Lai Tang, Jing Yang, and Xuejing Zhao
12.1 Introduction . . . 190
12.2 Construction of the Exact *F*-tests and the Generalized *F*-test . . . 192
12.3 A Monte Carlo Study and a Real Example . . . 194
12.3.1 Empirical Power Performance . . . 194
12.3.2 An Illustrative Application . . . 196
12.4 Concluding Remarks . . . 197
References . . . 198

13 Estimating the Location Vector for Spherically Symmetric Distributions . . . 201
Jian-Lun Xu
13.1 Introduction . . . 201
13.2 Main Results . . . 203
13.3 Extensions to Other Loss Functions and the Unknown Scale Case . . . 206
13.4 Discussion . . . 207

13.5 Proofs . . . 208
References . . . 214

14 On Equidistant Designs, Symmetries and Their Violations in Multivariate Models . . . 217
Milan Stehlík, Mirtha Pari Ruiz, Silvia Stehlíková, and Ying Lu
14.1 Introduction . . . 217
14.2 On Uniform Optimal Designs . . . 218
14.3 On Symmetric Multivariate Distributions and Beyond . . . 219
14.3.1 Pseudoexponential Models for Dose Finding Studies . . . 219
14.3.2 Asymmetric Cultural Distance Measures on Linguistic Sequences . . . 220
References . . . 224

15 Estimation of Covariance Matrix with ARMA Structure Through Quadratic Loss Function . . . 227
Defei Zhang, Xiangzhao Cui, Chun Li, and Jianxin Pan
15.1 Introduction . . . 227
15.2 Estimation Process . . . 230
15.3 Numerical Experiments . . . 235
15.3.1 Simulation Studies . . . 235
15.3.2 Real Data Analysis . . . 237
15.4 Conclusions . . . 238
References . . . 239

Part IV Data Mining

16 Depth Importance in Precision Medicine (DIPM): A Tree and Forest Based Method . . . 243
Victoria Chen and Heping Zhang
16.1 Introduction . . . 243
16.2 Methods . . . 245
16.2.1 Overview . . . 245
16.2.2 Depth Variable Importance Score . . . 245
16.2.3 Split Criteria . . . 247
16.2.4 Random Forest . . . 248
16.2.5 Best Predicted Treatment Class . . . 248
16.2.6 Splits by Variable Type . . . 248
16.2.7 Comparison Method . . . 249
16.2.8 Implementation . . . 250
16.3 Simulation Studies . . . 250
16.3.1 Methods . . . 250
16.3.2 Scenarios . . . 250
16.3.3 Results . . . 254

16.4 Analysis of CCLE Data . . . 256
16.5 Discussion . . . 257
References . . . 258

17 Bayesian Mixture Models with Weight-Dependent Component Priors . . . 261
Elaheh Oftadeh and Jian Zhang
17.1 Introduction . . . 261
17.2 Methodology . . . 263
17.2.1 Mixture of Univariate Normals . . . 264
17.3 A Simulation Study . . . 271
17.3.1 Adjusted Rand Index . . . 271
17.3.2 Simulated Data . . . 272
17.3.3 Results . . . 272
17.4 Application to a Real Dataset . . . 274
17.5 Conclusion . . . 275
References . . . 275

18 Cosine Similarity-Based Classifiers for Functional Data . . . 277
Tianming Zhu and Jin-Ting Zhang
18.1 Introduction . . . 277
18.2 Functional Dissimilarity Measures . . . 279
18.3 Functional Cosine Similarity . . . 280
18.4 Cosine Similarity-Based Classifiers for Functional Data . . . 282
18.4.1 FCD-Based Centroid Classifier . . . 282
18.4.2 FCD-Based kNN Classifier . . . 283
18.4.3 Theoretical Properties of the FCD-Based Centroid Classifier . . . 283
18.5 A Simulation Study . . . 285
18.6 Application to Australian Rainfall Data . . . 287
18.7 Concluding Remarks . . . 289
18.8 Appendix . . . 289
References . . . 291

Part V Hypothesis Test and Statistical Models

19 Projection Test with Sparse Optimal Direction for High-Dimensional One Sample Mean Problem . . . 295
Wanjun Liu and Runze Li
19.1 Introduction . . . 295
19.2 Projection Test with Sparse Optimal Direction . . . 299
19.3 Simulation Studies . . . 301
19.4 Real Data Example . . . 306
References . . . 308

20 Goodness-of-fit Tests for Correlated Bilateral Data from Multiple Groups 311
Xiaobin Liu and Chang-Xing Ma
20.1 Introduction 311
20.2 Models for Correlated Bilateral Data 313
20.2.1 Independence Model 313
20.2.2 Rosner's Model 313
20.2.3 Equal Correlation Coefficients Model 314
20.2.4 Dallal's Model 315
20.2.5 Saturated Model 316
20.3 Methods for Goodness-of-Fit Test 316
20.4 Simulation Study 318
20.5 Real World Examples 319
20.6 Conclusions 326
References 326

21 A Bilinear Reduced Rank Model 329
Chengcheng Hao, Feng Li, and Dietrich von Rosen
21.1 Introduction 329
21.2 Model 331
21.2.1 Example 332
21.3 Estimation 334
21.4 Discussion 339
References 339

22 Simultaneous Multiple Change Points Estimation in Generalized Linear Models 341
Xiaoying Sun and Yuehua Wu
22.1 Introduction 341
22.2 Simultaneous Multiple Change Points Detection 343
22.2.1 The GLM with Multiple Change Points 343
22.2.2 The Method 344
22.2.3 The Consistency of the Proposed Estimator 345
22.3 An Algorithm 347
22.4 Simulation Studies 349
22.4.1 Two Specific Generalized Linear Models 349
22.4.2 GLMs with No Change Point 349
22.4.3 GLMs with Multiple Change Points 350
22.5 A Real Data Example 351
22.6 Discussion 352
Appendix A: A Single Change Point Detection and Estimation in GLM 353
Appendix B: Proof of Theorem 22.1 353
References 355

23 Data-Based Priors for Bayesian Model Averaging 357
M. Ai, Y. Huang, and J. Yu
23.1 Introduction 357
23.2 Sequential Adjustment of Priors 359
23.3 Priors Based on Historical Data 362
23.4 Simulations 365
23.4.1 Synthetic Data Analysis 365
23.4.2 Real Data Study 367
23.5 Concluding Remarks 369
Appendix 370
References 372

24 Quantile Regression with Gaussian Kernels 373
Baobin Wang, Ting Hu, and Hong Yin
24.1 Introduction 373
24.2 Main Results and Effects of Parameters 376
24.3 Error Analysis and Proofs of Main Results 378
24.3.1 Approximation Error 379
24.3.2 Insensitive Analysis 379
24.3.3 One Step-Iteration 380
24.3.4 Sample Error Estimate 382
24.3.5 Bounding the Total Error 384
References 386

Part I
Review of Kai-Tai Fang's Contribution

Chapter 1
Walking on the Road to the Statistical Pyramid

–Prof. Kai-Tai Fang's Contribution to Multivariate Statistics

Jianxin Pan, Jiajuan Liang, and Guoliang Tian

Abstract This paper reviews Prof. Kai-Tai Fang's major contribution to multivariate statistics in three aspects: generalized multivariate statistics; general symmetric multivariate distributions; growth curve models and miscellaneous fields. Generalized multivariate statistics is a large extension of traditional statistics with normal assumption. It aims to generalize the traditional statistical methodologies like parametric estimation, hypothesis testing, and modeling to a much wider family of multivariate distributions, which is called elliptically contoured distributions (ECD). General symmetric multivariate distributions form an even wider class of multivariate probability distributions that includes the ECD as its special case. Growth curve models (GCM) includes statistical methods that allow for consideration of inter-individual variability in intra-individual patterns of change over time. Outlier detection and identification of influential observations are important topics in the area of the GCM. Miscellaneous fields cover major contributions that Prof. Fang made in various areas of multivariate statistics beyond the three aspects mentioned above.

1.1 Statistics in China Before 1980's

Professor Pao-Lu Hsu (1910–1970) is generally considered as the founder of probability and statistics in China. It is known that Prof. Hsu was the first teacher to offer courses in probability and statistics in the old "Southwest United University" in

J. Pan
University of Manchester, Manchester, UK
e-mail: Jianxin.Pan@manchester.ac.uk

J. Liang (✉)
University of New Haven, West Haven, USA
e-mail: jliang@newhaven.edu

G. Tian
Southern University of Science and Technology, Shenzhen, China
e-mail: tiangl@sustc.edu.cn

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_1

Kunming of China in 1940's during the Sino-Japanese war in World War II [9]. Prof. Hsu finished his Ph.D. study in the University College of London of the United Kingdom in 1938 and pursued his research in the United States in the last few years of 1940's. He returned to Peking University in 1947 and taught for more than 20 years there. In late 1950's, Kai-Tai Fang was one of Prof. Hsu's students in a series of seminar classes in probability and statistics [32]. Since then, Kai-Tai Fang developed more and more interests in statistics and devoted his lifetime career to statistics. During the early development of probability and statistics in China between early 1960's and before 1980's, probability and statistics were considered as a small unit in any mathematics departments of universities in China. Because of the serious shortage of teachers in probability and statistics, professors in this small unit mainly focused on teaching before 1980's. Prof. Kai-Tai Fang was one of the very few lecturers who insisted on doing research in the old "closed society" before 1977 although research topics are mainly focused on application of statistics in the industrial area, see, for example [18, 34, 35, 58, 59, 99]. The year of 1977 was the most memorable year in the history of the higher education in China since 1949 when the whole academia was re-open after the 10-year "Cultural Revolution". Since 1977, both theoretical and applied research in all areas of science and technology was highly recognized in academic institutions of China. Based on his non-stopping efforts in pursuing probability and statistics research during the lost ten years, Prof. Fang became one of the leading researchers in mathematical statistics and its applications in various areas in China. While his research in both theoretical and applied statistics had been continuing in 1970's [19, 20, 36, 37, 100, 118], Prof. Fang also collaborated with his colleagues in writing statistics textbooks to meet the urgent need of statistical education in China in late 1970's [37], (Fang et al. 1979).

The last three years (1977, 1978, 1979) of 1970's is generally considered as a period of academic revival of Chinese higher education after the ten-year "Cultural Revolution" (1966–1976). Many scientists and researchers burst out a kind of never-seen energy in pursuing new knowledge and accomplishments after being forced out of their academic life for ten years. Prof. Fang belonged to the small group of researchers who could focus most of their time on research and never stopped along their research directions. The strong basis laid down from Prof. Pao-Lu Hsu's seminar classes helped Fang's research throughout the early years in his statistical career. The statistical foundation knowledge trained from Prof. Hsu's classes and his never-give-up ambition in pursuing high-quality statistical career turned out to equip Fang with inexhaustible resources in his future years of being a highly productive statistician and a well-known statistical educator. Sections 1.2, 1.3 and 1.4 will introduce Prof. Fang's creative contributions to multivariate statistics in the last two decades of the 20th century.

1.2 Multivariate Analysis and Generalized Multivariate Statistics

After Prof. Pao-Lu Hsu opened the statistical door for young Chinese statisticians and led them into the realm of classical statistics in late 1950's, a number of Prof. Hsu's students grew up in late 1970's. Among these students, Prof. Yao-Ting Zhang (1933–2007) [31] and Prof. Kai-Tai Fang made significant contributions to inheritance of Prof. Hsu's major idea in multivariate statistical analysis and its application in various areas. Both Profs. Zhang and Fang not only brought with their own prolific research accomplishments but also trained a large number of graduate students and statistical practitioners from various institutions of Chinese higher education. Facing the almost empty of statistical textbooks and readings in the higher education of China in late 1970's and the early years of 1980's, both Profs. Zhang and Fang, cooperating with their colleagues, published a few urgent-needed statistical textbooks to meet the needs of college students and postgraduate students in their beginning study in statistics, for example [29, 38, 62, 63, 68, 70–72, 101, 123]. All of these early statistical textbooks and readings greatly enriched the urgent needs for students in Chinese higher education in the whole 1980's. By training graduate students and organizing statistical seminars and workshops in various directions, Prof. Fang took the leading role in developing new research directions in multivariate statistics and statistical education during the last twenty years of the 20th century. Profs. Fang and Zhang helped open numerous Chinese young statisticians' eyes in entering the realm of modern multivariate statistics and its applications through their productive research accomplishments and comprehensive statistical education. Profs. Zhang and Fang are generally considered as the pioneers and initiators of multivariate statistics and statistical education after Prof. Pao-Lu Hsu in the 20th century of China.

1.2.1 Development of the Theory of Elliptically Contoured Distributions

With the rapid development of the statistical science and computer science in the last two or three decades of the 20th century, classical statistics under the normal assumption can no longer meet the needs of high-dimensional data analysis. Statisticians in the world have long realized the phenomenon and reality of fat-tailed distributed data. Normal-theory-based statistical methods become doubtable when applied to this kind of data. Modern computer technology and algorithms make it possible to analyze a large amount of high-dimensional data beyond the classical normal assumption. Some early research on extending the normal-theory-based statistical methods to the ones under a wider class of probability distributions, which is called the elliptically contoured distributions (ECD for simplicity), includes [2–8, 11–15, 28, 33, 40, 55, 69, 71, 73–75, 112, 113, 122, 124].

The stochastic representation method plays an important role in the development of the theory on ECD. For example, the p-dimensional normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ has a stochastic representation

$$\boldsymbol{x} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{y}, \tag{1.1}$$

where $\boldsymbol{A}\boldsymbol{A}' = \boldsymbol{\Sigma}$, $\boldsymbol{y}$ has the standard normal distribution $N_p(\boldsymbol{o}, \mathrm{i})$ (i stands for the identity matrix), and "$\stackrel{\mathrm{d}}{=}$" denotes that the two sides of the equality have the same probability distribution. Equation (1.1) is called the stochastic representation of the multivariate normal distribution. One can pay attention to the fact that for any constant $p \times p$ orthogonal matrix $\boldsymbol{\Gamma}$, it is always true $\boldsymbol{\Gamma}y \stackrel{\mathrm{d}}{=} \boldsymbol{y}$ for $\boldsymbol{y} \sim N_p(\boldsymbol{o}, \mathrm{i})$. The probability distribution of $\boldsymbol{y}$ is said to have rotational invariance or to have spherical symmetry. The idea of spherical symmetry can be extended to the general case by defining a family of random vectors satisfying spherical symmetry:

$$\mathscr{S}_p(\phi) = \{\boldsymbol{x} : \ \boldsymbol{\Gamma}\boldsymbol{x} \stackrel{\mathrm{d}}{=} \boldsymbol{x} \ \text{ for any constant } p \times p \text{ orthogonal matrix } \boldsymbol{\Gamma}\}, \tag{1.2}$$

where $\phi(\cdot)$ stands for the characteristic function of a distribution. $\mathscr{S}_p(\phi)$ is called the family of spherically symmetric distributions or simply called spherical distributions. It is obvious that $\mathscr{S}_p(\phi)$ includes that the standard normal distribution $N_p(\boldsymbol{o}, \mathrm{i})$ and some commonly known multivariate distributions such as the multivariate Student t-distribution with zero mean and identity covariance matrix. It is known that $\boldsymbol{x} \in \mathscr{S}_p(\phi)$ if and only if

$$\boldsymbol{x} \stackrel{\mathrm{d}}{=} R\mathbf{u}^{(p)}, \tag{1.3}$$

where $\mathbf{u}^{(p)}$ stands for the uniform distribution on the surface of the unit sphere in R^p (the p-dimensional real space), that is , $\|\mathbf{u}^{(p)}\| = 1$ ($\|\cdot\|$ stands for the usual Euclidean norm), and $R > 0$ is a random variable that is independent of $\mathbf{u}^{(p)}$. Equation (1.3) is called the stochastic representation for a spherical distribution. For any nontrivial $\boldsymbol{x} \in \mathscr{S}_p(\phi)$ with $P(\boldsymbol{x} = \boldsymbol{o}) = 0$, it is always true that

$$\boldsymbol{x} \stackrel{\mathrm{d}}{=} \|\boldsymbol{x}\| \cdot \frac{\boldsymbol{x}}{\|\boldsymbol{x}\|}, \tag{1.4}$$

where $\|\boldsymbol{x}\|$ and $\boldsymbol{x}/\|\boldsymbol{x}\|$ are independent, and $\boldsymbol{x}/\|\boldsymbol{x}\| \stackrel{\mathrm{d}}{=} \mathbf{u}^{(p)}$.

Equation (1.1) is a linear transformation of the standard normal $N_p(\boldsymbol{o}, \mathrm{i})$ and gives a family of general multivariate normal distributions by choosing different linear transformations. This idea can be applied to the distributions in $\mathscr{S}_p(\phi)$ and gives a bigger family of distributions:

$$ECD_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi) = \{\boldsymbol{x};\ \boldsymbol{x} \stackrel{\mathrm{d}}{=} \boldsymbol{\mu} + \boldsymbol{A}\boldsymbol{y},\ \boldsymbol{y} \in \mathscr{S}_p(\phi),\ \boldsymbol{\mu} \in R^p,\ \boldsymbol{A}\boldsymbol{A}' = \boldsymbol{\Sigma}\}. \tag{1.5}$$

$ECD_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$ is called the family of elliptically contoured distributions or simply called elliptical distributions. The Eq. (1.1) with $\boldsymbol{y} \in \mathscr{S}_p(\phi)$ is called the stochastic

representation for an elliptical distribution. One can imagine that an elliptical distribution would have similar properties to those of the normal distribution $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. For example, if $\boldsymbol{x} \in ECD_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$ possesses a probability density function $f(\boldsymbol{x})$, it must have the form

$$f(\boldsymbol{x}) = c\boldsymbol{\Sigma}^{-\frac{1}{2}} g[(\boldsymbol{x} - \boldsymbol{\mu})' \boldsymbol{\Sigma}^{-1} (\boldsymbol{x} - \boldsymbol{\mu})], \tag{1.6}$$

where $g(\cdot) > 0$ is a scalar function and $c > 0$ is a normalizing constant. For example, if $\boldsymbol{x} \sim N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $g(x) = \exp(-x/2)$.

The method of stochastic representation used in Eqs. (1.1)–(1.5) plays an important role in developing some theory on ECD. Some statistical inference on the mean parameter $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ in $ECD_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \phi)$ was developed by Fang and his collaborators. Their comprehensive outcomes are summarized in [4–6, 77]. Some goodness-of-fit methods for spherical symmetry (a subfamily of ECD) were developed by Fang and his collaborators, for example [79, 88, 91, 92, 125, 127]. Some major approaches to testing spherical and elliptical symmetry were summarized in [56] and updated by [30].

1.2.2 Application of the Theory of Spherical Matrix Distributions

Prof. Fang's contribution to the area of multivariate analysis and generalized multivariate statistics, including papers, monographs, and textbooks, has been cited by many international researchers in developing new statistical methodologies for data analysis. For example [80, 85–87], employed the major theory of spherical matrix distributions in [77] to developed a class of exact multivariate tests for normal statistical inference. These tests can be still effectively applicable under high dimension with a small sample size, which may be smaller than the dimension of sample data. The tests developed by Läuter and his associates provide exact solutions to multivariate normal mean comparisons under high dimension with a small sample size. These tests extend the traditional Hotelling's T^2-test to the multiple mean comparisons as in multivariate analysis of variance (so-called MANOVA) and general linear tests for regression coefficients in multivariate regression models. Their tests are still applicable with fair power performance even in the case that the sample size is smaller than the dimension of sample data, see [84].

An $n \times p$ random matrix $\boldsymbol{X}$ is said to have a left-spherical matrix distribution, denote by $\boldsymbol{X} \sim LS_{n\times p}(\phi)$, if for any constant orthogonal matrix $\boldsymbol{\Gamma}$ $(n \times n)$

$$\boldsymbol{\Gamma X} \stackrel{\mathrm{d}}{=} \boldsymbol{X}. \tag{1.7}$$

It is known that $\boldsymbol{X} \sim LS_{n\times p}(\phi)$ if and only if $\boldsymbol{X}$ has the stochastic representation

$$\boldsymbol{X} \stackrel{\mathrm{d}}{=} \boldsymbol{U}V, \tag{1.8}$$

where $\boldsymbol{U}$ $(n \times p)$ is independent of $V\,(p \times p)$ and $\boldsymbol{U} \sim \mathscr{U}^{(n\times p)}$, which is uniformly distributed on the Stielfel manifold

$$\mathscr{D}(n, p) = \{\boldsymbol{H}_{n\times p} : \boldsymbol{H}'\boldsymbol{H} = I_p\}. \tag{1.9}$$

If $\boldsymbol{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)'$ $(n \times p)$ consists of i.i.d. observations from $N_p(\boldsymbol{0}, \boldsymbol{\Sigma})$, then $\boldsymbol{X} \sim LS_{n\times p}(\phi)$ and $\boldsymbol{X}$ has a stochastic representation (1.8). For any random matrix $\boldsymbol{D}_{p\times q}$ $(q \le p)$, which is a function of $\boldsymbol{X}$ in the quadratic form $\boldsymbol{D} = f(\boldsymbol{X}'\boldsymbol{X})$, it can be proved that $\boldsymbol{X}\boldsymbol{D} \sim LS_{n\times p}(\phi)$. So $\boldsymbol{X}\boldsymbol{D}$ also has a stochastic representation similar to (1.8), say, $\boldsymbol{X}\boldsymbol{D} \stackrel{\mathrm{d}}{=} \boldsymbol{U}\boldsymbol{A}$ and $\boldsymbol{U} \sim \mathscr{U}^{(n\times q)}$ that is independent of $\boldsymbol{A}$ $(q \times q)$. As a result of this stochastic representation, any affine-invariant statistic $T(\cdot)$ satisfies $T(\boldsymbol{X}\boldsymbol{D}) \stackrel{\mathrm{d}}{=} T(\boldsymbol{U})$, whose distribution is uniquely determined no matter how to choose the quadratic function $\boldsymbol{D} = f(\boldsymbol{X}'\boldsymbol{X})$. One can always choose $q \le p$ as dimension reduction for $\boldsymbol{U} \sim \mathscr{U}^{(n\times q)}$. For example, let $q = \min(n,\ p) - 1$, this will make a statistic $T(\boldsymbol{X}\boldsymbol{D})$ applicable for the case of high dimension with a small sample size, even for $p \le n$. This is the main idea in constructing Läuter and his associates' parametric tests.

By using the idea of spherical matrix distribution in [77, 85] and his associates' (1998) approach to constructing multivariate parametric tests, Prof. Fang led his graduate students and colleagues to develop a class of nonparametric goodness-of-fit tests for multivariate normality for the case of high dimension with a small sample size, including some graphical methods for detecting non-normality with confidence regions, and a class of tests for spherical symmetry. The representative papers are: [54, 57, 91–93]. Fang's approach to constructing multivariate tests and graphical methods for goodness-of-fit purpose was further developed by his graduate students and associates, see, for example [1, 89, 90, 94–98]. These papers are all based on the comprehensive study in [52, 77].

1.3 General Multivariate Symmetric and Related Distributions

Beyond the ECD are some classes of general multivariate symmetric distributions. A systematic summary of general multivariate continuous distributions can be dated back to [81]. Prof. Fang's research in constructing new classes of continuous multivariate symmetric distributions and their statistical inference started in 1980's, see, for example [33, 69].

1.3.1 From Spherical Distributions to the l_1-norm Symmetric Distributions

A general continuous multivariate symmetric distribution is usually constructed by a nonnegative random combination of a multivariate uniform distribution on the surface of a unit generalized sphere. By changing the distance measure for defining the unit generalized sphere, we can construct different families of continuous multivariate symmetric distributions. By applying a linear transformation to the stochastic representation of a general continuous multivariate symmetric distribution, one can obtain an even more general continuous multivariate symmetric distribution. For example, an elliptically contoured distribution (ECD) is obtained by applying a linear transformation to the stochastic representation of a spherically symmetric distribution. Fang and Fang [42] proposed a new family of multivariate exponential distributions. Based on their result [43], constructed different families of multivariate distributions related to the exponential distribution. We follow [43] notation to define the family of distributions given by

$$F_n = \{L(z) : \ z \stackrel{\mathrm{d}}{=} R\mathbf{u}, \ R \geq 0 \text{ is independent of } \mathbf{u}\}, \tag{1.10}$$

where $\mathbf{u} = (U_1, \ldots, U_n)'$ is uniformly distributed on the l_1-norm unit sphere constrained to the positive quadrant

$$\mathscr{S}_+^1 = \{z = (z_1, \ldots, z_n)' : \ z_i \geq 0 \ (i = 1, \ldots, n), \ \|z\|_1 = \sum_{i=1}^{n} z_i = 1\}, \tag{1.11}$$

where $\|z\|_1 = \sum_{i=1}^n z_i$ is called the l_1-norm of z with nonnegative components. Fang and Fang [43] proved that for any $z = (Z_1, \ldots, Z_n)' \in F_n$, its survival function

$$P(Z_1 > z_1, \ldots, Z_n > z_n) \tag{1.12}$$

only depends on the l_1-norm $\|z\|_1 = \sum_{i=1}^n z_i$. As a result, a new family of distributions can be constructed:

$$T_n = \{L(z) : \ z = (Z_1, \ldots, Z_n)' \in R_+^n, \ P(Z_1 > z_1, \ldots, Z_n > z_n) = h(\|z\|_1)\}, \tag{1.13}$$

where $R_+^n = \{z = (z_1, \ldots, z_n)' : \ z_i \geq 0 \ (i =, \ldots, n)\}$. Fang and Fang [43] proved T_n contains a subfamily of symmetric multivariate distributions:

$$\begin{aligned} D_{n,\infty} = \{L(z) : \ & z \stackrel{\mathrm{d}}{=} R\boldsymbol{x}, \ R \geq 0 \text{ is independent of } \boldsymbol{x} = (X_1, \ldots, X_n) \\ & \text{consisting of i.i.d. } X_i \sim \exp(\lambda)\}, \end{aligned} \tag{1.14}$$

where exp(λ) stands for the exponential distribution with parameter $\lambda > 0$. $D_{n,\infty}$ is actually the family of mixtures of exponential distributions. Fang and Fang [43] proved the interesting relationship between the three families of distributions:

$$D_{n,\infty} \subset T_n \subset F_n, \tag{1.15}$$

which means that F_n is the largest family of distributions that contains T_n as its subset and T_n contains $D_{n,\infty}$ as its subset. Fang and Fang [43] obtained the general formulation of the survival function of $z = (Z_1, \ldots, Z_n)' \in F_n$:

$$P(Z_1 > a_1, \ldots, Z_n > a_n) = \int_{\|\boldsymbol{a}\|_1}^{+\infty} (1 - \|\boldsymbol{a}\|_1/r)^{n-1} dG(r), \tag{1.16}$$

where $G(r)$ is the distribution function of R in the stochastic representation (1.10), $\boldsymbol{a} = (a_1, \ldots, a_n)' \in R_+^n$. If $z = (Z_1, \ldots, Z_n)' \in F_n$ has a density function, it must have the form of $f(\|z\|_1)$ $(z \in R_+^n)$ that depends only on the l_1-norm. Fang and Fang [44] obtained the distributions of the order statistics from the family of multivariate l_1-norm symmetric distributions. Fang and Fang [45] proposed the exponential matrix distribution. Fang and Fang [46] studied statistical inference on the location and scale parameters of the multivariate l_1-norm symmetric distributions. Fang and Fan [39] studied large sample properties for distributions with rotational symmetries. Fang and Fang [16] obtained a characterization property of multivariate l_1-norm symmetric distributions. Fang and Xu [73] constructed a class of multivariate distributions including the multivariate logistic. Fang et al. [52] summarizes most of the current findings on symmetric multivariate and related distributions. The idea of defining the general distribution family F_n in (1.10) was generalized to the l_p-norm symmetric distributions by [121], and was further generalized to the L_p-norm symmetric distributions by [117].

1.3.2 Other Related Multivariate Distributions

Fang and his collaborators' research on the direction of multivariate symmetric and related distributions continued throughout the 1990's and after. For example Fang and Fang [47] constructed a class of generalized Dirichlet distributions; Fang et al. [49] constructed a family of bivariate distributions with non-elliptical contours; Fang et al. [53] introduced the L_1-norm symmetric distributions to the topic of L_1-norm statistical analysis. Kotz et al. [82] applied the method of vertical density representation to a class of multivariate symmetric distributions and proposed a new method for generating random numbers from these distributions. Rosen et al. [115] proposed an approach to extending the complex normal distribution. Zhu et al. [126] proposed a new approach to testing symmetry of high-dimensional distributions. Fang et al. [49] constructed a family of bivariate distributions with non-elliptical contours. Fang et al. [76] proposed a new approach to generating multivariate distributions by

using vertical density representation. Fang et al. [50] developed a copula method for constructing meta-elliptical distributions with given marginals. Their copula method has been cited by many international scholars in different areas, see for example, scholar.google.com [50], "The meta-elliptical distributions with given marginals" has been cited for 296 times. Among various methods for constructing multivariate distributions, the copula method is one of the most cited methods for constructing a multivariate distribution with given marginals, see, for example [83].

Fang et al. [50] idea for constructing the meta-type ECD is based on the well-known property of ECD. If $\boldsymbol{z} = (Z_1, \ldots, Z_n)' \sim ECD_n(\boldsymbol{0}, \boldsymbol{R}, g)$ with a density-generating function $g(\cdot)$ as in (1.6) and correlation matrix $\boldsymbol{R}$, the marginal p.d.f. (probability density function) of each component Z_i $(i = 1, \ldots, n)$ is given by

$$q_g(z) = \frac{\pi^{(n-1)/2}}{\Gamma((n-1)/2)} \int_{z^2}^{+\infty} (y - z^2)^{(n-1)/2} g(y) dy \tag{1.17}$$

and a cumulative distribution function (c.d.f.) given by

$$Q_g(z) = \frac{1}{2} + \frac{\pi^{(n-1)/2}}{\Gamma((n-1)/2)} \int_0^z \int_{u^2}^{+\infty} (y - u^2)^{(n-1)/2} g(y) dy du. \tag{1.18}$$

Let $\boldsymbol{x} = (X_1, \ldots, X_n)'$ be a random vector with each component X_i having a continuous p.d.f. $f_i(x_i)$ and a c.d.f. $F_i(x_i)$. Let the random vector $\boldsymbol{z} = (Z_1, \ldots, Z_n)' \sim ECD_n(\boldsymbol{0}, \boldsymbol{R}, g)$. Suppose that

$$Z_i = Q_g^{-1}(F_i(X_i)), \quad i = 1, \ldots, n, \tag{1.19}$$

where $Q_g^{-1}(\cdot)$ is the inverse of $Q_g(\cdot)$ given by (1.18). Fang et al. [50] obtained the p.d.f. of $\boldsymbol{x} = (X_1, \ldots, X_n)'$ given by

$$h(x_1, \ldots, x_n) = \phi\Big(Q_g^{-1}(F_1(x_1)), \ldots, Q_g^{-1}(F_n(x_n))\Big) \prod_{i=1}^{n} f_i(x_i), \tag{1.20}$$

where ϕ is the n-variate density weighting function:

$$\phi(z_1, \ldots, z_n) = \frac{|\boldsymbol{R}|^{-\frac{1}{2}} g(\boldsymbol{z}' \boldsymbol{R}^{-1} \boldsymbol{z})}{\prod_{i=1}^{n} q_g(z_i)}. \tag{1.21}$$

If $\boldsymbol{x} = (X_1, \ldots, X_n)'$ has a p.d.f. given by (1.20), $\boldsymbol{X}$ is said to have a meta-elliptical distribution, denote by $\boldsymbol{X} \sim ME_n(\boldsymbol{0}, \boldsymbol{R}, g;\ F_1, \ldots, F_n)$. The family $ME_n(\boldsymbol{0}, \boldsymbol{R}, g;\ F_1, \ldots, F_n)$ includes various multivariate distributions, such as $ECD_n(\boldsymbol{0}, \boldsymbol{R}, g)$, the meta-Gaussian distributions and various asymmetric distributions by choosing suitable marginal c.d.f. $F_i(x_i)$. Fang et al. [50] obtained some interesting meta-elliptical distributions in the two-dimensional case. In general, $ME_n(\boldsymbol{0}, \boldsymbol{R}, g;\ F_1, \ldots, F_n)$ is such a big family of distributions that the exact p.d.f. of any given member is diffi-

cult to obtain. Today copula method has been comprehensively studied and has been applied to various fields, see, for example [10, 17, 103].

Based on the theory of spherical distributions developed by [52, 95] proposed a class of uniform tests for goodness of fit of the L_p-norm symmetric multivariate distributions. All of the research accomplishments from Fang and his collaborators have greatly enriched the theory of general symmetric multivariate and related distributions.

1.4 Directional Data Analysis, Occupancy Problem, Growth Curve Model, and Miscellaneous Directions

Entering the open age of the economic reform of China in late 1970's and 1980's, Prof. Fang's research topics were eradicating onto various directions. For example, to meet the needs of applied statistics in industry of China, Prof. Fang carried out a series of research projects in clustering analysis, occupancy problem, mathematical statistics and standardization, quality control, and graph analysis of multivariate observations. The research outcomes from these projects were summarized in papers: [19–25, 36, 60, 64].

1.4.1 Directional Data Analysis and Occupancy Problem

Directional data analysis is one of Prof. Fangs interests in late 1980's. Directional data occurs in many areas, namely the earth sciences, meteorology and medicine. It was a hot international research area in 1970's. A summary overview on directional data analysis was given by [102]. Let $\boldsymbol{x} = (x_1, \ldots, x_p)'$ be a direction on the surface of the unit sphere $S_p = \{\boldsymbol{x} \in R^p : \|\boldsymbol{x}\| = 1\}$ (R^p stands for the usual p-dimensional Euclidean space, $\|\cdot\|$ stands for the usual distance function). Some important topics in directional data analysis include the correlation analysis of data on any two different directions $\boldsymbol{x}$ and $\boldsymbol{y}$ on S_p and regression problem like $\boldsymbol{y}$ given $\boldsymbol{x}$. Fang led his graduate students to this research area that was brand new to Chinese statisticians in late 1980's. The major research outcomes were published in their series of papers [41, 52].

In addition to focusing his research on statistical theory and its applications, Prof. Fang also carried out research on probability theory and its applications. For example, occupancy in probability theory is about the problem of reasonably assigning a set of balls into a group of cells. Although the occupancy problem originated from simple probability theory, some practical problems on resource allocation can be reduced to the solution to some kind of occupancy problems. For example, the number of units in use in hotel rooms, apartment flats, or offices, or the number of persons using an undivided space, etc., can be described as a kind of occupancy problems. The optimal allocation of limited resources reduces to the solution to an occupancy

problem. Prof. Fang's research on the occupancy problem can be dated back to early 1980's, see, for example [25–27, 61].

1.4.2 Growth Curve Model and Miscellaneous Directions

The growth curve model (GCM for simplicity) is another research field in which Prof. Fang guided his graduate students in the middle of 1980's. A general review on GCM methodologies for data analysis was given by [114]. Among others, Prof. Fang's former Ph.D. student Jianxin Pan played the leading role in developing new GCM methodologies for data analysis. Outlier detection, discovery of influential observations, and covariance structure are important topics in the GCM theory. A general formulation of GCM is [104] defined by

$$\boldsymbol{Y}_{p\times n} = \boldsymbol{X}_{p\times m}\boldsymbol{B}_{m\times r}\boldsymbol{Z}_{r\times n} + \boldsymbol{E}_{p\times n}, \tag{1.22}$$

where where $\boldsymbol{X}$ and $\boldsymbol{Z}$ are known design matrices of rank $m < p$ and $r < n$, respectively, and the regression coefficient matrix $\boldsymbol{B}$ is unknown. Furthermore, the columns of the error matrix $\boldsymbol{E}$ are independent p-variate normal with a mean vector $\boldsymbol{0}$ and a common unknown covariance matrix $\boldsymbol{\Sigma} > \boldsymbol{0}$. The GCM formulation defined by (1.22) can be written as a matrix-variate normal distribution $\boldsymbol{Y} \sim N_{p\times n}(\boldsymbol{XBZ}, \boldsymbol{\Sigma} \otimes I_n)$ ("$\otimes$" stands for the Kronecker product). The maximum likelihood estimate (MLE) for the unknown coefficient matrix $\boldsymbol{B}$ and the unknown covariance matrix $\boldsymbol{\Sigma}$ can be easily obtained from the expression of the matrix normal distribution of GCM. Pan and Fang [104] employed the mean-shift regression model to develop an approach for multiple outlier detection. Pan and Fang [105] studied the influence of a subset of observations on the growth regression fittings by comparing empirical influence functions. Pan et al. [108] proposed the Bayesian local influence approach to develop a method for GCM model diagnostics with Rao's simple covariance structure. Pan et al. [109] studied the local influence assessment in GCM with unstructured covariance under an abstract perturbation scheme. Pan et al. [110] discussed the posterior distribution of the covariance matrix of GCM. Pan and Fang [106] extended the results in [108] from Rao's simple covariance structure to unstructured covariance. Pan et al. [111] applied projection pursuit techniques to multiple outlier detection in multivariate data analysis. A comprehensive study on the current development of GCM was summarized in [107].

Prof. Fang's research interest and accomplishments have been emanating from a number of areas and applications during 1990's. Besides his contributions to the areas of generalized multivariate analysis, theory on symmetric multivariate and related distributions, occupancy problems, directional data analysis, and growth curve modeling, Prof. Fang's miscellaneous and other significant contributions to statistics can be found from Fang's series of papers. Among the miscellaneous research directions, construction of effective algorithms for complex numerical computation in statistics became one of Prof. Fang's important research directions in 1990's. For

example [65, 66, 119], proposed the sequential algorithm for optimization problems and solving nonlinear equations [67]; proposed the general applications of number-theoretic methods in statistics [116]; proposed the neural computation on nonlinear regression analysis problems [51]; proposed some global optimization algorithms in statistics [120]; discussed the quasi-Monte Carlo approaches and their applications in statistics and econometrics; and [78] proposed a two-stage algorithm associated with number-theoretic methods for numerical evaluation of integrals. In addition to the major research areas, these miscellaneous research directions, as well as their related applications, have significantly enrich Prof. Fang's field of research.

Entering the new millenium of 2000, Prof. Fang led his graduate students and worked with his collaborators on the theory and applications of uniform design and general experimental designs –the biggest research area that Prof. Fang and his collaborators have been developing with the richest outcomes. One can refer to Prof. Fang and his collaborators' series of papers in 2000's. It is no doubt that the new millenium marks Prof. Kai-Tai Fang's biggest step to the statistical pyramid. We wish Prof. Fang would never stop marching to the peak of the statistical pyramid in his lifetime as a statistician.

References

1. Ai, M., Liang, J., Tang, M.L.: Generalized T_3-plot for testing high-dimensional normality. Front. Math. China **11**, 1363–1378 (2016)
2. Anderson, T.W., Fang, K.T.: Distributions of quadratic forms and Cochran's Theorem for elliptically contoured distributions and their applications. Technical report, No.53. ONR Contract N00014-75-C 0442, Department of Statistics, Stanford University, California (1982)
3. Anderson, T.W., Fang, K.T.: On the Theory of multivariate elliptically contoured distributions. Sankhya **49** Series A, 305–315 (1987)
4. Anderson, T.W., Fang, K.T.: On the theory of multivariate elliptically contoured distributions and their applications. In: Statistical Inference in Elliptically Contoured and Related Distributions, pp.1–23. Allerton Press Inc., New York (1990)
5. Anderson, T.W., Fang, K.T.: Inference in multivariate elliptically contoured distributions based on maximum likelihood. In: Statistical Inference in Elliptically Contoured and Related Distributions, pp. 201–216. Allerton Press Inc., New York (1990)
6. Anderson, T.W., Fang, K.T. Theory and Applications elliptically contoured and related distributions. In: The Development of Statistics: Recent Contributions from China, pp. 41–62. Longman, London (1992)
7. Anderson, T.W., Fang, K.T., Hsu, H.: Maximum likelihood estimates and likelihood ratio criteria for multivariate elliptically contoured distributions. Can. J. Stat. **14**, 55–59 (1986)
8. Cambanis, S., Huang, S., Simons, G.: On the theory of elliptically contoured distributions. J. Multivar. Anal. **11**, 368–385 (1981)
9. Chen, D., Olkin, I.: Pao-Lu Hsu (Hsu, Pao-Lu): the grandparent of probability and statistics in China. Stat. Sci. **27**, 434–445 (2012)
10. Cherubini, U., Luciano, E., Vecchiato, W.: Copulas Methods in Finance. Wiley (2004)
11. Dawid, A.P.: Spherical matrix distributions and a multivariate model. J. R. Stat. Soc. (B) **39**, 254–261 (1977)
12. Dawid, A.P.: Extendibility of spherical matrix distributions. J. Multivar. Anal. **8**, 559–566 (1978)

13. Fan, J.Q., Fang, K.T.: Minimax estimator and Stein Two-stage estimator of location parameters for elliptically contoured distributions. Chin. J. Appl. Prob. Stat. **1**, 103–114 (1985)
14. Fan, J.Q., Fang, K.T.: Inadmissibility of sample mean and regression coefficients for elliptically contoured distributions. Northeast. Math. J. **1**, 68–81 (1985)
15. Fan, J.Q., Fang, K.T.: Inadmissibility of the usual estimator for location parameters of spherically symmetric distributions. Chin. Sci. Bull. **32**, 1361–1364 (in Chinese) (1987)
16. Fan, J.Q., Fang, K.T.: Inadmissibility of the usual estimator for location parameters of spherically symmetric distributions. Chin. Sci. Bull. **34**, 533–537 (1989)
17. Fan, Y., Patton, A.J.: Copulas in econometrics. Ann. Rev. Econ. **6**, 179–200 (2014)
18. Fang, K.T.: An introduction to some nonparametric statistical methods. Math. Pract. Theory **5**, 58–66 (1972)
19. Fang, K.T.: Clustering analysis. Math. Pract. Theory (I and II) **4** (No. 1), 66–80, and (No. 2), 54–62 (1978)
20. Fang, K.T.: Mathematical statistics and standardization (II–VI). Stand. Bull. **2, 3, 4**, 13–22, 13–20, 12–30 (1978)
21. Fang, K.T.: Quality control. Stand. Bull. **2**, 19–30 (1979)
22. Fang, K.T.: Process capability index C_p in quality control. Stand. Bull. **1**, 10–17 (1980)
23. Fang, K.T.: Clustering methodology and applications, pp. 12–23. Mathematical Geology, Special Issue (I). Geological Publishing House, Beijing (1980)
24. Fang, K.T.: Graph analysis of multivariate observations (I and II). Math. Pract. Theory, **7** (No. 3), 63–71, (No. 4), 42–48 (1981)
25. Fang, K.T.: Restricted occupancy problem. J. Appl. Prob. **19**, 707–711 (1982)
26. Fang, K.T.: A restricted occupancy problem and its central limit theorem. Kexue Tongbao **27**, 572–573 (1982)
27. Fang, K.T.: Occupancy problems. In: Kotz, S., Johnson, N.L. (eds.) Encyclopedia of Statistical Sciences, vol. 6, pp. 402–406. Wiley, New York (1985)
28. Fang, K.T.: A review: on the theory of elliptically contoured distributions. Adv. Math. **16**, 1–15 (1987)
29. Fang, K.T.: Applied Multivariate Analysis. East China Normal University Press, Shanghai (1989). (in Chinese)
30. Fang, K.T.: Spherical and elliptical symmetry, test of. In: Johnson, N.L., Kotz, S. (eds.) Encyclopedia of Statistical Sciences, vol. 12, 2nd edn, pp. 7924–7930. Wiley, New York (2006)
31. Fang, K.T.: Professor Yao-ting Zhang and "multivariate statistical analysis". Stat. Educ. **5**, a special issue for memory of Professor Yao-ting Zhang (2008)
32. Fang, K.T.: Professor P.L. Hsu—my supervisor forever. In: Morality and Articles Modelling in Human, pp. 406–413. Peking University Press (2010)
33. Fang, K.T., Chen, H.F.: Relationships among classes of spherical matrix distributions. Acta Mathematicae Applicatae Sinica (English Series) **1**, 139–147 (1984)
34. Fang, K.T., Dai, S.S., et al. (under the team name): Basic Methods of Mathematical Statistics. Science Press, Beijing (1973, 1974, 1979). (in Chinese)
35. Fang, K.T., Dong, Z.Q., Han, J.Y.: The structure of stationary queue without after-effect. Acta Mathematicae Applicatae and Computation Sinica **2**, 84–90 (1965)
36. Fang, K.T., et al.: (under the team name): The national adult dress standardization by the use of the conditional distribution theory. Stand. Bull. **4**, 9–19 (1977)
37. Fang, K.T., et al.: (under the team name): The Analysis of Variance. Science Press, Beijing (in Chinese) (1977)
38. Fang, K.T., et al.: (under the team name): The Analysis of Variance. Science Press, Beijing (in Chinese) (1981)
39. Fang, K.T., Fan, J.Q.: Large sample properties for distributions with rotational symmetries. Northeastern Math. J. **4**, 379–388 (1988)
40. Fang, K.T., Fan, J.Q., Xu, J.L.: The distributions of quadratic forms of random matrix and applications. Chin. J. Appl. Prob. Stat. **3**, 289–297 (1987)

41. Fang, K.T., Fan, J.Q., Quan, H., Xiang, J.T.: Statistical analysis for directional data (I–VI). Appl. Stat. Manag. **8** (1), 59–61; (2) 58–65; (3) 57–64; (4) 56–65; (5) 56–64; (6) 57–64 (1989); **9** (1) 56–64; (2) 59–65 (1990)
42. Fang, K.T., Fang, B.Q: A new family of multivariate exponential distributions. Kexue Tongbao, **31**, 1510–1511 (1986)
43. Fang, K.T., Fang, B.Q.: Some families of multivariate symmetric distributions related to exponential distribution. J. Multivar. Anal. **24**, 109–122 (1988)
44. Fang, B.Q., Fang, K.T.: Distributions of order statistics of multivariate l_1-norm symmetric distribution and Applications. Chin. J. Appl. Prob. Stat. **4**, 44–52 (1988)
45. Fang, K.T., Fang, B.Q.: Families of Exponential matrix distributions. Northeast. Math. J. **4**, 16–28 (1988)
46. Fang, B.Q., Fang, K.T.: Maximum likelihood estimates and likelihood ratio criteria for location and scale parameters of the multivariate l_1-norm symmetric distributions. Acta Math. Appl. Sinica (English Series) **4**, 13–22 (1988)
47. Fang, K.T., Fang, B.Q.: A class of generalized symmetric Dirichlet distributions. Acta Math. Appl. Sinica (English Ser.) **4**, 316–322 (1988)
48. Fang, K.T., Fang, B.Q.: A characterization of multivariate l_1-norm symmetric distributions. Stat. Prob. Lett. **7**, 297–299 (1989)
49. Fang, K.T., Fang, H.B., von Rosen, D.: A family of bivariate distributions with non-elliptical contours. Commun. Stat.: Theory Methods **29**, 1885–1898 (2000)
50. Fang, H.B., Fang, K.T., Kotz, S.: The meta-elliptical distributions with given marginals. J. Multivar. Anal. **82**, 1–16 (2002)
51. Fang, K.T., Hickernell, F.J., Winker, P.: Some global optimization algorithms in statistics. In: Du, D.Z., Zhang, X.S., Cheng, K. (eds.) Lecture Notes in Operations Research, pp. 14–24. World Publishing Corporation (1996)
52. Fang, K.T., Kotz, S., Ng, K.W.: Symmetric Multivariate and Related Distributions. Chapman and Hall Ltd., London and New York (1990)
53. Fang, K.T., Kotz, S., Ng, K.W.: On the L_1-norm distributions. In: Dodge, Y. (ed.) L_1-Statistical Analysis and Related Methods, pp. 401–413. Elsevier Science Publishers, North Holland, Amsterdam (1992)
54. Fang, K.T., Li, R., Liang, J.: A multivariate version of Ghosh's T_3-plot to detect non-multinormality. Comput. Stat. Data Anal. **28**, 371–386 (1998)
55. Fang, K.T., Liang, J.: Inequalities for the partial sums of elliptical order statistics related to genetic selection. Can. J. Stat. **17**, 439–446 (1989)
56. Fang, K.T., Liang, J.: Tests of Spherical and elliptical symmetry. In: Johnson, N.L., Kotz, S. (eds.) Encyclopedia of Statistical Sciences, Update, vol. 3, pp. 686–691. Wiley, New York (1999)
57. Fang, K.T., Liang, J., Hickernell, F.J., Li, R.: A stabilized uniform Q-Q plot to detect non-multinormality. In: Hsiung, A.C., Ying, Z., Zhang, C.H. (eds.) Random Walk, Sequential Analysis and Related Topics, pp. 254–268. World Scientific, New Jersey (2007)
58. Fang, K.T., Liu, Z.X.: (under the team name): Orthogonal experimental design. Nonferrous Metal **8**, 39–56 (1974)
59. Fang, K.T., Liu, C.W.: The use of range in analysis of variance. Math. Pract. Theory **1**, 37–51 (1976)
60. Fang, K.T., Ma, F.S.: Splitting in cluster analysis and its applications. Acta Mathematicae Applicatae Sinica **5**, 339–534 (1982)
61. Fang, K.T., Niedzwiecki, D.: A unified approach to distributions in restricted occupancy problem. In: Sen, P.K. (ed.) Contributions to Statistics, Essays in Honour of Professor Norman Lloyd Johnson, pp. 147–158. North Holland Publishing Company (1983)
62. Fang, K.T., Pan, E.P.: Clustering Analysis. Geological Publishing House, Beijing (1982). (in Chinese)
63. Fang, K.T., Quan, H., Chen, Q.Y.: Applied Regression Analysis. Science Press, Beijing (1988). (in Chinese)

64. Fang, K.T., Sun, S.G.: Discriminant analysis by distance. Acta Mathematics Applicatae Sinica **5**, 145–154 (1982)
65. Fang, K.T., Wang, Y.: A sequential algorithm for optimization and its applications to regression analysis. In: Yang, L., Wang, Y. (eds.) Lecture Notes in Contemporary Mathematics, pp. 17–28. Science Press, Beijing (1990)
66. Fang, K.T., Wang, Y.: A sequential algorithm for solving a system of nonlinear equations. J. Comput. Math. **9**, 9–16 (1991)
67. Fang, K.T., Wang, Y., Bentler, P.M.: Some applications of number-theoretic methods in statistics. Stat. Sci. **9**, 416–428 (1994)
68. Fang, K.T., Wu, C.Y.: Mathematical Statistics and Standardization. Technical Standardization Press, Beijing (in Chinese) (1981)
69. Fang, K.T., Wu, Y.H.: Distributions of quadratic forms and generalized Cochran's Theorem. Math. Econ. **1**, 29–48 (1984)
70. Fang, K.T., Xing, K.F., Liu, G.Y.: Precision of Test Methods Determination. Chinese Standardization Press, Beijing (1988). (in Chinese)
71. Fang, K.T., Xu, J.L., Teng, C.Y.: Likelihood ratio criteria testing hypotheses about parameters of a class of elliptically contoured distributions. Northeastern Math. J. **4**, 241–252 (1988)
72. Fang, K.T., Xu, J.L.: The Mills' ratio of multivariate normal distributions and spherical distributions. Acta Mathematicae Sinica **30**, 211–220 (1987)
73. Fang, K.T., Xu, J.L.: A class of multivariate distributions including the multivariate logistic. J. Math. Res. Exposition **9**, 91–100 (1989)
74. Fang, K.T., Xu, J.L.: Likelihood ratio criteria testing hypotheses about parameters of elliptically contoured distributions. Math. Econ. **2**, 1–9 (1985)
75. Fang, K.T., Xu, J.L.: Statistical Distributions. Science Press, Beijing (1987). (in Chinese)
76. Fang, K.T., Yang, Z.H., Kotz, S.: Generation of multivariate distributions by vertical density representation. Statistics **35**, 281–293 (2001)
77. Fang, K.T., Zhang, Y.T.: Generalized Multivariate Analysis. Science Press and Springer-Verlag, Beijing and Berlin (1990)
78. Fang, K.T., Zheng, Z.K.: A two-stage algorithm of numerical evaluation of integrals in number-theoretic methods. J. Comput. Math. **17**, 285–292 (1999)
79. Fang, K.T., Zhu, L.X., Bentler, P.M.: A necessary test of goodness of fit for sphericity. J. Multivar. Anal. **45**, 34–55 (1993)
80. Glimm, E., Läuter, J.: On the admissibility of stable spherical multivariate tests. J. Multivar. Anal. **86**, 254–265 (2003)
81. Johnson, N.L., Kotz, S.: Distributions in Statistics: Continuous Multivariate Distributions. Wiley (1972)
82. Kotz, S., Fang, K.T., Liang, J.: On multivariate vertical density representation and its application to random number generation. Statistics **30**, 163–180 (1997)
83. Kotz, S., Seeger, J.P.: A new approach to dependence in multivariate distributions. In: Dallglio, G., Kotz, S., and Salinetti, G. (eds.) Advance in Probability Distributions with Given Marginals, pp. 13–50. Kluwer Academic, Dordrecht (1991)
84. Kropf, S., Läuter, J., Kosea, D., von Rosen, D.: Comparison of exact parametric tests for high-dimensional data. Comput. Stat. Data Anal. **53**, 776–787 (2009)
85. Läuter, J.: Exact t and F tests for analyzing studies with multiple endpoints. Biometrics **52**, 964–970 (1996)
86. Läuter, J., Glimm, E., Kropf, S.: New multivariate tests for data with an inherent structure. Biom. J. **38**, 5–23 (1996)
87. Läuter, J., Glimm, E., Kropf, S.: Multivariate tests based on left-spherically distributed linear scores. Ann. Stat. **26**, 1972–1988 (1998)
88. Li, R.Z., Fang, K.T., Zhu, L.X.: Some Q-Q probability plots to test spherical and elliptical symmetry. J. Comput. Graph. Stat. **6**(4), 435–450 (1997)
89. Liang, J.: Exact F-tests for a class of elliptically contoured distributions. J. Adv. Stat. **1**, 212–217 (2016)

90. Liang, J.: A Generalized F-test for the mean of a class of elliptically contoured distributions. J. Adv. Stat. **2**, 10–15 (2017)
91. Liang, J., Fang, K.T.: Some applications of Läuter's technique in tests for spherical symmetry. Biom. J. **42**, 923–936 (2000)
92. Liang, J., Fang, K.T., Hickernell, F.J.: Some necessary uniform tests for sphericcal symmetry. Ann. Inst. Stat. Math. **60**, 679–696 (2008)
93. Liang, J., Li, R., Fang, H., Fang, K.T.: Testing multinormality based on low-dimensional projection. J. Stat. Plann. Infer. **86**, 129–141 (2000)
94. Liang, J., Ng, K.W.: A multivariate normal plot to detect non-normality. J. Comput. Graph. Stat. **18**, 52–72 (2009)
95. Liang, J., Ng, K.W., Tian, G.: A class of uniform tests for goodness-of-fit of the multivariate L_p-norm spherical distributions and the l_p-norm symmetric distributions. Ann. Inst. Stat. Math. **71**, 137–162 (2019)
96. Liang, J., Tang, M.L.: Generalized F-tests for the multivariate normal mean. Comput. Stat. Data Anal. **57**, 1177–1190 (2009)
97. Liang, J., Tang, M.L., Chan, P.S.: A generalized Shapiro-Wilk W Statistic for testing high-dimensional normality. Comput. Stat. Data Anal. **53**, 3883–3891 (2009)
98. Liang, J., Tang, M.L., Zhao, X.: Testing high-dimensional normality based on classical skewness and kurtosis with a possible small sample size. Commun. Stat.-Theory Methods **48**(23), 5719–5732 (2019)
99. Liu, C.W., Fang, K.T.: How to use statistical papers (I and II). Math. Pract. Theory **3**, 49–55; **4**, 55–61 (1976)
100. Liu, C.W., Fang, K.T.: Yates' algorithm and its application in 2^n-type orthogonal array. Math. Pract. Theory **3**, 9–18 (1977)
101. Liu, C.W., Dai, S.S., Fang, K.T.: Elements of Probability Papers. Science Press, Beijing (1980). (in Chinese)
102. Mardia, K.V.: Directional data analysis: an overview. J. Appl. Stat. **15**, 115–122 (1988)
103. Nelsen, R.B.: An Introduction to Copulas. Springer (2006)
104. Pan, J.X., Fang, K.T.: Multiple outlier detection in growth curve model with unstructured covariance matrix. Ann. Inst. Stat. Math. **47**, 137–153 (1995)
105. Pan, J.X., Fang, K.T.: Detecting influential observations in growth curve model with unstructured covariance. Comput. Stat. Data Anal. **22**, 71–87 (1996)
106. Pan, J.X., Fang, K.T.: Bayesian local influence in growth curve model with unstructured covariance. Biom. J. **41**, 641–658 (1999)
107. Pan, J.X., Fang, K.T.: Growth Curve Models and Statistical Diagnostics. Springer, New York (2002)
108. Pan, J.X., Fang, K.T., Liski, E.P.: Bayesian local influence in the growth curve model with Rao's simple covariance structure. J. Multivar. Anal. **58**, 55–81 (1996)
109. Pan, J.X., Fang, K.T., Rosen, D.V.: Local infuence assessment in the growth curve model with unstructured covariance. J. Stat. Plann. Infer. **62**, 263–278 (1997)
110. Pan, J.X., Fang, K.T., Rosen, D.V.: On the posterior distribution of the covariance matrix of the growth curve model. Stat. Prob. Lett. **38**, 33–39 (1998)
111. Pan, J.X., Fung, W.K., Fang, K.T.: Multiple outlier detection in multivariate data using projection pursuit techniques. J. Stat. Plann. Infer. **83**, 153–167 (2000)
112. Quan, H., Fang, K.T.: Unbiasedness of some testing hypotheses in elliptically contoured population. Acta Mathematicae Applicatae Sinica **10**, 215–234 (1987)
113. Quan, H., Fang, K.T., Teng, C.Y.: The applications of information function for spherical distributions. Northeast. Math. J. **5**, 27–32 (1989)
114. Rosen, D.V.: The growth curve model: a review. Commun. Stat.-Theor. Methods **20**, 2791–2822 (1991)
115. Rosen, V.D., Fang, H.B., Fang, K.T.: An extension of the complex normal distribution. In: Johnson, N.L., Balakrishnan, N. (eds.) Advances in the Theory and Practice of Statistics: A volume in Honor of Samuel Kotz, pp. 415–427. Wiley, New York (1997)

116. Shen, S.Y., Fang, K.T.: Neural computation on nonlinear regression analysis problems. Int. J. Math. Stat. Sci. **3**(2), 155–178 (1995)
117. Song, D., Gupta, A.K.: L_p-norm uniform distribution. Proc. Am. Math. Soc. **125**(2), 595–601 (1997)
118. Sun, S.G., Fang, K.T.: The test for additional information in multivariate analysis. Acta Mathematicae Applicatae Sinica **3**, 81–91 (1977)
119. Wang, Y., Fang, K.T.: A sequential number-theoretic methods for optimization and its applications in statistics. In: The Development of Statistics: Recent Contributions from China, pp. 139–156. Longman, London (1992)
120. Winker, P., Fang, K.T.: Randomness and quasi-Monte Carlo approaches: some remarks on fundamentals and applications in statistics and econometrics. Jahrbücher für Nationalökonomie und Statistics **218**, 215–228 (1999)
121. Yue, X., Ma, C.: Multivariate l_p-norm symmetric distributions. Stat. Prob. Lett. **24**, 281–288 (1995)
122. Zhang, H.C., Fang, K.T.: Some properties of left-spherical and right-spherical matrix distributions. Chin. J. Appl. Prob. Stat. **3**, 97–105 (1987)
123. Zhang, Y.T., Fang, K.T.: An Introduction to Multivariate Analysis. Science Press, Beijing (1982, 1999). (in Chinese)
124. Zhang, Y., Fang, K.T., Chen, H.F.: On matrix elliptically contoured distributions. Acta Math. Scientia **5**, 341–353 (1985)
125. Zhu, L.X., Fang, K.T., Bhatti, M.I., Bentler, P.M.: Testing Sphericity of a high-dimensional distribution based on bootstrap approximation. Pakistan J. Stat. **11**, 49–65 (1995)
126. Zhu, L.X., Fang, K.T., Li, R.Z.: A new approach for testing symmetry of a high-dimensional distribution. Bull. Hong Kong Math. Soc. **1**, 35–46 (1997)
127. Zhu, L.X., Fang, K.T., Zhang, J.T.: A projection NT-type test for spherical symmetry of a multivariate distribution. In: Tiit, E.M., Kollo, T., and Niemi, H. (eds.) New Trends in Probability and Statistics, Multivariate Statistics and Matrices in Statistics, vol. 3, pp. 109–122. VSP-TEV, Utrecht, The Netherlands (1995)

Chapter 2
The Contribution to Experimental Designs by Kai-Tai Fang

Min-Qian Liu, Dennis K. J. Lin, and Yongdao Zhou

Abstract Professor Kai-Tai Fang has a wide research interest including applications of number-theoretic methods in statistics, distribution theory, experimental design, multivariate analysis and data mining. This paper only focuses on his contribution to experimental design. He proposed the method of visualization analysis for orthogonal designs in 1970. Inspired by three big military projects in 1978, he cooperated with Prof. Yuan Wang and proposed a new type of design of computer experiments, uniform design by utilized the number-theoretic methods. The uniform design can be also regarded as a kind of fractional factorial design, supersaturated design and design of experiments with mixture. In the past decades, the theory and applications of uniform designs have been developed rapidly by Kai-Tai Fang and his collaborators. In 2008, together with Professor Yuan Wang, Kai-Tai Fang received the 2008 State Natural Science Award at the Second Level, the highest level award in this kind of State award in that year. This paper focuses on the contribution of Kai-Tai Fang to experimental designs such as uniform designs, orthogonal designs, supersaturated designs and computer experiments.

2.1 Introduction

During the early 1970s, researches from Peking University and the Institute of Mathematics, Chinese Academy of Sciences, attempted to promote and apply orthogonal design to the industrial sector. In 1972, Kai-Tai Fang had the opportunity to go to the Tsingtao Beer Factory and other factories. He supervised their engineers to apply the

M.-Q. Liu · Y. Zhou (✉)
School of Statistics and Data Science, Nankai University, Tianjin 300071, China
e-mail: ydzhou@nankai.edu.cn

M.-Q. Liu
e-mail: mqliu@nankai.edu.cn

D. K. J. Lin
Department of Statistics, The Pennsylvania State University, University Park, PA 16802, USA
e-mail: dkl5@psu.edu

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_2

orthogonal design to industrial experiments. During the consultancy process Kai-Tai Fang found that the engineers had difficulty to understand statistical methods, especially in calculating the ANOVA table without the help of computers or calculators in that time. Therefore, he realized the need for statisticians to simplify the complicated statistical theory and methods, and proposed the method of "Visualization Analysis" for analytical use on experimental data. Very soon this method was commonly used in the Mainland. He also suggested to use the range instead of sum of squares in ANOVA table, called as "the range analysis", see Fang and Liu [25]. The range analysis is simple to understand and easy to compute.

During the consultancy process Kai-Tai Fang met many case studies with multiple factors, large experimental domains and non-linear relationships between the response and factors. Some experiments can not reach the goal for several years. Faced with these complicated cases Kai-Tai Fang considered several issues: (1) the number of levels should be more than 2 (3–5 for example); (2) Considering all the possible factors in the first stage; (3) ranking importance of the factors and interactions for choosing recommended level-combination. By these considerations he helped the engineers to solve a number of complicated experiments. Kai-Tai Fang with his colleague Mr Liu summarized their experience into a Notes for giving lecture to engineers. Late, this Notes had been published in the journal, see Fang and Liu [24].

The most difficult problems Kai-Tai Fang met in 1978 can not be solved using the orthogonal designs. These problems gave a strong motivation for the establishment of the theory and method of uniform designs.

In summary, Kai-Tai Fang has authored and co-authored 25 monographs and textbooks, and published more than 300 papers, among which 5 monographs and more than 100 papers are on the research field of experimental designs. The purpose of this paper is to introduce Fang's contribution to uniform designs, orthogonal designs and supersaturated designs. The paper is organized as follows. Sections 2.2–2.5 introduce the contribution to uniform designs, orthogonal designs and supersaturated designs by Kai-Tai Fang, respectively. Some material is chosen from the paper "A Conversation with Kai-Tai Fang" by Loie et al. [50].

2.2 The Contribution to Uniform Designs

In 1978, Kai-Tai Fang took part in three major missile-related projects covering land, sea and aerospace. In these projects the true model between the response and factors can be numerical expressed by solving a system of differential equations. It needed a long computation time by a computer. It turned out the idea of computer experiments. Due to the Cultural Revolution there was no any information about the design of computer experiments from outside of China. Kai-Tai Fang and Yuan Wang considered to choose a certain number of experiments in the domain and find an approximate model to replace the true one. For example, one project needs a design with 6 factors some of which having at least 18 levels on a large experimental

domain. Since the experiment was quite expensive and the speed of computer was quite slow (one experiment in one day), they wanted a design with at most 50 runs. Again, it was highly challenging. It needed a new method that could approximate a complicated system by a simple method with required accuracy. The great challenge was a motivating force to Kai-Tai Fang.

Kai-Tai Fang collaborated with Prof. Yuan Wang and borrowed the idea of number-theoretic methods to put experimental points uniformly on the domain and proposed the uniform design after a three-month hard working. Applying the uniform design to one of the three projects, 31 runs were arranged for the 6 factors each having 31 levels, and a satisfactory result was achieved. This method made that it was possible to calculate an accurate answer in 0.00001 s with the required accuracy. Eventually, the three projects were successful and won several nationwide awards. Kai-Tai Fang and Prof. Wang published two papers for introducing the uniform design theory in Chinese and English [4, 60], respectively. The new type of experimental designs was proposed since then. It was both time- and cost-saving and provided a valuable alternative design in computer experiments as well as laboratory experiments [17, 18, 23, 38]. During the 1970s, especially just after the Cultural Revolution in China, many scholars in China were still adhering to the modeling of the traditional experimental designs for data analysis, however, Kai-Tai Fang used regression analysis for modelling. Although the uniform design approach was not quite supported by few scholars in the experimental design, but it was greatly welcomed by the engineers. Several years later, the method of uniform designs has being used extensively in the mainland. Not only was it used for military purposes, but also it was adopted by and for civilians.

The idea of uniform design was from the overall mean regression model and the number-theoretic methods (Quasi-Monte Carlo methods). However, the uniformity is a geometric concept, not a statistical criterion. How to set up a solid theory is a very difficult target. Kai-Tai Fang had a difficult time during 1990–1996 after he moved to Hong Kong Baptist University. In fact, 90% of his academic pursuits has focused on uniform design since 1993. The progress was slow at the beginning. After several years, his collaboration with several scholars led to the discovery of a breakthrough.

In the following, we introduce the contribution to uniform designs by Kai-Tai Fang in the aspects of uniformity measures, construction methods of uniform designs and the relationship among different types of designs. Recently, Fang et al. [26] published a monograph that introduces the theory of the uniform design in details, and collects recent development in this direction.

2.2.1 Uniformity Measures

Assume $y = f(\mathbf{x})$ be the true model of a system on a domain $\mathscr{X} = C^s = [0, 1]^s = [0, 1] \times \cdots \times [0, 1]$, where $\mathbf{x} = (x_1, \ldots, x_s)$ are variables/factors and y is response. Let $\mathscr{P} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a set of n design points on C^s. One important issue is to estimate the overall mean of $f(\mathbf{x})$, i.e., $E(y) = \int_{C^s} f(\mathbf{x})d\mathbf{x}$. A natural idea is to use

the sample mean of $\mathscr{P}$, $\bar{y}(\mathscr{P}) = \frac{1}{n}\sum_{i=1}^{n} y_i$ to estimate $E(y)$, where $y_i = f(\mathbf{x}_i)$, $i = 1, \ldots, n$. The difference between $E(y)$ and the sample mean $\bar{y}(\mathscr{P})$ has following upper bound

$$|\bar{y}(\mathscr{P}) - E(y)| \leq V(f)D^*(\mathscr{P}), \tag{2.1}$$

where $V(f)$ is the total variation of the function f in the sense of Hardy and Krause (see Hua and Wang [45]; Niederreiter [54]), and $D^*(\mathscr{P})$ is the star discrepancy of $\mathscr{P}$ proposed by Weyl [63], which does not depend on f. The inequality (2.1) is the famous Koksma-Hlawka Inequality in quasi-Monte Carlo methods, and it is tight in some cases. If $V(f)$ is bounded in the experimental domain, then one may choose $\mathscr{P}$ with n design points on C^s such that its star discrepancy $D^*(\mathscr{P})$ is as small as possible and we can minimize the upper bound of the difference in (2.1). Fang [4] and Wang and Fang [60] called a design to be a uniform design if it has the smallest star discrepancy in the design space.

However, the star discrepancy has some shortcomings. Kai-Tai pointed out that it is not invariant under rotation of the coordinates, and is not easy to compute. He discussed this problem with his colleague Prof. Fred J. Hickernell. Hickernell [42, 43] used the tool of reproducing kernel Hilbert space, to generalize the definition of discrepancy and proposed different types of discrepancies. Among them the wrap-around L_2-discrepancy (WD) and centered L_2-discrepancy (CD) are popularly used. Fang et al. [17] gave the following requirements for a reasonable measure of uniformity.

C_1 It is invariant under permuting factors and/or runs.
C_2 It is invariant under rotation of the coordinates.
C_3 It can measure not only uniformity of $\mathscr{P}$ over C^s, but also the projection uniformity of $\mathscr{P}$ over C^u, where u is a non-empty subset of $\{1, \ldots, s\}$.
C_4 There is some reasonable geometric meaning.
C_5 It is easy to compute.
C_6 It satisfies the Koksma-Hlawka-like inequality.
C_7 It is consistent with other criteria in experimental design.

It has been known that the star discrepancy satisfies C_1, C_3, C_4 and C_6 and that both the WD and CD satisfy the requirements $C_1 - C_7$. Later, Zhou et al. [70] considered the following two additional requirements for a uniformity measure.

C_8 Sensitivity on a shift for one or more dimensions.
C_9 Less curse of dimensionality.

Zhou et al. [70] also showed that CD does not satisfy the requirement C_9 and WD does not satisfy the requirement C_8. Then, they proposed another type of discrepancy, called mixture discrepancy (MD). The MD can satisfy $C_1 - C_9$, which means that the MD can overcome the shortcomings of WD and CD, and MD may be the more reasonable measure of uniformity.

In many physical or practical situations, it prefers to have an experimental domain with a finite number of levels. Then, it is requested to give some discrepancies for

experimental domain with finite candidates directly. Hickernell and Liu [44] and Fang et al. [20] proposed a discrepancy, called discrete discrepancy, which is also defined by a special kernel. Qin and Fang [57] further discussed the property of the discrete discrepancy and the construction methods of uniform designs. Besides, Zhou et al. [71] proposed the Lee discrepancy for finite numbers of levels. The discrete discrepancy is better for two-level designs and the Lee discrepancy can be used for multi-level designs.

It is known that a measure of uniformity plays a key role in the theory of uniform designs, Kai-Tai Fang, Fred J. Hickernell and their collaborators proposed different types of discrepancies, which greatly develop the theory of uniform designs. Based on those discrepancies, many relationships between uniform designs and other type of designs were shown by Kai-Tai Fang and his collaborators.

Given a type of discrepancies, a tight lower bound is useful for the construction of uniform designs, since it can be served as a benchmark during the searching procedure. Kai-Tai Fang and his collaborators gave many lower bounds for different types of discrepancies, see [28, 32, 35, 37].

2.2.2 Construction Methods of Uniform Designs

For the convenient use of uniform designs in practice, uniform design tables are very useful. Kai-Tai Fang and his collaborators Mingyao Ai, Gennian Ge, Fred J. Hickernell, Runze Li, Min-Qian Liu, Xuan Lu, Chang-Xing Ma, Jianhui Ning, Jianxin Pan, Hong Qin, Yu Tang, Yuan Wang, Xiaoqun Wang, Peter Winker, Aijun Zhang, Yongdao Zhou, etc., gave many construction methods, which include the following three approaches: (i) Quasi-Monte Carlo methods [4, 16, 73]; (ii) Combinatorial methods [9, 11–13, 13]; (iii) Numerical search [35, 64, 65, 68, 69].

The Quasi-Monte Carlo methods are popularly used to construct uniform designs, since the first group of uniform designs were generated from the number-theoretic methods. Among them, the good lattice point (glp) method and the glp method with power generator are firstly used by Fang [4]. The main idea of glp method for constructing an n-point s-factor design is to find a generator vector $(h_1, \ldots, h_s)$, where h_i is coprime with n and is $h_1, \ldots, h_s$ are different with each other. Then, the ith run of a glp set is determined by $d_{ij} = ih_j \pmod{n}$, which means a glp set is fully determined by the generator vector. One may find a best generator vector under some uniformity criterion. Moreover, given the parameters including the number of runs n and the number of factors s the uniformity of the design constructed by the glp method may have some space to improve. For example, based on a glp set, [73] showed that the linear level permutation technique can improve the space-filling property under the uniformity criterion and maximin distance criterion.

From 2000, Kai-Tai Fang began the collaboration with Gennian Ge from Suzhou University and Min-Qian Liu from Nankai University to link up combinatorial designs and uniform designs. Combinatorial construction methods are powerful to construct uniform designs under the discrete discrepancy, i.e., the resulting designs

by those methods reach the minimum values of discrete discrepancy in many cases. The main tool of the combinatorial methods is the equivalence between an asymmetrical uniform designs with constant number of coincidences between any two rows and a uniformly resolvable design (URD). Therefore, given a URD, we can obtain a uniform design without any computational search. There are some miscellaneous known results on the existence of URDs, readers can refer to [11, 13] and the references therein for these results. The combinatorial methods can construct symmetric and asymmetric uniform designs, as well as supersaturated uniform designs. Some proposed construction methods by Kai-Tai Fang and his collaborators employed the following tools.

(A) Resolvable balanced incomplete block designs [9, 12, 13]
(B) Room squares [8]
(C) Resolvable packing designs [10, 27]
(D) Large sets of Kirkman triple systems [10]
(E) Super-simple resolvable t-designs [14]
(F) Resolvable group divisible designs [11]
(G) Latin squares [34]
(H) Resolvable partially pairwise balanced designs [36]

Here, (A)–(E) introduced the approaches for constructing symmetrical uniform designs, and (F)–(H) for asymmetrical cases. Most of those construction methods can obtain uniform designs under the discrete discrepancy.

The combinatorial methods only work for some special parameters n, s and $q_1, \ldots, q_s$. It is worth to give some construction methods of uniform designs for any given parameters. Kai-Tai Fang invited Peter Winker from Germany, a doctoral student then and a professor now, to cooperate for the numerical searching methods, which can satisfy such a requirement. Peter Winker is one of the experts on the threshold-accepting (TA) method. Winker and Fang [64] applied the TA for calculation of the star discrepancy and Winker and Fang [65] applied the TA for numerical searching uniform designs. This method uses the hard thresholds to accept the new solution in the neighborhood of current solution rather than some probability to accept the new solution in the simulation annealing method. Fang and Ma [29] and Fang et al. [31] used the TA algorithm to find uniform design tables under the WD and CD, respectively. Fang et al. [28] reexpressed the formulas of the WD and CD as functions of column balance, and also as functions of Hamming distances of the rows. And they also developed an efficient updating procedure for the local search heuristic threshold accepting based on these formulations of the WD and CD. Later, Fang et al. [35] proposed an efficient balance-pursuit heuristic algorithm to find many new uniform designs, especially with high levels. It was seen that the new algorithm is more powerful than the existing traditional threshold accepting algorithm. Fang et al. [32] also used the balance-pursuit heuristic algorithm to obtain many uniform designs. This algorithm uses some combinatorial properties of inner structures required for a uniform design. Moreover, Fang et al. [15] constructed uniform designs via an adjusted threshold accepting algorithm under the mixture discrepancy.

Later, Zhou et al. [69] reformed the optimization method for searching uniform designs into a zero-one quadratic integer program problem, and used some local searching methods to obtain the solution of such a problem, as well as the corresponding uniform design. Moreover, Fang et al. [23] found that many orthogonal designs can be generated by TA under the CD. Their results imply the so called "uniformly orthogonal design" by Fang and Ma [29], "Uniform fractional factorial designs" by Tang et al. [59].

2.3 More About Uniform Designs

In this section, more aspects of uniform designs are shown. We will show the contribution of Kai-Tai Fang on the topic of the connection between uniform designs and other types of designs, uniform designs for experiments with mixture and the application of uniform designs.

2.3.1 Connection Between Uniform Designs and Other Types of Designs

The uniform design theory was first proposed from Quasi-Monte Carlo method, and it is a deterministic method. It seems that the uniform design theory is totally different with orthogonal designs which have much statistical meaningfulness. Based on many research results of uniform designs, Kai-Tai Fang came up with the conjecture that most orthogonal designs are uniform in a certain sense. If this conjecture is true, we could link up orthogonal design with uniform design and obtain a vast development potential for uniform designs.

Kai-Tai Fang collaborated with several scholars and led to the discovery of a breakthrough. First, Kai-Tai Fang and Peter Winker found that such a conjecture was true in many cases, i.e., many existing orthogonal designs are also uniform designs. The result is based on the measure of uniformity proposed by Fang's colleague, Fred J. Hickernell. This discovery was of mutual benefit to both Hickernell and Fang. For Hickernell, his proposed measure of uniformity was initially not appreciated by many researchers in Quasi-Monte Carlo field but his measure became important in theory of uniform designs. For Fang, the measure of uniformity helped to prove that many existing orthogonal designs are uniform designs.

It still had one step to complete the proof of such a conjecture, i.e., we need a mathematical proof. Then, Kai-Tai Fang invited Rahul Mukerjee, Professor of the Indian Institute of Management in Calcutta, to HKBU for the collaboration in this topic. Rahul is a worldwide expert in the filed of experimental design. After two weeks, Rahul told Kai-Tai that the conjecture is not always true, even for a two-level factorial case. However, he showed an excellent result that it exists some relation-

ship between uniformity and orthogonality. Usually, the wordlength pattern and the criterion "minimum aberration" are popularly used to measure the orthogonality of a regular design, and the CD can be used to assess the uniformity of a design. Kai-Tai and Rahul established an analytic relationship between the CD and wordlength pattern for regular designs. This discovery was immediately published in a top statistical journal, Biometrika, see Fang and Mukerjee [33]. It opened up an entirely new area that linked up uniform design and factorial design, an area in which Kai-Tai Fang collaborated with Chang-Xing Ma and others, and published more than 20 papers during 1999–2004. For example, Ma et al. [53] showed that the equivalence between the uniformity and orthogonality is only true in some special cases.

Tang et al. [59] gave the relationship between the CD and the generalized wordlength pattern of a three-level fractional factorial design, and also showed that minimum aberration designs have low discrepancies on average. Later, Zhou and Xu [72] obtained the close relationship between any discrepancy defined by the tool of reproduced kernel Hilbert space and the generalized wordlength pattern, which can measure the orthogonality of a nonregular design.

Moreover, Zhang et al. [67] used the majorization framework to generalize and unify classical criteria for comparisons of balanced lattice designs, which include fractional factorial designs, supersaturated designs and uniform designs. Fang and Ma [30] showed the relationship between uniformity, aberration and correlation in regular fractions 3^{s-1}. Furthermore, Ma et al. [52] used the CD to efficiently detect the isomorphism of fractional factorial designs.

Furthermore, the blocking design is an important type of experimental designs. Blocking experiments emphasize the balance among blocks, treatments or groups. Such a balance is easy to intuitively understand, and has a simple formula in data analysis. However, it needs to be proven in theory. Under the guide of Prof. Kai-Tai Fang, Liu and Chan [46] used the discrete discrepancy to prove that balanced incomplete blocking designs are the most uniform ones among all binary incomplete block designs. Liu and Fang [47] considered a certain kind of resolvable incomplete blocking designs, obtained a sufficient and necessary condition for such a blocking design is the most uniform in the sense of a discrete discrepancy measure, proposed a construction method for such designs via a kind of U-type designs, and set up an important bridge between resolvable incomplete blocking designs and U-type designs.

2.3.2 Uniform Designs for Experiments with Mixture

Usually, the experimental domain of uniform designs is a hypercube. Kai-Kai Fang and Yuan Wang firstly considered uniform designs for experiments with mixture [38, 61], i.e., the experimental domain becomes a simplex. Later, Fang and Yang [39] discussed uniform designs of experiments with restricted mixtures.

For constructing uniform design of experiments with mixtures, the uniformity criterion should be given first. There are two types of uniformity criteria, indirect

and direct methods. One indirect method for measuring the uniformity of designs with mixtures is to measure the uniformity of the corresponding design on the hypercube C^{s-1} by a special transformation, see the F-discrepancy in Fang and Wang [38]. Ning et al. [56] proposed another uniformity criterion, DM2-discrepancy, for direct measuring the uniformity of designs with mixtures. Ning et al. [55] gave some construction method for the uniform designs with mixture on simplex.

2.3.3 Application of Uniform Designs

The achieved breakthrough in relation to uniform designs by Kai-Tai Fang and his collaborators won an international recognition. For example, the Handbook of Statistics (Volume 22) included the topic of uniform designs as a chapter, see Fang and Lin [19]. The Encyclopedia of Statistics Science (Second Edition) had chosen the aspect of uniform design as an entry, see Fang [5]. Both the Handbook of Engineering Statistics and the International Encyclopedia of Statistical Science by Springer also invited Kai-Tai Fang to write a chapter on uniform design for engineers, see Fang and Chan [7] and Fang [6], respectively. Moreover, Encyclopedia on Statistics in Quality and Reliability also invited Kai-Tai Fang to introduce the topic of uniform experimental designs, see Fang and Hickernell [3]. Uniform designs also won national acclaim. The Uniform Design Association of China (UDAC), as a branch of the Chinese Mathematical Society, was founded in 1994. The UDAC organized the national conferences, training courses, workshops and other activities to meet the calls to promote the applications of uniform designs.

In application-wise, there were numerous successful applications of uniform designs in China. With the keyword "uniform design", you can find thousands of published case studies from the academic database China national knowledge infrastructure (CNKI), which collects most of the important academic journals in China. The application of uniform designs by Ford Motor Co. Ltd in USA is an exemplary application of this method. In Ford, Dr. Agus Sudjianto introduced to Kai-Tai Fang that the technique had become a critical enabler for them to execute "Design for Six Sigma" to support the new product development, in particular, for the automotive engine design. Moreover, it was told that computer experiments using uniform designs have become the standard practices at Ford Motor Co. Ltd to support the early stage of the production design before the availability of the hardware. As a result, Fang et al. [17] published a textbook/monograph entitled "Design and Modeling for Computer Experiments", in which many case studies were from the real cases in Ford Motor Co. Ltd. In 2001 the 50th Gordon Research Conference: the Statistics in Chemistry & Chemical Engineering invited the topic "Uniform design for simulation experiments" as one of the nine topics, and each topic was given 3.5 h for introduction and discussion. Kai-Tai Fang, Professors Dennis K. J. Lin and Yizhen Liang (a chemist) formed a panel for this topic.

From the website of CNKI, there are 5660 papers used uniform designs to solve their problems between the period 2000–2018, see Fig. 2.1. There are also more than

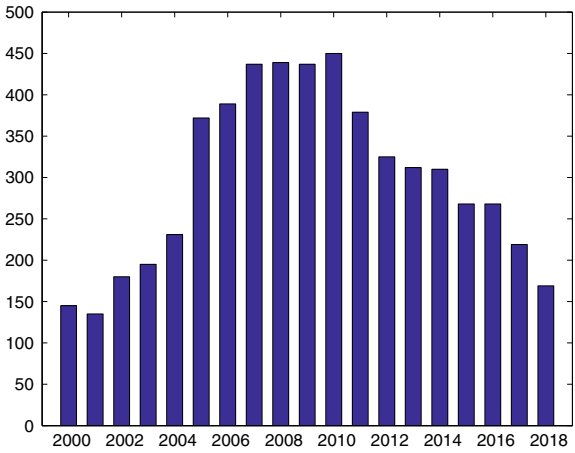


Fig. 2.1 The number of publications with the topic of uniform designs in CNKI

2000 citations of uniform designs from ISI Web of Science. Moreover, from the Google Scholar, the number of the citations of Kai-Tai Fang's publications is more than 14,000 times, and most of them are the citations of the papers about experimental designs, especially the topics of uniform designs.

2.4 The Contribution to Orthogonal Designs

During the process of promoting the common use of orthogonal designs, Kai-Tai Fang encountered quite a number of complicated multi-factor and non-linear issues. The engineers were unable to identify a satisfactory combination values of the parameters for a long time. An example was a porcelain insulator factory in Nanjing. The factory had a team whose job is to assign the conduction of the experiments continually for identifying a satisfactory combination values of the parameters. Although they achieved much knowledge in their experiments, they still failed to get a suitable combination of the values of the parameters to satisfy the requirement. At that time, the factory received a large number of orders for glass insulators but was unable to deliver the products. In view of the complexity of the issue, Kai-Tai Fang adhered to the principle of "big net catching big fish", and he conducted a 25-run experiment and arranged the six 5-level factors by an orthogonal design.

Such a design is a saturated design, which can not estimate all the main effects of the six factors, as well as none of the interaction effects can be estimated. However, in those 25 runs, all the responses of a special level-combination fulfill all the requirements. That was a great news to the factory in-charge. Should one liken the outcome to winning the US lottery or was it significant? In fact, using an orthogonal design to conduct 25 experiments actually represented 15,625 experiments, thus greatly increasing the probability of attaining an ideal technical/manufacturing condition.

The power of fractional factorial designs is that the experimental points have a good representation. Since then, Kai-Tai Fang used the same strategy to solve many of the "lasting, major and difficult" problems of the factories. This success encouraged Kai-Tai Fang to initiate the theory and method of uniform designs.

There are many criteria for assessing the property of orthogonal designs, such as minimum aberration [41], which is based on the wordlength pattern and can only be used for the comparison of regular designs. For extending such a criterion for nonregular designs, Kai-Tai Fang and Chang-Xing Ma used the MacWilliams identities to obtain the generalized wordlength pattern and the corresponding generalized minimum aberration criterion [51]. Independently, Xu and Wu [66] also obtained the generalized wordlength pattern by ANOVA models. The obtained generalized wordlength patterns by the two different ways are equivalent to each other for symmetrical nonregular designs. Additionally, the result in Xu and Wu [66] still works for asymmetrical designs. Later, Fang et al. [40] gave an effective algorithm for generation of factorial designs with generalized minimum aberration.

Moreover, Kai-Tai Fang cooperated with Lingyou Chan and Peter Winker to consider the relationship between orthogonal designs and optimal designs. They verified that each orthogonal array is an optimal design for a special polynomial regression models, see Chan et al. [1]. Liu et al. [49] showed the connections among different criteria for asymmetrical fractional factorial designs. Fang et al. [22] provided a theoretical justification for the optimal foldover plans for two-level designs, including the regular 2^{s-p}, nonregular, saturated and supersaturated designs.

2.5 The Contribution to Supersaturated Designs

A supersaturated design is essentially a fractional factorial design in which the number of potential effects is greater than the number of runs. A supersaturated design can be firstly used to screen the important factors in an experiment. Cooperated with Dennis K.J. Lin and Min-Qian Liu, Kai-Tai Fang gave a new criterion, $E(f_{NOD})$-criterion, for comparing supersaturated designs from the viewpoint of orthogonality and uniformity, see Fang et al. [20]. They also showed that the $E(f_{NOD})$-criterion is the generalization of the popularly used $E(s^2)$ and $ave\chi^2$ criteria for two- and three-level supersaturated designs, respectively. Moreover, Kai-Tai Fang also gave other criteria for assessing supersaturated designs such as $Ave(|f|)$, $Ave(f^2)$ and f_{max}, see Fang et al. [21].

Based on those criteria, Kai-Tai Fang and his collaborators gave many construction methods for multi-level and mixed-level supersaturated designs and investigated the properties of the obtained designs. The construction methods include the fractions of saturated orthogonal arrays (FSOA) method, the cyclic construction method, collapsing a U-type uniform design to an orthogonal array, and the global optimization algorithm, the threshold accepting algorithm, and the aforementioned combinatorial methods, see [2, 8, 10, 13, 14, 20, 21, 58]. Those results have high citations according to the ISI web of science. Moreover, Liu and Fang [48] used a uniform

mixed-level supersaturated design to study a case in computer experiments, and explored the efficiency of supersaturated designs for screening important factors and building the predictors.

2.6 Conclusion

Kai-Tai Fang's contribution in the field of experimental designs includes the theoretical development and practical application of orthogonal designs, uniform designs and supersaturated designs. Moreover, he also has some contribution on other types of designs. For example, he showed that the optimal representative point method via quantizer is superior to using other methods (including orthogonal array) to design outer array points in Taguchi's product-array designs, see [62]. In a word, among his contributions, the most important one of Kai-Tai Fang is that he first proposed the uniform design with Yuan Wang. Uniform design becomes an important type of experimental designs which has great theoretical significance and application value. The uniform experimental design can be regarded as a fractional factorial design with model uncertainty, a space-filling design for computer experiments, a robust design against the model specification, a supersaturated design and can be applied to experiments with mixtures. Moreover, in the era of big data, experimental designs will also play an important role for the analysis of big data. Uniform designs also have such a chance to be used for dealing with their problem. For example, one can use uniform designs for the subsampling of big data.

Acknowledgments The three authors would like to express their sincere thanks for the help from Prof. Kai-Tai Fang. The cooperation among the authors and Kai-Tai is very effectively and pleasantly. This work was supported by the National Natural Science Foundation of China (Grant Nos. 11771220 and 11871288), Natural Science Foundation of Tianjin (19JCZDJC31100), National Ten-Thousand Talents Program, Tianjin Development Program for Innovation and Entrepreneurship, and the Tianjin "131" Talents Program.

References

1. Chan, L.Y., Fang, K.T., Winker, P.: An equivalence theorem for orthogonality and optimality. Technical report math-186, Hong Kong Baptist University (1998)
2. Chen, J., Liu, M.Q., Fang, K.T., Zhang, D.: A cyclic construction of saturated and supersaturated designs. J. Statist. Plan. Infer. **143**, 2121–2127 (2013)
3. Fang, K., Hickernell, F.J.: Uniform experimental design. In: Encyclopedia on Statistics in Quality and Reliability, vol. 4, pp. 2037–2040 (2008)
4. Fang, K.T.: The uniform design: application of number-theoretic methods in experimental design. Acta Math. Appl. Sinica **3**, 363–372 (1980)
5. Fang, K.T.: Uniform Designs. Encyclopedia of Statistics, pp. 8841–8850. Wiley, New York (2006)
6. Fang, K.T.: Uniform experimental design. In: Lovric, M. (ed.) International Encyclopedia of Statistical Science Springer (2011)

7. Fang, K.T., Chan, L.Y.: Uniform design and its industrial applications. In: Pham, H. (ed.) Springer Handbook of Engineering Statistics, pp. 229–247 (2006)
8. Fang, K.T., Ge, G.N., Liu, M.Q.: Construction of $E(f_{NOD})$-optimal supersaturated designs via Room squares. Calcutta Statistical Association Bulletin **52**, 71–84 (2002a)
9. Fang, K.T., Ge, G.N., Liu, M.Q.: Uniform supersaturated design and its construction. Sci. China Ser. A **45**, 1080–1088 (2002b)
10. Fang, K.T., Ge, G.N., Liu, M.Q.: Construction of optimal supersaturated designs by the packing method. Sci. China Ser. A **47**, 128–143 (2004)
11. Fang, K.T., Ge, G.N., Liu, M.Q., Qin, H.: Optimal supersaturated designs and their constructions, Technical report MATH-309, Hong Kong Baptist University (2001)
12. Fang, K.T., Ge, G.N., Liu, M.Q., Qin, H.: Construction on minimum generalized aberration designs. Metrika **57**, 37–50 (2003)
13. Fang, K.T., Ge, G.N., Liu, M.Q., Qin, H.: Combinatorial constructions for optimal supersaturated designs. Discrete Math. **279**, 191–202 (2004a)
14. Fang, K.T., Ge, G.N., Liu, M.Q., Qin, H.: Construction of uniform designs via super-simple resolvable t-designs. Utilitas Mathematica **66**, 15–32 (2004b)
15. Fang, K.T., Ke, X., Elsawah, A.M.: Construction of uniform designs via an adjusted threshold accepting algorithm. J. Complex. **43**, 28–37 (2017)
16. Fang, K.T., Li, J.K.: Some new results on uniform design. Chin. Sci. Bull. **21** (1921–1924). English version in **40**, 268–272 (1995)
17. Fang, K.T., Li, R., Sudjianto, A.: Design and Modeling for Computer Experiments. Chapman and Hall/CRC, New York (2006)
18. Fang, K.T., Lin, D.K.J.: Theory and applications of the uniform design. J. Chin. Stat. Assoc. **38**(4), 331–351 (2000) (in Chinese)
19. Fang, K.T., Lin, D.K.J.: Uniform designs and their application in industry. In: Rao, C., Khattree, R. (eds.) Handbook on Statistics 22: Statistics in Industry, pp. 131–170 (2003)
20. Fang, K.T., Lin, D.K.J., Liu, M.Q.: Optimal mixed-level supersaturated design. Metrika **58**, 279–291 (2003)
21. Fang, K.T., Lin, D.K.J., Ma, C.X.: On the construction of multi-level supersaturated designs. J. Statist. Plann. Inference **86**, 239–252 (2000)
22. Fang, K.T., Lin, D.K.J., Qin, H.: A note on optimal foldover design. Statist. Probab. Lett. **62**, 245–250 (2003)
23. Fang, K.T., Lin, D.K.J., Winker, P., Zhang, Y.: Uniform design: theory and applications. Technometrics **42**, 237–248 (2000)
24. Fang, K.T., Liu, Z.X.: Design for orthogonal designs. Non-ferrous Metal **8**, 39–56 (1974) (in Chinese)
25. Fang, K.T., Liu, C.W.: The use of range in analysis of variance. Math. Pract. Theory **1**, 37–51 (1976)
26. Fang, K.T., Liu, M.Q., Qin, H., Zhou, Y.: Theory and Application of Uniform Experimental Designs. Springer, Singapore (2018)
27. Fang, K.T., Lu, X., Tang, Y., Yin, J.: Constructions of uniform designs by using resolvable packings and coverings. Discrete Math. **274**, 25–40 (2004)
28. Fang, K.T., Lu, X., Winker, P.: Lower bounds for centered and wrap-around l_2-discrepancies and construction of uniform. J. Complex. **20**, 268–272 (2003)
29. Fang, K.T., Ma, C.X.: Wrap-around L_2-discrepancy of random sampling, Latin hypercube and uniform designs. J. Complex. **17**, 608–624 (2001)
30. Fang, K.T., Ma, C.X.: Relationship between uniformity, aberration and correlation in regular fractions 3^{s-1}. In: Fang, K.T., Hickernell, F.J., Niederreiter, H. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 2000, pp. 213–231. Springer (2002)
31. Fang, K.T., Ma, C.X., Winker, P.: Centered L_2-discrepancy of random sampling and Latin Hypercube design, and construction of uniform designs. Math. Comp. **71**, 275–296 (2002)
32. Fang, K.T., Maringer, D., Tang, Y., Winker, P.: Lower bounds and stochastic optimization algorithms for uniform designs with three or four levels. Math. Comp. **75**, 859–878 (2006)

33. Fang, K.T., Mukerjee, R.: A connection between uniformity and aberration in regular fractions of two-level factorials. Biometrika **87**, 193–198 (2000)
34. Fang, K.T., Shiu, W.C., Pan, J.X.: Uniform designs based on latin squares. Statist. Sinica **9**, 905–912 (1999)
35. Fang, K.T., Tang, Y., Yin, J.X.: Lower bounds for wrap-around l_2-discrepancy and constructions of symmetrical uniform designs. J. Complex. **21**, 757–771 (2005)
36. Fang, K.T., Tang, Y., Yin, J.X.: Resolvable partially pairwise balanced designs and their applications in computer experiments. Utilitas Mathematica **70**, 141–157 (2006)
37. Fang, K.T., Tang, Y., Yin, J.X.: Lower bounds of various criteria in experimental designs. J. Statist. Planing and Inferences **138**, 184–195 (2008)
38. Fang, K.T., Wang, Y.: Number-Theoretic Methods in Statistics. Chapman and Hall, London (1994)
39. Fang, K.T., Yang, Z.H.: On uniform design of experiments with restricted mixtures and generation of uniform distribution on some domains. Statist. Probab. Lett. **46**, 113–120 (2000)
40. Fang, K.T., Zhang, A., Li, R.: An effective algorithm for generation of factorial designs with generalized minimum aberration. J. Complex. **23**, 740–751 (2007)
41. Fries, A., Hunter, W.G.: Minimum aberration 2^{k-p} designs. Technometrics **22**, 601–608 (1980)
42. Hickernell, F.J.: A generalized discrepancy and quadrature error bound. Math. Comp. **67**, 299–322 (1998a)
43. Hickernell, F.J.: Lattice rules: how well do they measure up? In: Hellekalek, P., Larcher, G. (eds.) Random and Quasi-Random Point Sets, pp. 106–166. Springer, New York (1998b)
44. Hickernell, F.J., Liu, M.Q.: Uniform designs limit aliasing. Biometrika **89**, 893–904 (2002)
45. Hua, L.K., Wang, Y.: Applications of Number Theory to Numerical Analysis. Springer and Science Press, Berlin and Beijing (1981)
46. Liu, M.Q., Chan, L.Y.: Uniformity of incomplete block designs. Int. J. Mater. Prod. Technol. **20**, 143–149 (2004)
47. Liu, M.Q., Fang, K.T.: Some results on resolvable incomplete block designs. Sci. China Ser. A **48**, 503–512 (2005)
48. Liu, M.Q., Fang, K.T.: A case study in the application of supersaturated designs to computer experiments. Acta Math Sci **26B**, 595–602 (2006)
49. Liu, M.Q., Fang, K.T., Hickernell, F.J.: Connections among different criteria for asymmetrical fractional factorial designs. Statist. Sinica **16**, 1285–1297 (2006)
50. Loie, A.W.L., Li, L., Puntanen, S., Styan, G.P.H.: A conversation with Kai-Tai Fang. In: Souvenir Booklet of the 24th International Workshop on Matrices and Statistics, pp. 1–39 (2015)
51. Ma, C.X., Fang, K.T.: A note on generalized aberration in factorial designs. Metrika **53**, 85–93 (2001)
52. Ma, C.X., Fang, K.T., Lin, D.K.J.: On isomorphism of fractional factorial designs. J. Complex. **17**, 86–97 (2001)
53. Ma, C.X., Fang, K.T., Lin, D.K.J.: A note on uniformity and orthogonality. J. Statist. Plann. Inference **113**, 323–334 (2003)
54. Niederreiter, H.: Random number generation and Quasi-Monte Carlo methods. In: SIAM CBMS-NSF Regional Conference. Applied Mathematics, Philadelphia (1992)
55. Ning, J.H., Fang, K.T., Zhou, Y.D.: Uniform design for experiments with mixtures. Comm. Statist. Theory Methods **40**, 1734–1742 (2011)
56. Ning, J.H., Zhou, Y.D., Fang, K.T.: Discrepancy for uniform design of experiments with mixtures. J. Statist. Plann. Inference **141**, 1487–1496 (2011)
57. Qin, H., Fang, K.T.: Discrete discrepancy in factorial designs. Metrika **60**, 59–72H (2004)
58. Tang, Y., Ai, M., Ge, G., Fang, K.T.: Optimal mixed-level supersaturated designs and a new class of combinatorial designs. J. Statist. Plann. Inferences **137**, 2294–2301 (2007)
59. Tang, Y., Xu, H., Lin, D.K.L.: Uniform fractional factorial designs. Ann. Statist. **40**, 891–907 (2012)
60. Wang, Y., Fang, K.T.: A note on uniform distribution and experimental design. Chin. Sci. Bull. **26**, 485–489 (1981)

61. Wang, Y., Fang, K.T.: Uniform design of experiments with mixtures. Sci. China Ser. A **39**, 264–275 (1996)
62. Wang, Y.J., Lin, D.K.J., Fang, K.T.: Designing outer array points. J. Q. Technol. **27**, 226–241 (1995)
63. Weyl, H.: *ü*ber die gleichverteilung der zahlem mod eins. Math. Ann. **77**, 313–352 (1916)
64. Winker, P., Fang, K.T.: Application of threshold accepting to the evaluation of the discrepancy of a set of points. SIAM Numer. Anal. **34**, 2038–2042 (1997)
65. Winker, P., Fang, K.T.: Optimal *u*-type design. In: Niederreiter, H., Zinterhof, P., Hellekalek, P. (eds.) Monte Carlo and Quasi-Monte Carlo Methods 1996, pp. 436–448. Springer, Berlin (1998)
66. Xu, H.Q., Wu, C.F.J.: Generalized minimum aberration for asymmetrical fractional factorial designs. Ann. Statist. **29**, 1066–1077 (2001)
67. Zhang, A., Fang, K.T., Li, R., Sudjianto, A.: Majorization framework for balanced lattice designs. Ann. Statist. **33**, 2837–2853 (2005)
68. Zhou, Y.D., Fang, K.T.: An efficient method for constructing uniform designs with large size. Comput. Stat. **28**(3), 1319–1331 (2013)
69. Zhou, Y.D., Fang, K.T., Ning, J.H.: Constructing uniform designs: a heuristic integer programming method. J. Complex. **28**, 224–237 (2012)
70. Zhou, Y.D., Fang, K.T., Ning, J.H.: Mixture discrepancy for quasi-random point sets. J. Complex. **29**, 283–301 (2013)
71. Zhou, Y.D., Ning, J.H., Song, X.B.: Lee discrepancy and its applications in experimental designs. Statist. Probab. Lett. **78**, 1933–1942 (2008)
72. Zhou, Y.D., Xu, H.: Space-filling fractional factorial designs. J. Amer. Statist. Assoc. **109**, 1134–1144 (2014)
73. Zhou, Y.D., Xu, H.: Space-filling properties of good lattice point sets. Biometrika **102**, 959–966 (2015)

Chapter 3
From "Clothing Standard" to "Chemometrics"

Review of Prof. Kai-Tai Fang's Contributions in Data Mining

Ping He, Xiaoling Peng, and Qingsong Xu

Abstract This paper reviews Prof. Kai-Tai Fang's contributions in data mining. Since the 1970s, Prof. Fang has been committed to applying statistical ideas and methods to deal with large amounts of data in practical projects. By analyzing more than 400,000 pieces of data, he found representative clothing indicators and established the first adult clothing standard in China; through cleaning and modeling steelmaking data from steel mills all over the country, he revised the national standard for alloy structural steel; by studying various data in chemometrics, he introduced many new advanced statistical methods to improve the identification and classification of chemical components, established more effective models for the relationship between quantitative structure and activity, and promoted the application of the traditional Chinese medicine (TCM) fingerprint in TCM quality control. Professor Fang and his team's research achievements in data mining have been highly appreciated by relevant experts. This article is written to celebrate the 80th birthday of Prof. Kaitai Fang.

P. He (✉) · X. Peng
Beijing Normal University and Hong Kong Baptist University United International College, Zhuhai, China
e-mail: heping@uic.edu.hk

X. Peng
e-mail: xlpeng@uic.edu.hk

Q. Xu
Department of Mathematics and Statistics, Central South University, Changsha, China
e-mail: qsxu@csu.edu.cn

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_3

3.1 Introduction

With the rapid development of computer technology, data collection and storage in various industries has greatly improved, and the amount of data has increased dramatically. Various new statistical applications have emerged, showing their vital role in the development of all fields. In this revolutionary data-driven evolution, statistics, as the core foundation and technology of data science, has unprecedentedly developed in recent years. Many new theories and methods of statistics have been proposed to deal with new types of data that have broken through the scales and types of classical data.

In China, as the economy developed, so did the mutual integration of various disciplines and the rapid development of computer technology, which resulted in the government, enterprises and society continuously expanding and promoting the application of statistics. The technology of data mining receives increasing attention and is widely used in various fields. All of these are inseparable from the efforts and contributions of statisticians.

Professor Kai-Tai Fang has always been committed to research on data mining and promoting the application of statistics in practical fields. He, in cooperation with experts in other fields, undertook many practical projects, such as establishing the first Chinese adult clothing standards by analyzing body size measurements, assessing and optimizing the national standard for alloy structural steel through processing the steel-making data from all state-owned steel mills, identifying the composition and structure of the compounds via machine learning put into mass spectral databases, and improving the efficiency of traditional Chinese medicine (TCM) quality control by analyzing TCM fingerprints. As well as these accomplishments, Prof. Fang wrote multiple books. He together with Prof. Yuan Wang published *Number-Theoretic Methods in Statistics* [7]. The theories and methods introduced in the book belong to the intersection of number theory, statistics, and computer science and widely used in computer simulation experiment, agriculture, industry, medicine, and high-tech innovation. Also, the book *Introduction to Multivariate Statistical Analysis* [9] by Profs. Fang and Yaoting Zhang not only has been adopted as a textbook by many universities, but also has been recommended as one of the main reference books for those engaged in work related to statistics and data mining.

3.2 Establishment of the First Chinese Adult Clothing Standard

In the early 1970s, as the Chinese population grew, making clothing needed to be automated, controlled, and standardized for mass production to meet the growing demand. However, at that time, China did not have a national-level clothing standard to provide a reliable basis for the design specifications needed to produce garments. For this reason, Chinas Ministry of Textiles, Ministry of Light Industry and State

Administration of Standards launched a cooperative project to set the first national adult clothing standard.

A clothing standard is a formulation of a series of specifications based on the shape and size of the human body. To create a standard, peoples body types needed to be divided into several categories where the most representative measurements were calculated for each category. Then, clothes would be designed according to these specified parameters to meet the needs of most people. A good clothing standard could not contain too many specifications and also had to accommodate for a wide range of shapes so that people with ordinary bodies could easily buy clothes that fit. Crafting a clothing standard at the national level was an undertaking that required many experts.

Originally, the project members undertaking this monumental task were composed of senior tailors and relevant industry leaders from Beijing, Shanghai, Tianjin and other cities. They conducted surveys on the human body in more than ten provinces across the country. Using stratified sampling, the members selected more than 400,000 men and women for body measurements; men had twelve measurements and women had fourteen. However, they did not know how to set the clothing standard using the vast data, so they just made a preliminary analysis of the figures and then drafted a standard mainly based on their experience. Due to the lack of data-driven evidence and theoretical bases, this clothing standard was not adopted by the State Administration of Standards.

Researchers realized that mathematical and statistical methods should be used to analyze the data and thus provide a reliable basis for the creation of standards. In 1974, Prof. Kaitai Fang was invited to join the standards-setting research team. He found that principal component analysis (PCA), which was widely used in the world to develop the clothing standard, was not suitable for the development of the Chinese standard. The main reason was that the principle components in PCA were too difficult for Chinese clothing workers to understand and master at the time. Therefore, Prof. Fang developed a brand-new method to formulate easy-to-implement Chinese clothing standards [3].

Professor Fang introduced a statistical idea to develop the clothing standard: if a variable can well represent others, then given the variable, the conditional standard deviation of these other variables should be small, and vice versa. Using this idea and optimal design theory in multivariate analysis, He proposed a method to sequentially found the variables with the smallest generalized conditional variance. In the article [3], the data of adult womens clothing in Beijing were adopted as an example. The results showed that height, chest circumference and waist circumference were the most representative measurement sites among all the body measurements and could be used for further body type classification. Then, the most representative values were calculated in each category.

According to the classification of body type and the corresponding representative measurements, manufacturers could mass-produce garments in large batches. The developed clothing standard ensured that most people could buy ready-to-wear clothing and that only a small fraction of people with special body types would need tailored clothing. On behalf of the project team, Prof. Fang reported the Beijing

clothing standard formulated by his method to the three state offices (the Ministry of Textiles, the Ministry of Light Industry and the State Administration of Standards) where the standard received unanimous approval.

Later, numerous clothes were produced for most body types according to the project's results. The market feedback was excellent, and thus the method proposed by Prof. Fang to calculate the Chinese adult clothing standard was a success. After many years of development, the team calculated and formulated clothing standards for each of China's sub-regions. Their project "Series of Standards for Chinese Adult Heads" won the 1980 Science and Technology Achievement 3rd Prize Award issued by the Beijing Government; their project "Chinese Adult Clothing Standards" won the 1982 Special Prize of Light Industry of the People's Republic of China. On December 10th, 1988, the national standard GB10000-88 for the Chinese adult body size was officially released.

It is worth mentioning that, in 1982, Prof. Fang came up with a new method to further improve the clothing standard and theoretically defined the concept of representative points for statistical distributions. He also provided numerical algorithm of computing representative points for univariate normal distribution [6]. However, it was a pity that his theory was highly overlapping with that of Prof. Cox [2], a famous British statistician. Professor Fang wasn't discouraged and instead proposed NTLBG algorithm [10] based on number theory and k-means algorithm in 1994, which can obtain the representative points of elliptically contoured distributions. In 2014, he and his students used representative points and random sampling significantly improved the efficiency of the Monte Carlo method [16].

3.3 Revision of the National Standard for Alloy Structural Steel

Alloy structural steel is made metallurgically by adding elements chromium, manganese, nickel, molybdenum silicon and other elements to steel. During the manufacturing process, the contents of these elements must be controlled within a certain scope to ensure that the five mechanical properties, including characteristics such as strength and elasticity, of the produced steel to meet the requirements.

During the 1960s, there already was a national standard for the range that each elements contents should fall into. This standard was used by more than 10 factories all over the country to produce the same kind of alloy structure steel, however, the results were inconsistent. While some steel mills produced a high proportion of qualified alloy steel, other steel mills had low proportions of qualified alloy steel even if the content of all elements, such as carbon, chromium, manganese, and etc. were full compliance with the national standard: Qiqihar Steel Plant, for instance, produced qualified steel only 38% of the time. Thus, many steel mills entertained doubts about the national standard for the scope of contents of elements. This standard was introduced from the Soviet Union. At that time, no one knew its principles, and

the reasons for the inconsistent production qualified rate. Since refining a batch of low quality alloy structural steel is expensive, the manufacturers not meeting the national standard would suffer heavy economic losses. Therefore, it was important and urgent to study whether the standard being used was reasonable or whether there was room for improvement.

In 1973, Prof. Kai-Tai Fang and his colleagues took on a project from China’s Ministry of Metallurgy to review the national standard for the ranges of each element present when forging alloy structural steel. The volume of data collected in this project was enormous; it included relevant data from all state-owned steel mills. It took the team, composed of professors and engineers, half a year to preprocess the copious amounts of various data. For example, some steel mills used the proportion of steel elements to estimate whether the alloy steel was qualified based on experience, and if the estimate was disqualified, a different treatment would be carried out in the steels quenching process to make it meet the standard. This special kind of treatment make these data need to be eliminated when preprocessing.

After cleaning the data, Prof. Fang and his colleagues built regression models, predicted the five mechanical properties of steel with its elements and did five-fold integral calculation before getting the qualified rate. Finally they found that: the national standard is scientific and reasonable; the combination ratio of elements in steel will affect mechanical properties’ rates of meeting the standard, therefore, in the steel making process, one should try to choose the combination which produces the highest qualified rates; and it is correct of some steel mills’ experience that whether the steel is qualified or not can be estimate by the element ratio [5].

From this project, Prof. Fang also came up with some new statistics theoretical problems. He refined the metallurgical problem into a extremum problem based on multi-dimensional normal probability density, and proved the necessary condition for the existence of solutions to the extremum problem in 1979 [8]. Also in this paper, he presented an effective algorithm for finding solutions to the extremum problems. While solving these problems, Prof. Fang faced difficulties in calculating the probability of a multidimensional normal distribution. He had to calculate multiple integrals without a computer’s aid since the technology at that time could not meet the demands of these problems. Using a suggestion from Prof. Yuan Wang, a famous mathematician, Prof. Fang employed the good lattice method (GLM) [4] and efficiently calculated the multiple integrals. This method also laid the foundation for Prof. Fang’s later invention: uniform experimental design.

Although data mining has become popular in the last three decades, Prof. Fang pioneered similar work in the 1970s, when domestic computer technology was still lagging. Through communicating extensively with experts in related field, he carefully preprocessed data, analyzed data, constructed model, tested model and finally obtained convincing conclusions. He not only successfully completed the actual project, but also refined the specific problems into general theory. The general problems were studied and discussed, and solutions that may arise in other practical work were proposed for similar problems.

Speaking of the standard, Prof. Fang also devoted himself in the promotion of standard of statistics. The International Organization for Standardization (ISO) set the

standard of ISO5725 to measure the testing precision of an instrument with repeatability and reproducibility. In order to introduce this standard to China, the Standard Administration of China established a specialized committee, in which Prof. Fang was designated as the chair. He explained the statistical theory of the standard to committee members in detail and spent more than two years completing the national standard GB/T6379 [14]. This project was awarded the second prize of Standardization Administration of China. This is a standard of statistics, which mainly includes model of variance analysis of random effect, the elimination of abnormal data and linear regression. Afterwards, Prof. Fang participated in international ISO5725 committee on behalf of China on many occasions.

3.4 Contributions to Chemometrics

Professor Kai-Tai Fang's contribution to chemometrics began by collaborating with Prof. Yizeng Liang, a celebrated analytical chemist from Central South University: Prof. Liang received the Chemometrics Lifetime Achievement Award at the XVI International Conference on Chemometrics in Analytical Chemistry. He has been committed to working on the application of statistical methods in analytical chemistry and has a profound understanding of statistical theories and methods. In 1995, during Prof. Liang's visit to the chemistry department of Hong Kong Baptist University (HKBU), he met Prof. Fang, who was then a professor in the mathematics department of HKBU.

When he learned that Prof. Fang was working on the application of number-theoretic methods in statistics, Prof. Liang was fascinated and studied Prof. Fang's book *Number-Theoretic Methods in Statistics* [7], which had just been published by Profs. Fang and Yuan Wang. In 1996, Profs. Liang and Fang published an article together in the academic journal *Analyst* [24] which proposed a robust multivariate calibration method based on the least median of squares (LMS) and the number-theoretic methods in optimization (SNTO) algorithm. Compared with the least squares method, the proposed method significantly reduces the computational complexity when the analysis system has two or three components. In addition, the method is more robust when there are more abnormal values.

The SNTO algorithm mentioned above is an optimization algorithm for multivariate functions proposed by Profs. Fang and Wang. It uses number theory to evenly distribute points in the search space, and gradually reduces the optimization search space by sequential compression to find the global optimal solution of multivariate functions [7]. In 1997, Profs. Fang and Liang further applied the SNTO algorithm to the constrained background bilinear decomposition model for the quantitative analysis of analytical samples containing unexpected interference [35]. In this paper, the SNTO algorithm was compared with another global optimization algorithm: variable step size generalized simulated annealing (VSGSA). The results showed that when the two methods achieve the same analytical accuracy, the SNTO algorithm is simpler, clearer and easier to implement, making it a practical tool in chemometrics.

In the same year, Prof. Liang advised his colleagues to adopt a sequential uniform design (SUD) procedure for the separation of five dithiocarbamate (DTC) compounds by capillary electrophoresis (CE). The CE technique was unable to separate these five DTC compounds when changing one variable at a time, whereas they were completely separated by using SUD method [23]. In addition, the SUD procedure was introduced as a promising candidate for experimental design in nonlinear multivariate calibration with ANN [36].

Regarding Prof. Fang’s contribution to chemometrics, the application of uniform experiment design in chemical experiments must be mentioned. In 1998, Prof. Liang and his Ph.D. student Qingsong Xu were invited by Prof. Fang to visit the mathematics department of HKBU. They noticed that Atkinson, a well-known statistician in optimal experimental design, recommended using D optimal design and T optimal design to estimate the kinetic rate in reversible chemical reactions [1]. Thus, Profs. Fang and Liang decided to compare the performance of orthogonal experimental design (OD), D optimal design (DOD) and uniform experimental design (UD) in reversible chemical reactions.

Their studies showed that for nonlinear reversible reaction kinetic models, DOD usually performs best if the initial value is not far from the true parameter and the random error is not large. It’s sensitivity to the choice of initial values is a drawback: if the initial value is far from the true parameter, then the parameter estimate is likely to fail. When compared with DOD, OD is less sensitive to the location of initial parameters, but as the random error increases to a certain level, OD faces a similar problem: if the initial value is far from the real parameter, the parameter estimation is also likely to fail. When there is no prior information about the location of real parameters and random noise intensity, UD always performs best among the three designs. The results were summarized in the article [30].

In 2001, Profs. Fang, Liang and Dr. Qingsong Xu published an article “Uniform design and its applications in chemistry and chemical engineering” [25]. The article has had a significant impact in chemistry and chemical engineering: currently, the SCI citation rate has reached 255. In the same year, Prof. Fang was invited by the 50th Gordon Research Conference titled “The Statistics in Chemistry and Chemical Engineering” to deliver a one hour lecture to introduce uniform design. Many chemists and chemical engineers have shown interest in UD and hope to develop an in-depth understanding of UD. Since then, Prof. Fang still receives letters from chemists, requesting to construct uniform experimental tables for their research.

Professors Fang and Liang are largely celebrated today for their work together in chemometrics. In 2016, Prof. Ruqin Yu, fellow of the Chinese Academy of Sciences and the former President of Hunan University, invited Prof. Fang to write an article for the special issue of the *Journal of Chemometrics* in China. In his letter to Dr. Qingsong Xu, Prof. Yu said:

> I am thinking about a problem: throughout the history of chemometrics in China, there are few Chinese original contributions, and most of them are applications of methods proposed by others. You and Prof. Liang worked with Prof. Fang to develop the application of uniform experimental design in chemistry. Uniform experimental design is an entirely original Chinese innovation. It must be stressed that, somehow in this special topic, Chinese scientists still have original innovations in chemometrics. The value of uniform experimental design itself is that it is an original innovation in mathematics.

In 2018, Prof. Fang and Dr. Xu published an article in the *Journal of Chemometrics* reviewing the development of uniform design in chemometrics and its various applications [32].

3.5 Research Group's Further Contributions to Chemometrics

After Profs. Kai-Tai Fang, Yizeng Liang, and Dr. Qingsong Xu published their 2001 article about uniform design in chemistry and chemical engineering, Profs. Fang and Liang further strengthened their cooperation in data mining in chemometrics. In 2002, they organized a series of research seminars where Prof. Fang's five Ph.D. students and Prof. Liang's six Ph.D. students attended all of them together. They also invited each other's Ph.D. students to attend the other school for at least a month to deepen their understanding of the other field, strengthen the discussion and exchange ideas. Together, the two groups have done a series of work involving many aspects of chemometrics.

The research group led by Profs. Fang and Liang worked on an important area in traditional chemometrics: the topological structure representation of organic compounds. In the article [20], various matrix representations, topological indices and atomic properties of topological structures were summarized and the shortcomings of topological indices were discussed. Then in the articles [21, 22], they combined projection pursuit and number theory to mine the hidden structural feature information in the space formed by multiple topological indices, which is associated with some certain chemical properties.

Mass spectrometry is another important aspect of chemometrics and has always been one of the essential methods for the identification and characterization of compounds. With the development of mass spectrometry technology, databases containing mass spectra of a large number of compounds and their other chemical information were established, such as NIST Library and Wiley Library. When the mass spectrum of the compound to be identified already exists in the mass spectrometry database, the computer retrieval method usually performs well. However, existing mass spectrometry libraries contain only a small fraction of the number of compounds: the Chemical Abstracts Service describes more than 200 million natural compounds. Therefore, when the mass spectrum of the compound is not in the existing mass spectrometry database, experts hope that some substructures of the compound can be identified by studying the existing mass spectrometry library.

The research group was also engaged in this research area. Their article [28] proposed a method for detecting the corresponding compound substructure by searching the peak combination of the mass spectrum and then using that peak combination for further compound identification and classification. They also proposed a method which applied the combination of the sliced inverse regression (SIR) and a decision tree to the mass spectrometry data for the identification of the substructure of the compound, which is published in [17].

Professors Fang and Liang’s research group also studied quantitative structural activity relationships (QSAR). QSAR research has always been an important branch of chemometrics, which aims to establish a quantitative relationship between active proprieties of compounds and their structural parameters through appropriate mathematical statistics methods. They adopted SCAD, a variable selection method with the oracle property, in [26] and selected 12 out of 128 topological indexes to establish the explainable connection between molecular boiling point and molecular structure. The articles [15, 34] introduced the Kriging model which is derived from geostatistics and the improved empirical Kriging model into the study of QSAR. Then they combined SCAD with the Kriging model and established the empirical Kriging model with selected important variables, as shown in the article [27]. This scheme has been applied in QSAR research and obtained better results than prior research.

As well as applying traditional statistical methods to where they were needed, the research group introduced the latest and most advanced statistical methods into chemometrics. In the article [31], they employed a two-step multivariate adaptive regression spline (TMARS) to show the relationship between the alkane retention index and the molecular structure. Later, in the subsequent three articles [18, 19, 29], they applied a boosting algorithm to improve the classification performance for the different types of chemical data.

The last contribution to chemometrics mentioned in this paper is Profs. Fang, Liang and the research group’s work with Traditional Chinese Medicine (TCM). They promoted and studied the data mining of TCM. Using statistical analysis and pattern recognition to indicate the authenticity of herb medicine and its main components has been widely used in the field of quality control for Chinese herbal medicines. A TCM fingerprint refers to the chromatogram or spectrogram that can be used to identify the chemical characteristics of TCM which has been properly treated by certain analytical means. In the paper [11], Prof. Fang employed a bootstrap method to estimate the probability distribution of the correlation coefficients of TCM fingerprints between the unknown test samples and the standard fingerprints under the assumption that they belong to the same category, and thus provided the test’s critical value for evaluating whether the fingerprint of an unknown test sample is qualified. In addition, he assessed the phylogenetic relationships of Lycium samples via random amplified polymorphic DNA (RAPD) and entropy theory in the paper [33].

In July 2010, Prof. Fang was invited to deliver a speech titled “A Challenge Research Direction in Biostatistics-Chinese Medicine” at the First Joint Biostatistics Symposium. The speech introduced the current status and challenges of research on the fingerprints of traditional Chinese medicine. In order to better communicate with international counterparts, some papers by the research group were selected

for publication in a special issue of the *Journal of Data Science* for data mining in chemometrics. In addition, Prof. Fang and his other collaborator Prof. Yu compiled two volumes titled *Data Mining and Bioinformatics in the field of Chemistry and Traditional Chinese Medicine* [12, 13], which have been highly regarded by international peers.

3.6 Summary

Professor Kai-Tai Fang has been engaged in the field of statistics for decades. He not only has devoted himself to the research of statistical theory, but also has been committed to promoting the development of statistical applications.

With a statistician's astute insight, Prof. Fang saw data mining would be an important research field of statistics and addressed the topic earlier than most statisticians. He has put a lot of energy into actively learning new theories, developing new methods, and courageously putting them into practice. Through in-depth communication with experts in other fields, he has studied and solved many practical problems. Even in the 1970s when computer technology wasn't developed in China, Prof. Fang persisted in overcoming various difficulties and successfully completed national projects.

Professor Fang realized early that in the era of data, statisticians need to adapt themselves to the development of the world, actively embrace data science and carry out research on statistical theories and methods based on actual needs. This attitude is demonstrated as various collaborative research by Prof. Fang's research groups. They made important contributions to the further application and development of statistics in chemometrics.

Acknowledgments This work was partially supported by Guangdong Natural Science Foundation (No. 2018A0303130231) and Guangdong Innovation and Enhancement Project: Education Research Programme (R5201920).

References

1. Atkinson, A.C., Bogacka, B., Bogacki, M.B.: D- and T-optimum designs for the kinetics of a reversible chemical reaction. Chemom. Lab. Syst. **43**, 185–198 (1998)
2. Cox, D.R.: Note on grouping. J. Am. Stat. Assoc. **52**, 543–547 (1957)
3. Fang, K.T.: Using conditional distributions to formulate a national clothing standard. J. Appl. Math. **2**, 63–74 (1976). (Chinese)
4. Fang, K.T.: Uniform design—application of number theory method in experimental design. J. Appl. Math. **3**, 363–372 (1980). (Chinese)
5. Fang, K.T.: Endless and unbending journey to statistics research: the oral autobiography of Kai-tai Fang. Hunan Education Press (2016) (in Chinese)

6. Fang, K.T., He, S.: The problem of selecting a given number of representative points in a normal population and a generalized mill’s ratio. Technical report, No. 5, Department of Statistics, Stanford University, USA (1982)
7. Fang, K.T., Wang, Y.: Number-Theoretic Methods in Statistics. Chapman and Hall, London (1994)
8. Fang, K.T., Wu, C.Y.: A probability extremum problem. J. Appl. Math. **2**, 132–148 (1979) (in Chinese)
9. Fang, K.T., Zhang Y.T.: Introduction to Multivariate Statistical Analysis. Science Press (1982) (in Chinese)
10. Fang, K.T., Bentler, P.M., Yuan, K.H.: Applications of number-theoretic methods to quantizers of elliptically contoured distributions. In: Multivariate Analysis and Its Applications, IMS Lecture Notes—Monograph Series, pp. 211–225 (1994)
11. Fang, K.T., Liang, Y.Z., Yin, X.L., Chen, K., Lu, G.H.: Critical value determination on similarity of fingerprints. Chemom. Intell. Lab. Syst. **82**, 236–240 (2006)
12. Fang, K.T., Liang, Y.Z., Yu, R.Q. (eds.): Data Mining and Bioinformatics in Chemistry and Chinese Medicines. Hong Kong Baptist University (2003)
13. Fang, K.T., Liang, Y.Z., Yu, R.Q. (eds.): Data Mining and Bioinformatics in Chemistry and Chinese Medicines, vol. 2, Hong Kong Baptist University (2004)
14. Fang, K.T., Xiang, K.F., Liu, G.Y.: Precision of Test Method. China Standard Press, Beijing (1988)
15. Fang, K.T., Yin, H., Liang, Y.Z.: New approach by Kriging methods to problems in QSAR. J. Chem. Inform. Model. **44**, 2106–2113 (2004)
16. Fang, K.T., Zhou, M., Wang, W.J.: Applications of the representative points in statistical simulations. Sci. China Ser. A **57**, 2609–2620 (2014)
17. He, P., Fang, K.T., Xu, C.J.: The classification tree combined with SIR and its applications to classification of mass spectra. J. Data Sci. **1**, 425–445 (2003)
18. He, P., Fang, K.T., Liang, Y.Z., Li, B.Y.: A Generalized boosting algorithm and its application to two-class chemical classification problem. Analytica Chimica Acta **543**, 181–191 (2005)
19. He, P., Xu, C.J., Liang, Y.Z., Fang, K.T.: Improving the classification accuracy in chemistry via boosting technique. Chem. Intell. Lab. Syst. **70**, 39–46 (2004)
20. Hu, Q.N., Liang, Y.Z., Fang, K.T.: The matrix expression, topological index and atomic attribute of molecular topological structure. J. Data Sci. **1**, 361–389 (2003)
21. Hu, Q.N., Liang, Y.Z., Peng, X.L., Yin, H., Fang, K.T.: Structural interpretation of a topological index. 1. External factor variable connectivity index (EFVCI). J. Chem. Inf. Comput. Sci. **44**, 437–446 (2004)
22. Hu, Q.N., Liang, Y.Z., Yin, H., Peng, X.L., Fang, K.T.: Structural interpretation of a topological index. 2. The molecular connectivity index, the Kappa index, and the atom-type E-state index. J. Chem. Inf. Comput. Sci. **44**, 1193–1201 (2004)
23. Lee, A.W.M., Chan, W.F., Yuen, F.S.Y., Tse, P.K., Liang, Y.Z., Fang, K.T.: An example of a sequential uniform design: application in capillary electrophoresis. Chem. Intell. Lab. Syst. **39**, 11–18 (1997)
24. Liang, Y.Z., Fang, K.T.: Robust multivariate calibration algorithm based on least median squares and sequential number theoretic optimization method. Anal. Chem. **121**, 1025–1029 (1996)
25. Liang, Y.Z., Fang, K.T., Xu, Q.S.: Uniform design and its applications in chemistry and chemical engineering. Chem. Intell. Lab. Syst. **58**, 43–57 (2001)
26. Peng, X.L., Hu, Q.N., Liang, Y.Z.: Variable selection via nonconcave penalty function in structure-boiling points correlations. J. Mol. Struct.: THEOCHEM, **714**, 235–242 (2005)
27. Peng, X.L., Yin, H., Li, R., Fang, K.T.: The application of Kriging and empirical Kriging based on the variables selected by SCAD. Analytica Chimica Acta **578**, 178–185 (2006)
28. Tang, Y., Liang, Y.Z., Fang, K.T.: Data mining in chemometrics: sub-structures learning via peak combinations searching in mass spectra. J. Data Sci. **1**, 481–496 (2003)
29. Varmuza, K., He, P., Fang, K.T.: Boosting applied to classification of mass spectral data. J. Data Sci. **1**, 391–404 (2003)

30. Xu, Q.S., Liang, Y.Z., Fang, K.T.: The effects of different experimental designs on parameter estimation in the kinetics of a reversible chemical reaction. Chem. Intell. Lab. Syst. **52**, 155–166 (2000)
31. Xu, Q.S., Massart, D.L., Liang, Y.Z., Fang, K.T.: Two-step multivariate adaptive regression spline for modeling a quantitative relationship between gas chromatography retention indices and molecular descriptors. J. Chromatogr. A **998**, 155–167 (2003)
32. Xu Q.S., Xu Y.D., Li L., Fang K.T.: Uniform experimental design in Chemometrics. J. Chem. e3020 (2018). https://doi.org/10.1002/cem.3020
33. Yin, X.L., Fang, K.T., Liang, Y.Z., Wong, R.N.S., Ha, A.W.Y.: Assessing phylogenetic relationships of Lycium samples using RAPD and entropy theory. Acta Pharmacologica Sinica **26**, 1217–1224 (2005)
34. Ying, H., Li, R.Z., Fang, K.T., Liang, Y.Z.: Empirical Kriging models and their applications to QSAR. J. Chem. **20**, 1–10 (2007)
35. Zhang, L., Liang, Y.Z., Yu, R.Q., Fang, K.T.: Sequential number-theoretic optimization (SNTO) method applied to chemical quantitative analysis. J. Chem. **11**, 267–281 (1997)
36. Zhang, L., Liang, Y.Z., Jiang, J.H., Yu, R.Q., Fang, K.T.: Uniform design applied to nonlinear multivariate calibration by ANN. Analytica Chimica Acta **370**, 65–77 (1998)

Chapter 4
A Review of Prof. Kai-Tai Fang's Contribution to the Education, Promotion, and Advancement of Statistics in China

Gang Li and Xiaoling Peng

Abstract As an eminent leader in the field of statistics, Prof. Kai-Tai Fang has made impactful contributions to the application, promotion, education and advancement of statistics in China. Under his leadership, his team had completed some of the China's hallmark industrial projects through novel applications of statistics and developments of new statistical methodologies. He has authored/coauthored a series of best-selling modern statistics textbooks, taught numerous workshops and short courses, and mentored a large number of students. He has been active in promoting scholastic exchanges and organizing national and international statistics conferences. He has also served on the leadership of many national and international statistics organizations and on the editorial boards of many major statistical journals. This article provides a selective review of Prof. Fang's contributions to the education, promotion, and advancement of statistics in China.

4.1 Background

Since the early twentieth century, statistics has seen a flourishing development, and the modern data-centric statistical data science has received extensive recognitions with widespread applications in all industries. In past decades, more and more Chinese statisticians started to show their talents in international statistical academia, and gained unprecedented recognition and attention. As one of the most influential pioneers of statistics in China, Prof. Kai-Tai Fang has dedicated himself to the education, promotion, and advancement of statistics in China during his entire

G. Li
Departments of Biostatistics and Biomedicine, UCLA, Los Angeles, CA 90095-1772, USA
e-mail: vli@ucla.edu

X. Peng (✉)
Division of Science and Technology, BNU-HKBU United International College, Zhuhai 519087, China
e-mail: xlpeng@uic.edu.hk

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_4

academic career. In celebration of his 80th birthday, this paper reviews the impactful contributions of Prof. Fang in three areas: application and popularization of statistics; discipline construction of statistics and cultivation of statistical talents; and statistical education in China.

4.2 Development and Popularization of Statistics Through Applications

Professor Fang did his graduate study in statistics in the Institute of Mathematics at the Chinese Academy of Sciences (IMCAS). Since graduation from IMCAS, Prof. Fang and his peers have dedicatedly devoted themselves to the development and popularization of statistics in China.

4.2.1 The Early Days of Statistical Popularization and Education in China

In 1964, the Institute of Mathematics at Chinese Academy of Sciences organized a team to conduct a collaborative research for Anshan steelworks (now known as Ansteel), in which Prof. Fang participated as a graduate student. In order to estimate the capacity of a stove of molten steel, a young engineer spent several months collecting large amounts of data but struggled to make sense of the data. Professor Fang thought that non-linear regression models might work here. However, although his undergraduate major in Peking University was in probability and statistics, back then the curriculum mostly focused on theory rather than application. For example, linear regression analysis was covered by only a 2 h lecture. After some research, Prof. Fang found a needed non-linear regression model in Huazhang Zhou's book entitled *Applied Mathematical Statistics of Industrial Technology*, which was appropriate for analysis of the molten steel data and led to a good estimation of the capacity of molten steel. The success of this project drastically stimulated the desire of engineers and technicians for statistical knowledge. To help them learn statistics, Prof. Fang wrote an easy-to-understand handout named "Six Lectures on Mathematical Statistics" tailored to their needs, which was later published via Anshan Metal Association.

In 1963, Beijing Vinylon Factory purchased a fully automated factory from Japan. But the engineers and technicians in that factory were only told the process of production, not the principles behind it. Therefore, during the cultural revolution, Prof. Fang and his colleagues from IMCAS were invited to Beijing Vinylon Factory to help them understand the principles underlying the production process. Using orthogonal designs, Profs. Fang and Ping Cheng, Chairman of Probability and Statistics Research Lab at IMCAS, helped the factory to "decipher" many rules of the production process as detailed in [2]. The engineers and technicians were fancinated by the power of

statistics, especially the power of experimental design and regression analysis, and expressed great interesting in learning statistics. To help them learn the statistical principles and applications, Prof. Fang wrote a lecture notes based on cases studies, instead of abstract theories. During the same time period, many peer institutes, such as chemistry, biological physics and developmental biology, within the Chinese Academy of Sciences, have also became aware of the importance of statistics in their research and approached IMCAS for help to provide statistical training for their researchers. In response to their requests for statistical education and training, Prof. Fang and his colleagues held free statistical workshops inside the Chinese Academy of Sciences, training more than 100 research fellows from various disciplines. In the 1970s, realizing the increasing demand for the application of statistics, Prof. Ping Cheng suggested to combine and improve the previous lecture notes into a textbook of common mathematical statistical methods [6], which was published via Science Press in 1973. This book was issued more than 200,000 copies and sold out very quickly. It was published again in 1974–1979 and became a bestseller. During that historical period of China, a book can only be published in the name of the collective and the author did not receive any royalty. The only reward to the author(s) was 300 free copies of the book from the publisher. Professor Fang generously donated these books to readers who needed them.

From 1971 to 1975, Prof. Fang participated in many promotional events for the orthogonal design. Along the way, he learned at first hand that the analysis of variance method was not easily understood by the majority of engineers. In the age with no computers and electronic calculators, it was difficult to calculate the ANOVA table. To address these issues, Prof. Fang created "visual analysis of orthogonal design", which made it much easier for engineers to understand the principles of orthogonal design and data analysis using charts [7]. It is worth noting that in 1976, Prof. Fang was assigned by IMCAS to run a TV show to introduce and advertise the orthogonal design to the general audience on the China Central Television (CCTV). CCTV allocated 17 min for Prof. Fang to present a lecture on orthogonal design, which was aired during the prime time following the CCTV news. Because of Prof. Fang's well prepared and delivered lecture, the show was a big success and attained desired effect of promoting orthorgonal design in China.

4.2.2 Determination and Examination of National Standards

The formulation of national standards requires a solid theoretical foundation and abundant experience in dealing with practical problems. In the 1970s, Prof. Fang participated in three national projects related to national standards: the examination of national standards for alloy structural steel, the establishment of Chinese adult clothing standard, and the introduction of standard for precision of testing methods.

In 1973, Prof. Fang and his colleague Prof. Chuanyi Wu from IMCAS collaborated on a project from department of alloy structural steel to review national standards for alloy structural steel. Professor Fang and his colleagues built regression models to

predict five mechanical properties of steel with its chemical elements and used five-fold integration to calculate the qualifying rate of steel. Finally, Prof. Fang and his colleagues reached three conclusions: (1) the Chinese national standards for alloy structural steel are scientifically sound; the combination of different elements in steel will affect the qualifying rate of mechanical properties and has to be optimized; (2) the previous empirical methods from Beijing Steel and Fushun steel mill are indeed reasonable. This project started with a huge amount of data, went through laborious data cleaning, modeling, and testing, and finally arrived at convincing conclusions, which can be viewed as an early form of the data mining in the modern age. This collaboration further led to several of Prof. Fang's theoretical research projects later on (see [9]).

Manufacture and research often involve a wide variety of instruments. The International Organization for Standardization (ISO) set the standard of ISO5725 to measure the testing precision of an instrument in terms of its repeatability and reproducibility. To translate this standard to China, the Standard Administration of China (SAC) appointed a special committee with Prof. Fang being the chair. Professor Fang explained to the committee members the statistical principles underlying the ISO5725 and led the committee to complete the Chinese national standard GB/T6379 after two years of hard work. This standard indeed relies heavily on statistical methods including analysis of variance with random effect, the elimination of outliers, and linear regression. Together with GB/T6379, the committee also published a monograph [14] to explain the statistical theory and methods used for GB/T6379. This project was later awarded the second prize from SAC. Recognizing his important contribution and indispensible role in developing GB/T6379, Prof. Fang was later asked to serve on the international ISO5725 committee as a China representative.

In the 1970s, the Ministry of Light Industry, the Ministry of Textile, and the Standard Administration of China, jointly set up a working team to establish China's first clothing standard. Measurements were taken from 400,000 people using stratified random sampling. During the project implementation, Prof. Fang noticed that the popular principal component analysis (PCA) method was not most suitable for the Chinese data. As an alternative, he suggested to use the conditional distribution combined with D-optimal design and successfully developed a new novel method for establishing the Chinese clothing standards [18]. The developed standard adopted in China and became effective in 1977. It was awarded a special prize by the Ministry of China Light Industry. Afterwards, Prof. Fang also collaborated with the Institute of Ancient Vertebrate and Institute of Ancient Human at Chinese Academy of Sciences and developed the Chinese head type standard, which received the Beijing Science and Technology Progress Award. In 1982, to improve the clothing standard for body type, Prof. Fang introduced the concept of statistical distribution representative points and derived a numerical algorithm to compute representative points for univariate normal distribution. Unfortunately he later found out that his work was highly overlapping with that of Cox [1]. Not being discouraged, Prof. Fang further pursued this idea and proposed the NTLBG algorithm [15] based on number theory and k-means algorithm in 1994 to compute the representative points of multivariate symmetrical distributions (elliptically contoured distributions). In 2014, Prof. Fang and his

students used representative points and random sampling to significantly improve the efficiency of the Monte Carlo method [16]. In 2015, Prof. Fang and his students discovered a seemingly impossible property of arcsine distribution representative points [17]. Their further research on this property led to a modified definition of distributional representative points for further improvement of resampling efficiency [26].

4.2.3 Number Theory Methods in Statistics

Another example of Prof. Fang's many impactful contributions to the development and application of statistics is his novel introduction and development of number theory methods in statistics. The number theory, founded by Kolmogorov, concerns how to produce a point set uniformly distributed in a high-dimensional rectangle. This method was largely used in high dimensional numerical integration. Famous Chinese mathematicians Luogeng Hua and Yuan Wang have made important contributions to this field. In 1978, Profs. Fang and Yuan Wang first applied the number theory method in computer experiments to create uniform design, which broke ground for a brand-new research area in experimental design. In 1988, Institute of Applied Mathematics at Chinese Academy of Sciences and Second Artillery Force (renamed as People's Liberation Army of China Rocket Force in 2015) began a series of collaborations, which involved estimation of the probability of geometric flow patterns and related optimization problems. Most of these problems have no analytical solutions. If stochastic simulation is used, uniform grids need to be designed on geometric flow patterns, which requires the expansion of number theory methods from hyperrectangle to different geometric manifolds. Motivated by these applications, Prof. Wang and Fang submitted a grant proposal "the applications of number-theoretic methods in statistics", which was funded by The National Natural Science Foundation of China and Hong Kong government. They successfully solved the problems needed for national defense construction and won the second prize of progress in science and technology from Chinese Academy of Sciences in 1989 and the first prize of progress in science and technology from People's Liberation Army of China in 1992. Their pioneer research has promoted the wide applications of number-theoretic methods in statistics. For example, the SNTO algorithm to solve the optimization calculation of multivariate non-linearity was developed for parameter estimate, maximum likelihood estimation, sequential test design, optimal design of experiment, and so on in terms of non-linear regression. It is also used for the statistical inferences of multivariate projection, including multivariate normality hypothesis testing, multivariate ellipsoidal contour distribution testing, and multivariate maximum likelihood estimation, etc. These results are included in monographs [10] and statistical encyclopedia [11].

4.3 Contributions to Statistical Education

Professor Fang always values statistical education and statistical talent cultivation and has devoted his entire career to statistical education in China.

4.3.1 Cultivating Outstanding Statistical Talents as a Pioneer of Statistics in China

In 1983, Prof. Fang began to supervise graduate students after coming back from a two-years academic visit of the US. At that point, the research section in probability and statistics in the Institute of Applied Mathematics at the Chinese Academy of Sciences recruited four master graduate students: Jianqing Fan, Hui Quan, Fanhui Kong, and Hongqing Zhang. They all chose Prof. Fang to be their academic advisor. Due to the shortage of academic advisors for graduate students in statistics in China, many universities sent their graduate students to Prof. Fang to study statistics. As a result, Prof. Fang's students during that period were not only from Chinese Academy of Sciences but also from other many universities in China. Among them, four were from Nankai University, two from Yunnan University, one from Wuhan University, one from Southeast University, two from National Bureau of Statistics of China, one from Beijing Institute of Technology, one from Soochow University, and one (the first author of this article) from Shandong College of Oceanography (now Ocean University of China). In addition, three junior researchers from Institute of Applied Mathematics, Biqi Fang, Ping Yan, and Xiaoming Chen, also joined Prof. Fang's research team.

Professor Luogeng Hua, a famous Chinese mathematician, once said "high quality seminars are essential to high quality research in a research institute". To help his students to get to the frontiers of statistics, Prof. Fang organized a journal club on multivariate statistical analysis which meets several times a week. In addition to the classical multivariate analysis theory, he chose some state of the art text books and papers on generalized multivariate analysis for his students to study and discuss. At the regular journal club meetings, students reported their own research results and actively engaged in discussions, which helped them to gain thorough understanding of the studied topics. To help his students, especially those from outside of Beijing, Prof. Fang offered personal assistance to arrange their dormitories and study space, and sometimes even subsidized their travel and lodging expenses out of his own pocket.

Under Prof. Fang's guidance, his students made quick progress in learning generalized multivariate statistical analysis, a relatively new area of research at that time. Thanks to the stimulating and highly collaborative research atmosphere, Prof. Fang's students from the journal club were highly productive. On average, each graduate student published 2–8 papers at graduation. From 1983 to 1988, Prof. Fang and his students from that cohort published more than 60 papers on generalized mul-

tivariate statistical analysis, 40 of which were included in "*Statistical Inference in Elliptically Contoured and Related Distributions*" [5], co-edited by Profs. Fang and T.W. Anderson.

In the early 1980s, China began to open its door to the world. Eager to learn from the world, studying abroad became a dream of many young students in those days, including Prof. Fang's students. Without any hesitation, Prof. Fang helped many of his students to study abroad to pursue their dreams. Many of his former students have become well known scholars in the international statistics community. Among them is Prof. Jianqing Fan, a chair professor in the Department of Statistics and Finance at Princeton University, who received the COPSS award, the highest honor for a statistician, and became the first Chinese chief editor of "*The Annals of Statistics*", a top statistics journal. Interestingly, Prof. Fang's another graduate student, Runze Li, now a professor at Pennsylvania State University, was also appointed as a chief editor of "*The Annals of Statistics*" in 2013.

Because of his extraordinary achievements, Prof. Fang was appointed as a Ph.D. advisor by the Degree Committee of the China State Council in 1985. The information brochure for Ph.D. Advisors at Chinese Academy of Sciences published in 1989 has the following description of Prof. Fang: "Prof. Fang is one of the pioneers of mathematical statistics in China", and "he has made world-class achievements not only in the theoretical development, but also applications of statistics."

In 1992, Prof. Fang joined the Department of Mathematics at Hong Kong Baptist College (now Hong Kong Baptist University) as chair professor. At that time, China's educational circles knew very little about universities in HK, and very few students from mainland went to Hong Kong to study or do research. Professor Fang made tremendous efforts to initiate and strengthen scholastic exchange programs between the mainland China and Hong Kong and encourage more mainland Chinese students to study in Hong Kong. He set an example by recruiting many Ph.D. students from mainland China including Jianxin Pan, Minyu Xie, Jiajuan Liang, Hongbin Fang, Guoliang Tian, Hong Qin, Yu Tang, Ping He, Xiaoling Peng (one of the authors of this paper), Hong Yin, Hongya Zhao, Xiaolin Yin, and a graduate student Aijun Zhang. Many of these students became well established leaders in statistics. For example, Jianxin Pan is now a chair professor at University of Manchester in the United Kingdom and a Turing Fellow of The Alan Turing Institute of the United Kingdom. During his three-year doctoral study in Hong Kong, he had published six papers in international journals. In 2002, Profs. Pan and Fang collaboratively published a monograph [23] in Springer. Besides, Prof. Fang also helped Prof. T.W. Anderson advising a Ph.D. student in Stanford University and advised an M.Phil student at the North Carolina University at Chapel Hill.

Professor Fang's research has received grant support from the Hong Kong government research fund many times. With these grant support, he invited many mainland China scholars to Hong Kong for research collaborations and cultivated a number of excellent young scholars for China, including Xiaoqun Wang, Gennian Ge, Chang-Xing Ma, and Min-Qian Liu among others.

4.3.2 *Creating Undergraduate Statistics Major for Liberal Arts Education*

In 2005, when Prof. Fang retired with honor from Hong Kong Baptist University, many famous universities from foreign countries invited him to join them. At the same time, Prof. Ching Fai Ng, the president of Hong Kong Baptist University, was working to set up a new liberal arts university in Zhuhai, named Beijing Normal University-Hong Kong Baptist University United International College (UIC), which became the first Chinese university jointly ran by Chinese and Hong Kong universities. In UIC, students receive comprehensive education in humanity, arts, science, social science with broad skills and focus on critical thinking and creativity cultivation. Professor Ching Fai Ng invited Prof. Fang to join UIC to found an undergraduate statistics major in UIC. Sharing a common goal, Prof. Fang decided to accept the invitation and joined UIC to build an undergraduate statistics major from the ground up.

Professor Fang worked tirelessly to forge a statistics major in UIC with a modern curriculum. With his broad experience in statistical research, application and education, Prof. Fang envisioned that a student majoring in statistics should have a sound mathematical foundation, be well trained in statistical thinking and methods, and be skillful in programming and using software for data analysis and modeling. With these principles in mind, he developed a statistics curriculum that is in line with the modern international standards. The UIC statistics curriculum includes not only core courses in mathematics, operations research, and statistics, but also courses in computer programing, data analysis and modeling. Furthermore, most of the statistical courses at UIC are designed to include group projects, which require students to apply statistical methods to analyze real data, write code implementation, and make an oral project presentation in English. The UIC statistics curriculum has well prepared students for job placement and graduate studies after graduation. It has also been well received among domestic and international experts.

Faculty recruitment was the most difficult in the first few years due to insufficient school funding and a shortage of qualified fluent English speaking statisticians in China. Many departments of statistics in China had to deal with the same issue at the time and were unable to offer a comprehensive range of elected courses in statistics. Utilizing his network, Prof. Fang successfully recruited four junior faculty to UIC. He led by example and taught four elective courses each year himself. In addition, he persuaded many of his old friends including Prof. Jianzhong Zhang of City University of Hong Kong, Prof. Yung Liang Tong of Georgia Institute of Technology from the United States, Prof. Philips Cheng of Academia Sinica from Taiwan, and Prof. Kai Fun Yu of FDA from the United States to offer elected courses for UIC students. As a result, students in the UIC statistics program benefited greatly from these courses taught by high quality and knowledgable instructors. It is worth noting that in addition to teaching, Prof. Fang had also served as Director of the Statistics Program at UIC for six years and handled administrative affairs routinely.

Besides teaching, Prof. Fang also placed great emphasis on cultivating student's research ability. He served as undergraduate thesis advisor for many senior students every year. He used summer breaks to run student seminars and set a high standard for their thesis. As a results, many of the UIC undergraduate students thesis were published in international journals [13, 17, 19–22, 24].

Over 70% of the students who graduated from the UIC statistics program went on for graduate study abroad, many in prestigious universities such as Columbia University, Johns Hopkins University, Purdue University, University of Pittsburgh, Georgetown University, University of Oxford, University College London, University of Manchester in the United Kingdom, Australian National University, University of Melbourne, University of Hong Kong, Hong Kong University of Science and Technology, and University of Tokyo. Professor Fang also helped to develop joint master degree programs with Department of Biomedicine at Georgetown University, Department of Applied Statistics at Victoria University of Wellington, New Zealand, and Hong Kong Baptist University.

In UIC, Prof. Fang also founded the Institute of Statistical and Computational Intelligence (ISCI) and invited some famous scholars including academician Jiaan Yan from Chinese Academy of Sciences, Prof. Jianqing Fan from Princeton University, Prof. Xiaoli Meng from Harvard University to give seminars. The Four-Dimensional Statistical Lab at ISCI has established a high reputation for offering comprehensive statistical consulting services within UIC and to industries and government agencies in the great Pearl River Delta area. His contributions to UIC were not limited to establishing the statistical major. With his assistance, the Division of Science and Technology at UIC successively established the Financial Mathematics major and the Data Science major in 2011 and 2017, respectively. In 2018, UIC statistical major has graduated over 400 students over a 10-year period. Their graduates entered a variety of industries inlcuding finance, biomedicine, and internet technology, and many of them became industry leaders. In the UIC statistics alumni conference held in December 2017, more than 100 alumni, teachers and students gathered at the UIC new campus to share their experiences and achievements at work and study and pay tribute to Prof. Fang for introducing them to the fascinating field of statistical data science and paving the way for their success.

4.3.3 Improving and Writing Statistical Textbooks

Influenced by the Soviet Union, the early statistical textbooks in China overemphasized on probability theory over statistical thinking and focused more on theory than application. Over the years of conducting collaborative research using statistics, Prof. Fang has written many popular statistics textbooks, such as *Introduction to Multivariate Statistical Analysis* [25], *Practical Multivariate Statistical Analysis* [3], *Statistical Distributions* [12], and *Orthogonal and Uniform Design* [8]. These books give a balanced account of statistical methods and their applications, with many real data examples. Because of his authority in the field of statistics and his outstanding

achievements in writing statistical textbooks, in 2009, China Higher Education Press invited him to be the chief editor for a new series of statistics textbooks. In 2010, the editorial board of modern statistics college textbooks was established and held its first meeting in Fuzhou. Professor Fang emphasized that a statistics textbook should modernize its contents, stress statistical thinking, and adapt to the characteristics of modern statistical education and the new requirements of the fast evolving data science era. It should be fun to study and easily accessible, with an emphasis on the application of statistics, supported by statistical software. Since its inception, this new textbook series has so far published 20 statistics textbooks. To enhance the research of graduate students and scholars majoring in statistics, China Higher Education Press decided to publish *Lecture Notes in Probability, Statistics and Data Science* in 2017, and also invited Prof. Fang as a chief editor. The editorial committee have already met twice and decided on the first collection of books to be published. Finally, Prof. Fang also served as a chief editor for *A Series of Modern Applied Mathematical Methods in the 20th Century*, supported by China Tianyuan Foundation and published by Science Press (see Appendix A.2C).

4.4 Academic Services

In recognition of his extraordinary achievements in mathematical statistics and its application, Prof. Fang has been invited to serve and lead many academic and professional organizations. Among others, he was Associate Director of the Institute of Applied Mathematics at Chinese Academy of Sciences, Chair of Department of Mathematics at Hong Kong Baptist University, Managing Director and Secretary—General of Chinese Mathematical Society, Managing Director and Secretary—General of the Chinese Society of Probability and Statistics, Director of the Sixth Division of National Standardization Committee of Mathematical Statistical Methods, Director of Multivariate Analysis Committee, Director of Uniform Design Division of Chinese Mathematical Society, Managing Director of Hong Kong Mathematical Society, Councilor of International Chinese Statistical Association, and Managing Director of International Statistical Association (see Appendix A.2A). He has provided dedicated services and made great contributions to management, academic exchanges and services, and development of the statistics profession in China.

Professor Fang also served on the editorial board of many international and domestic academic journals including *Journal of Multivariate Analysis*, *International Statistical Review*, *Statistical Sinica*, *Statistics & Probability Letters*. He was the vice chief editor of *Journal of Applied Mathematics* for many years (see Appendix A.2B for details of Prof. Fang's services in academic journals). He also reviews many papers for a variety of academic journals every year.

4.4.1 Reforms in the Institute of Applied Mathematics at Chinese Academy

Professor Fang has long been a researcher in the Institutes of Mathematics and Applied Mathematics (IMAS) at Chinese Academy of Sciences. In 1984, Chinese Academy of Sciences appointed him as the Asscociate Director of Institute of Applied Mathematics (IAM) at Chinese Academy of Sciences (equivalent to vice President of a university). This was a big challenge to him. He was in charge of the research projects and finance of the institute, and at the same time, served as the director of the academic committee. During that period, there were only very limited allocated research fund to support research and scholastic exchanges. To stimulate research, he proposed to give project leaders more freedom in managing research funds so they can have more opportunities in participatingin scholastic exchange activities. This reform gave researchers more incentive to bring in more outside research projects and greatly increased the institute's research revenue to support more research activities.

To bring in new talents to research, Prof. Fang led IAM to recruit more graduate students. By developing joint training programs with other universities, IMS recruited and trained a large number of young scholars, and reached to a total of 120 graduate students at one time. Dr. Fang efforts had fostered an active research environment that encouraged critical thinking, hard work and collaborations and resulted in productive research. During his visiting at Stanford University, Prof. Fang was inspired by their technical reports series, which offered a fast way of disseminating new research results, drastically accelerated further research, and protected the author's copyright. To introduce and help establishing a technical report series in his own institute, Prof. Fang donated a Latex software and bought a laser printer from the US, and reviewed every technical report to ensure its quality.

4.4.2 Organize Academic Conferences and Promote Research Communications

Since the 1970s, Prof. Fang has helped organizing over 30 domestic and international academic conferences (see Appendix A.1). He has spent tremendous amount of time and energy to develop and promote statistics in China. Here we briefly describe some selected conferences that Prof. Fang helped to organize during some special historical periods with lasting impact.

As early as the end of the China's "cultural revolution", Prof. Fang and other members of a multivariate analysis discussion group planned to organize a national academic conference at Mount Huangshan for the coming of a new era in China's science development. Hosting a conference requires financial planning, fundraising, administrative approval, hotel arrangement, purchase of return train tickets, and arrangement of local transportation, etc. These tasks may not seem difficult nowadays, but did require enormous amount of hard work and efforts at that time since

Prof. Fang held no administrative positions and had no administrative assistant. With the support of the IAM leadership and colleagues, Prof. overcame many difficulties and successfully organized the first national conference on Multivariate Statistical Analysis in China. Moving away form the past tradition of emphasizing theory over application, this conference stroke a balance between theory and application, and made them complement each other. Presenters included not only theoretical researchers from universities, but also those doing practical applications such as National Meteorological Center of CMA in China. The conference was full of stimulating discussions and concluded with unprecedented success. This conference also established Prof. Fang as a domestic leader in the field of multivariate statistical analysis. During his presidency on the multivariate analysis committee, he has directly involved in organizing six national academic conferences, which had lasting impacts in promoting multivariate statistical analysis and its applications in China.

The US had been leading the world in the development of statistical science and its applications. To help China learn from the US, Prof. Fang proposed to organize a "Sino-American Statistical Conference". His proposal received enthusiastic response and support from Prof. George C. Tiao of University of Chicago, as well as from the Institute of Applied Mathematics and Institute of System Science at Chinese Academy of Sciences, and Chinese Probability and Statistics Association. Professor Fang led his team and spent a lot of time and efforts to prepare for the conference from fundraising, contacting American peers, session organizing and conference logistics. The conference was successfully held In 1987, with over 230 participants (60 from the US and more than 170 from mainland China). At the conference, Prof. George C. Tiao brought up his idea of establishing an professional organization to bring together all the Chinese statisticians around the world, which led to the later establishment of the International Chinese Statistical Association (ICSA), the four largest statistical associations in the world today. Professor Tiao asked Prof. Fang to use his influence to promote ISCA in mainland China. Over the years, Prof. Fang has made countless contributions to promote ICSA and statistics in China and beyond. Notable examples include member recruitment, organizing the first ICSA International Conference in Hong Kong in 1990, and serving on editorial board of Statistical Sinica, the flagship journal of ISCA, and serving as an elected board member of ISCA.

In 1992 and 1997, as conference chair, Prof. Fang successfully organized two large International Multivariate Analysis Symposiums in Hong Kong. Many world-class statisticians attended the conferences, including 3 members of the US National Academy of Sciences, 22 fellows of the Institute of Mathematical Statistics (IMS) and 15 fellows of the American Statistical Association (ASA). These two conferences had made significant positive impacts on the Hong Kong statistics community. To commend Prof. Fang's outstanding contributions to statistics, an "International Conference on Statistics in Honor of Prof. Kai-tai Fang's 65th Birthday" was held in Hong Kong Baptist University on his 65th birthday in 2005. This conference was attended by more than 150 statisticians from more than 20 countries around the world such as China, the United States, the United Kingdom, Canada, among others. In 2014, ISCA awarded Prof. Fang the "2014 ICSA Outstanding Achievement Award".

Professor Fang is passionate about statistics and its application and has tirelessly devoted all his life and energy to the statistics profession [4]. Now aged 80, Professor shows no signs of slowing down. He is still active in doing research, advising graduate students, continuing his efforts to develop and mature the statistical major in UIC, and leading the way to improve and publish modern statistical textbooks in China. Professor Fang has set an example of a truly examplary and devoted statistician for the generations to come.

Acknowledgments Xiaoling Pengs work was partially supported by Guangdong Natural Science Foundation No. 2018A0303130231.

Appendix A.1: Academic Conferences Organized by Prof. Kai-Tai Fang

1 Sino-Japan conference on statistics, Member, Chinese organization committee, 1984, Beijing, China.
2 Sino-Japan symposium on statistics, Member of Chinese organization committee, 1986, Fukuoka, Japan.
3 Sino-US conference on statistics, Member of Chinese organization committee, 1987, Beijing, China.
4 Sino-Japan symposium on statistics, Member of Chinese organization committee, 1989, Tokyo, Japan.
5 Sino-Japan symposium on statistics, Member of Chinese organization committee, 1986, Okayama, Japan.
6 IMS conference, Organizer of Multivariate Analysis Under Non-normal Population, Colorado, USA, 1988.
7 Asian congress of mathematicians, Deputy Head of the Chinese Delegation, August 1998, Hong Kong, China.
8 The first conference on recent developments in statistics research, Hong Kong, December 1990.
9 International symposium on multivariate analysis and its applications, Chair of the organization committee, 1992, Hong Kong, China.
10 International workshop on Quasi-Monte Carlo methods and their applications, organizer, 1995, Hong Kong, China.
11 1997 International Symposium on Contemporary Multivariate Analysis and Its Applications, Chair, 1997, Hong Kong, China.
12 1999 Symposium on Theory of uniform Design and Its Applications, Chair, 1999, Hong Kong, China.
13 The 4th Monte Carlo and Quasi-Monte Carlo Conference in Scientific Computing), Chair, 2000, Hong Kong, China.
14 The 5th ICSA international conference, Member of the organizing committee, 2000, Hong Kong, China.

15 The 5th International Conference on Optimization Techniques and Applications, Member of the organizing committee, 2001, Hong Kong, China.
16 The 5th International Conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, Member of the organizing committee, Nov. 2002, Singapore.
17 International Conference on Applied Statistics, Actuarial Science and Financial Mathematics, Member of the organizing committee, Dec. 2002, Hong Kong, China.
18 2003 Symposium on The Uniform Experimental Design and its applications, Chair, Dec. 2003, Shenzhen, China.
19 The 6th ICSA International Conference, Honorary Advisor, Jul. 2003, Singapore.
20 International Conference on Chemometrics and Biometrics in Asia, Member of the organizing committee, Oct. 2003. Shanghai, China.
21 The International Congress of Chinese Mathematicians (ICCM), Member of the organizing committee, 2004, Hong Kong, China.
22 International Workshop on Applied Mathematics and Statistics, Chair, Dec. 2004, Hong Kong, China.
23 International Conference on the Future of Statistical Theory, Practice and Education, Member of the International Advisory Board, Dec. 2004–Jan. 2005, Hyderabad, India.
24 2005 Symposium on The Uniform Experimental Design and Its Applications, Member of the organizing committee, Aug. 2005, Jishou, China.
25 International Conference on Design of Experiments: Theory and Applications, Member of the international advisory committee, May 2005, Memphis, the United States.
26 The International Committee, International Conference: Statistics in the Technological Age, Member of the international advisory committee, Dec. 2005, Kuala Lumpur, Malaysia.
27 The 29th European Meeting Of Statisticians, Member of the international advisory committee, Jul. 2013, Budapest, Hungary.
28 The 24th International Workshop on Matrices and Statistics (IWMS), (Member of Scientific Organizing Committee, May 2015 Haikou, China.

Appendix A.2: Prof. Kai-tai Fang's Academic Services

A.2A Academic Organizations

1 Chinese Society of Probability and Statistics, Secretary-general, Oct. 1982–Oct. 1984.
2 Chinese Society of Probability and Statistics, Executive Director, 1982–1990.
3 International Statistical Institute (ISI), Elected Member, 1985–2009.

4 Chinese Mathematical Society, Executive Director, Vice Secretary-general, 1988–1992.
5 Chinese Organization for Standardization of Statistical Methods Applications, Committee Member, Director of the Sixth Chapter Committee, 1983–1990.
6 Chinese Association of Multivariate Statistical Analysis, Director, 1980–1990.
7 Chinese Committee of Mathematical Geosciences, Committee Member, 1978–1985.
8 International Association for Mathematical Geosciences, Director, 1981–1985.
9 Institute of Mathematical Statistics (IMS), Life Member, 1988–, Elected Member, 1992–.
10 International Chinese Statistical Association (ISCA), Life Member, 1988–, Director, 1990–1994.
11 Chinese Mathematical Society, President of Uniform Design, 1993–2003.
12 Chinese Mathematical Society, Honorary President of Uniform Design, 1993–.
13 Hong Long Institution of Science, Director, 1994–1998.
14 International Statistical Institute(ISI), Executive Director, 1995–1999.
15 Hong Kong Mathematical Society, Executive Director, 1994–1996.
16 Hong Kong Mathematical Society, Fellow, 1990–.
17 Southeast Asian Mathematical Society, Fellow, 1990–.
18 Hong Kong Statistical Society, Fellow, 1991–, Honorary Member, 2002–.
19 American Statistical Association (ASA), Fellow, 1993–, Elected Member, 2001–.
20 Institute of Mathematical Statistics (IMS), Member of the Academician Selection Committee, 2007–2009.
21 Statistics Research Association of Anhui Province, Honorary President, 2001–2004.
22 Experimental Design Chapter of China Statistics Research Association, Honorary President, 2010–.

A.2B Academic Journals

1 Acta Mathematicae Applicatae Sinica, Vice Director, 1985–1992, Editor, 1992–.
2 Chinese Journal of applied probability and statistics, Editor, 1985–1990.
3 Journal of Mathematical Research with Applications and Comments, Editor, 1986–.
4 Northeastern Mathematical Journal, Editor, 1985–.
5 Journal of Quantitative Economics, Editor, 1984–.
6 Mathematical Theory and Applied Probability, Editor, 1986–.
7 Statistics & Probability Letters, Editor, 1988–2005.
8 Statistica Sinica, 1993–1999, Editor, 2005–2012.
9 Journal of Multivariate Analysis, Editor, 2002–2007.
10 Statistics & Information Forum, Editor, 2009–.
11 International Statistical Review, Editor, 2009–2010.

A.2C Editor-in-Chief Series

1 Modern Applied Mathematics Methodology Series, Science Press, 1990–2004.
2 Higher Education Modern Statistics Series, Higher Education Press, 2010–.
3 Lecture Notes in Lecture Notes in Probability, Statistics and Data Science? For Higher Education Press, Beijing, 2017–2022.

References

1. Cox D.R.: Note on grouping. JASA. **52**, 543–547 (1957)
2. Fang, K.T. (in the name of collective): Application of statistics in Vinylon production. Practice and understanding of mathematics (1972)
3. Fang, K.T.: Practical Multivariate Statistical Analysis. East China Normal University Press, Shanghai (1989)
4. Fang, K.T.: Long Long Road to Success—Kaitai Fang's Self-narration, a Series of Chinese Oral History of Science in the 20th Century. Hunan Education Press, Changsha (2016)
5. Fang, K.T., Anderson, T.W. (eds.): Statistical Inference in Elliptically Contoured and Related Distributions. Allenton Press Inc., New York (1990)
6. Fang, K.T., Dai S.S.: Common Mathematical Statistical Methods. Science Press, Beijing (1973)
7. Fang, K.T., Liu, Z.X. (in the name of collective): Orthogonal experimental design. Nonferrous Metals **8**, 39–56 (1974)
8. Fang, K.T., Ma, C.X.: Orthogonal and Uniform Experimental Design. Science Press, Beijing (2001)
9. Fang, K.T., Wu, C.Y.: A probabilistic extreme value problem. J. Appl. Math. **2**, 132–148 (1979)
10. Fang, K.T., Wang, Y.: Number-Theoretic Methods in Statistics. Chapman and Hall, London (1994)
11. Fang, K.T., Wang, Y.: Number-Theoretic Methods. Encyclopedia of Statistics, vol. 2, pp. 993–998. Wiley, New York (1997)
12. Fang, K.T., Xu, J.L.: Statistical Distribution. Science Press, Beijing (1987)
13. Fang, K.T., Zheng, Y.X.: Magic squares. Math. Cult. **4**(3), 52–65 (2013)
14. Fang, K.T., Xiang, K.F., Liu, G.Y.: Precision of Test Method. China Standard Press, Beijing (1988)
15. Fang, K.T., Bentler, P.M., Yuan, K.H.: Applications of number-theoretic methods to quantizers of elliptically contoured distributions. In: Multivariate Analysis and Its Applications. IMS Lecture Notes—Monograph Series, pp. 211–225 (1994)
16. Fang, K.T., Zhou, M., Wang, W.J.: Applications of the representative points in statistical simulations. Sci. China Ser. A **57**, 2609–2620 (2014)
17. Jiang, J.J., He, P., Fang, K.T.: An interesting property of the arcsine distribution and its applications. Stat. Prob. Letters **105**, 88–95 (2015)
18. Ke, M., Fang, K.T.: Application of the theory of conditional distribution to revise the national standard for clothing. J. Appl. Math. **2**, 62–74 (1976)
19. Ke, X., Zhang, R., Ye, H.J.: Two- and three-level lower bounds for mixture L2-discrepancy and construction of uniform designs by threshold accepting. J. Complex. **31**(5), 741–753 (2015)
20. Lin, Y., Fang, K.T.: The main effect confounding pattern for saturated orthogonal designs. In: Metrika (2019) (to appear)
21. Lin, Z., Liu, S., Fang, K.T., Deng, Y.H.: Generation of all magic squares of order 5 and interesting patterns finding. Spec. Matrices **4**(1), 110–120 (2016)
22. Ma, X.Y., Fang, K.T., Deng, Y.H.: Some results on magic squares based on generating magic vectors and R-C similar transformations. Spec. Matrices **5**, 82–96 (2017)

23. Pan, J.X., Fang, K.T.: Growth Curve Models and Statistical Diagnostics. Springer, New York (2002)
24. Styan, G.P.H., Fang, K.T., Zhu, J.L., Lin, Z.Q.: Magic squares on stamps. Math. Cult. **6**(3), 109–118 (2015)
25. Zhang, Y.T., Fang, K.T.: Introduction to Multivariate Statistical Analysis. Science Press, Beijing (1982)
26. Zhou, Y.D., Fang, K.T.: FM-representative points. Scientia Sinica Mathematica **49**(2) (2019)

Part II
Design of Experiments

Chapter 5
Is a Transformed Low Discrepancy Design Also Low Discrepancy?

Yiou Li, Lulu Kang, and Fred J. Hickernell

Abstract Experimental designs intended to match arbitrary target distributions are typically constructed via a variable transformation of a uniform experimental design. The inverse distribution function is one such transformation. The discrepancy is a measure of how well the empirical distribution of any design matches its target distribution. This chapter addresses the question of whether a variable transformation of a low discrepancy uniform design yields a low discrepancy design for the desired target distribution. The answer depends on the two kernel functions used to define the respective discrepancies. If these kernels satisfy certain conditions, then the answer is yes. However, these conditions may be undesirable for practical reasons. In such a case, the transformation of a low discrepancy uniform design may yield a design with a large discrepancy. We illustrate how this may occur. We also suggest some remedies. One remedy is to ensure that the original uniform design has optimal one-dimensional projections, but this remedy works best if the design is dense, or in other words, the ratio of sample size divided by the dimension of the random variable is relatively large. Another remedy is to use the transformed design as the input to a coordinate-exchange algorithm that optimizes the desired discrepancy, and this works for both dense or sparse designs. The effectiveness of these two remedies is illustrated via simulation.

Y. Li
Depaul University, 2320 N. Kenmore Avenue, Chicago, IL 60614, USA
e-mail: yli139@depaul.edu

L. Kang · F. J. Hickernell (✉)
Illinois Institute of Technology, RE 220, 10 W. 32nd Street, Chicago, IL 60616, USA
e-mail: hickernell@iit.edu

L. Kang
e-mail: lkang2@iit.edu

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_5

5.1 Introduction

Professor Kai-Tai Fang and his collaborators have demonstrated the effectiveness of low discrepancy points as space filling designs [4–6, 11]. They have promoted discrepancy as a quality measure for statistical experimental designs to the statistics, science, and engineering communities [7–10].

Low discrepancy uniform designs, $\mathscr{U} = \{\boldsymbol{u}_i\}_{i=1}^N$, are typically constructed so that their empirical distributions, $F_{\mathscr{U}}$, approximate F_{unif}, the uniform distribution on the unit cube, $(0, 1)^d$. The discrepancy measures the magnitude of $F_{\text{unif}} - F_{\mathscr{U}}$. The uniform design is a commonly used space filling design for computer experiments [5] and can be constructed using JMP® [20].

When the target probability distribution for the design, F, defined over the experimental domain Ω, is *not* the uniform distribution on the unit cube, then the desired design, $\mathscr{X}$, is typically constructed by transforming a low discrepancy uniform design, i.e.,

$$\mathscr{X} = \{\boldsymbol{x}_i\}_{i=1}^N = \{\boldsymbol{\Psi}(\boldsymbol{u}_i)\}_{i=1}^N = \boldsymbol{\Psi}(\mathscr{U}), \qquad \boldsymbol{\Psi} : (0, 1)^d \to \Omega. \tag{5.1}$$

Note that F may differ from F_{unif} because $\Omega \neq (0, 1)^d$ and/or F is non-uniform. A natural transformation, $\boldsymbol{\Psi}(\boldsymbol{u}) = \big(\Psi_1(u_1), \ldots, \Psi_d(u_d)\big)$, when F has independent marginals, is the inverse distribution transformation:

$$\Psi_j(u_j) = F_j^{-1}(u_j), \quad j = 1, \ldots, d, \qquad \text{where } F(\boldsymbol{x}) = F_1(x_1) \cdots F_d(x_d). \tag{5.2}$$

A number of transformation methods for different distributions can be found in [2] and [11, Chap. 1].

This chapter addresses the question of whether the design $\mathscr{X}$ resulting from transformation (5.1) of a low discrepancy design, $\mathscr{U}$, is itself low discrepancy with respect to the target distribution F. In other words,

$$\textit{does small } F_{\text{unif}} - F_{\mathscr{U}} \textit{ imply small } F - F_{\mathscr{X}}? \tag{Q}$$

We show that the answer may be yes or no, depending on how the question is understood. We discuss both cases. For illustrative purposes, we consider the situation where F is the standard multivariate normal distribution, F_{normal}.

In the next section, we define the discrepancy and motivate it from three perspectives. In Sect. 5.3 we give a simple condition under which the answer to (Q) is yes. But, in Sect. 5.4 we show that under more practical assumptions the answer to (Q) is no. An example illustrates what can go wrong. Section 5.5 provides a coordinate exchange algorithm that improves the discrepancy of a candidate design. Simulation results illustrate the performance of this algorithm. We conclude with a brief discussion.

Table 5.1 Three interpretations of the discrepancy

Kernel Interpretation	Discrepancy $D(\mathscr{X}, \nu, K) = D(\mathscr{X}, \varrho, K)$
$K(\boldsymbol{t}, \boldsymbol{x}) = \langle \delta_{\boldsymbol{t}}, \delta_{\boldsymbol{x}} \rangle_{\mathscr{M}}$	$\lVert \nu - \nu_{\mathscr{X}} \rVert_{\mathscr{M}}$
$f(\boldsymbol{x}) = \langle K(\cdot, \boldsymbol{x}), f \rangle_{\mathscr{H}}$	$\sup_{f \in \mathscr{H} : \lVert f \rVert_{\mathscr{H}} \le 1} \left\lvert \int_{\Omega} f(\boldsymbol{x})\, \varrho(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} - \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x}_i) \right\rvert$
$K(\boldsymbol{t}, \boldsymbol{x}) = \mathrm{cov}\big(f(\boldsymbol{t}), f(\boldsymbol{x})\big)$	$\sqrt{\mathbb{E}\left\lvert \int_{\Omega} f(\boldsymbol{x})\, \varrho(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x} - \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x}_i) \right\rvert^2}$

5.2 The Discrepancy

Experimental design theory based on discrepancy assumes an experimental region, Ω, and a target probability distribution, $F : \Omega \to [0, 1]$, which is known a priori. We assume that F has a probability density, ϱ. It is convenient to also work with measures, ν, defined on Ω. If ν is a probability measure, then the associated probability distribution is given by $F(\boldsymbol{x}) = \nu((-\infty, \boldsymbol{x}])$. The Dirac measure, $\delta_{\boldsymbol{x}}$ assigns unit measure to the set $\{\boldsymbol{x}\}$ and zero measure to sets not containing $\boldsymbol{x}$. A design, $\mathscr{X} = \{\boldsymbol{x}_i\}_{i=1}^N$, is a finite set of points with empirical distribution $F_{\mathscr{X}} = N^{-1} \sum_{i=1}^N \mathbb{1}_{(-\infty, \boldsymbol{x}_i]}$ and empirical measure $\nu_{\mathscr{X}} = N^{-1} \sum_{i=1}^N \delta_{\boldsymbol{x}_i}$.

Our notation for discrepancy takes the form of

$$D(F_{\mathscr{X}}, F, K),\ D(\mathscr{X}, F, K),\ D(\mathscr{X}, \varrho, K),\ D(\mathscr{X}, \nu, K),\ D(\nu_{\mathscr{X}}, \nu, K),\ \text{etc.},$$

all of which mean the same thing. The first argument always refers to the design, the second argument always refers to the target, and the third argument is a symmetric, positive definite kernel, which is explained below. We abuse the discrepancy notation because sometimes it is convenient to refer to the design as a set, $\mathscr{X}$, other times by its empirical distribution, $F_{\mathscr{X}}$, and other times by its empirical measure, $\nu_{\mathscr{X}}$. Likewise, sometimes it is convenient to refer the target as a probability measure, ν, other times by its distribution function, F, and other times by its density function, ϱ.

In the remainder of this section we provide three interpretations of the discrepancy, summarized in Table 5.1. These results are presented in various places, including [14, 15]. One interpretation of discrepancy is the norm of $\nu - \nu_{\mathscr{X}}$. The second and third interpretations consider the problem of evaluating the mean of a random variable $Y = f(\boldsymbol{X})$, or equivalently a multidimensional integral

$$\mu = \mathbb{E}(Y) = \mathbb{E}[f(\boldsymbol{X})] = \int_{\Omega} f(\boldsymbol{x})\, \varrho(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}, \tag{5.3}$$

where $\boldsymbol{X}$ is a random vector with density ϱ. The second interpretation of the discrepancy is worst-case cubature error for integrands, f, in the unit ball of a Hilbert space.

The third interpretation is the root mean squared cubature error for integrands, f, which are realizations of a stochastic processes.

5.2.1 Definition in Terms of a Norm on a Hilbert Space of Measures

Let $(\mathscr{M}, \langle\cdot,\cdot\rangle_{\mathscr{M}})$ be a Hilbert space of measures defined on the experimental region, Ω. Assume that $\mathscr{M}$ includes all Dirac measures. Define the kernel function $K : \Omega \times \Omega \to \mathbb{R}$ in terms of inner products of Dirac measures:

$$K(\boldsymbol{t}, \boldsymbol{x}) := \langle \delta_{\boldsymbol{t}}, \delta_{\boldsymbol{x}} \rangle_{\mathscr{M}}, \qquad \forall \boldsymbol{t}, \boldsymbol{x} \in \Omega. \tag{5.4}$$

The squared distance between two Dirac measures in $\mathscr{M}$ is then

$$\|\delta_{\boldsymbol{x}} - \delta_{\boldsymbol{t}}\|^2_{\mathscr{M}} = K(\boldsymbol{t}, \boldsymbol{t}) - 2K(\boldsymbol{t}, \boldsymbol{x}) + K(\boldsymbol{x}, \boldsymbol{x}), \qquad \forall \boldsymbol{t}, \boldsymbol{x} \in \Omega. \tag{5.5}$$

It is straightforward to show that K is symmetric in its arguments and positive-definite, namely:

$$K(\boldsymbol{x}, \boldsymbol{t}) = K(\boldsymbol{t}, \boldsymbol{x}) \qquad \forall \boldsymbol{t}, \boldsymbol{x} \in \Omega, \tag{5.6a}$$

$$\sum_{i,k=1}^{N} c_i c_k K(\boldsymbol{x}_i, \boldsymbol{x}_k) > 0, \qquad \forall N \in \mathbb{N},\ \boldsymbol{c} \in \mathbb{R}^N \setminus \{\boldsymbol{0}\},\ \mathscr{X} \subset \Omega. \tag{5.6b}$$

The inner product of arbitrary measures $\lambda, \nu \in \mathscr{M}$ can be expressed in terms of a double integral of the kernel, K:

$$\langle \lambda, \nu \rangle_{\mathscr{M}} = \int_{\Omega \times \Omega} K(\boldsymbol{t}, \boldsymbol{x})\, \lambda(\mathrm{d}\boldsymbol{t}) \nu(\mathrm{d}\boldsymbol{x}). \tag{5.7}$$

This can be established directly from (5.4) for $\mathscr{M}_0$, the vector space spanned by all Dirac measures. Letting $\mathscr{M}$ be the closure of the pre-Hilbert space $\mathscr{M}_0$ then yields (5.7).

The discrepancy of the design $\mathscr{X}$ with respect to the target probability measure ν using the kernel K can be defined as the norm of the difference between the target probability measure, ν, and the empirical probability measure for $\mathscr{X}$:

$$
\begin{aligned}
D^2(\mathscr{X}, \nu, K) &:= \|\nu - \nu_{\mathscr{X}}\|^2 \\
&= \int_{\Omega\times\Omega} K(\boldsymbol{t}, \boldsymbol{x})\,(\nu - \nu_{\mathscr{X}})(\mathrm{d}\boldsymbol{t})(\nu - \nu_{\mathscr{X}})(\mathrm{d}\boldsymbol{x}) \\
&= \int_{\Omega\times\Omega} K(\boldsymbol{t}, \boldsymbol{x})\,\nu(\mathrm{d}\boldsymbol{t})\nu(\mathrm{d}\boldsymbol{x}) - \frac{2}{N}\sum_{i=1}^{N}\int_{\Omega} K(\boldsymbol{t}, \boldsymbol{x}_i)\,\nu(\mathrm{d}\boldsymbol{t}) \\
&\qquad + \frac{1}{N^2}\sum_{i,k=1}^{N} K(\boldsymbol{x}_i, \boldsymbol{x}_k). \qquad (5.8a)
\end{aligned}
$$

The formula for the discrepancy may be written equivalently in terms of the probability distribution, F, or the probability density, ϱ, corresponding to the target probability measure, ν:

$$
\begin{aligned}
D^2(\mathscr{X}, F, K) &= \int_{\Omega\times\Omega} K(\boldsymbol{t}, \boldsymbol{x})\,\mathrm{d}F(\boldsymbol{t})\mathrm{d}F(\boldsymbol{x}) - \frac{2}{N}\sum_{i=1}^{N}\int_{\Omega} K(\boldsymbol{t}, \boldsymbol{x}_i)\,\mathrm{d}F(\boldsymbol{t}) \\
&\qquad + \frac{1}{N^2}\sum_{i,k=1}^{N} K(\boldsymbol{x}_i, \boldsymbol{x}_k), \qquad (5.8b) \\
&= \int_{\Omega\times\Omega} K(\boldsymbol{t}, \boldsymbol{x})\,\varrho(\boldsymbol{t})\varrho(\boldsymbol{x})\,\mathrm{d}\boldsymbol{t}\mathrm{d}\boldsymbol{x} - \frac{2}{N}\sum_{i=1}^{N}\int_{\Omega} K(\boldsymbol{t}, \boldsymbol{x}_i)\,\varrho(\boldsymbol{t})\,\mathrm{d}\boldsymbol{t} \\
&\qquad + \frac{1}{N^2}\sum_{i,k=1}^{N} K(\boldsymbol{x}_i, \boldsymbol{x}_k). \qquad (5.8c)
\end{aligned}
$$

Typically the computational cost of evaluating $K(\boldsymbol{t}, \boldsymbol{x})$ for any $(\boldsymbol{t}, \boldsymbol{x}) \in \Omega^2$ is $\mathscr{O}(d)$, where $\boldsymbol{t}$ is a d-vector. Assuming that the integrals above can be evaluated at a cost of $\mathscr{O}(d)$, the computational cost of evaluating $D(\mathscr{X}, \nu, K)$ is $\mathscr{O}(dN^2)$.

The formulas for the discrepancy in (5.8) depend inherently on the choice of the kernel K. That choice is key to answering question (Q). An often used kernel is

$$
K(\boldsymbol{t}, \boldsymbol{x}) = \prod_{j=1}^{d}\left[1 + \frac{1}{2}\left(|t_j| + |x_j| - |x_j - t_j|\right)\right]. \qquad (5.9)
$$

This kernel is plotted in Fig. 5.1 for $d = 1$. The distance between two Dirac measures by (5.5) for this kernel in one dimension is

$$
\|\delta_x - \delta_t\|_{\mathscr{M}} = \sqrt{|x - t|}.
$$

The discrepancy for the uniform distribution on the unit cube defined in terms of the above kernel is expressed as

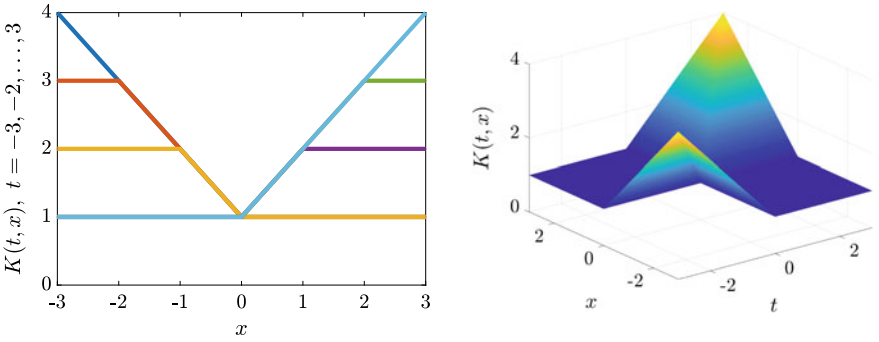


Fig. 5.1 The kernel defined in (5.9) for $d = 1$

$$
\begin{aligned}
D^2(\mathscr{U}, F_{\text{unif}}, K) &= \int_{(0,1)^d \times (0,1)^d} K(\boldsymbol{t}, \boldsymbol{x})\, \mathrm{d}\boldsymbol{t}\mathrm{d}\boldsymbol{x} - \frac{2}{N} \sum_{i=1}^{N} \int_{(0,1)^d} K(\boldsymbol{t}, \boldsymbol{u}_i)\, \mathrm{d}\boldsymbol{t} \\
&\quad + \frac{1}{N^2} \sum_{i,k=1}^{N} K(\boldsymbol{u}_i, \boldsymbol{u}_k) \\
&= \left(\frac{4}{3}\right)^d - \frac{2}{N} \sum_{i=1}^{N} \prod_{j=1}^{d} \left[1 + u_{ij} - \frac{u_{ij}^2}{2}\right] \\
&\quad + \frac{1}{N^2} \sum_{i,k=1}^{N} \prod_{j=1}^{d} \left[1 + \min(u_{ij}, u_{ik})\right].
\end{aligned}
$$

5.2.2 *Definition in Terms of a Deterministic Cubature Error Bound*

Now let $(\mathscr{H}, \langle \cdot, \cdot \rangle_{\mathscr{H}})$ be a reproducing kernel Hilbert space (RKHS) of functions [1], $f : \Omega \to \mathbb{R}$, which appear as the integrand in (5.3). By definition, the reproducing kernel, K, is the unique function defined on $\Omega \times \Omega$ with the properties that $K(\cdot, \boldsymbol{x}) \in \mathscr{H}$ for any $\boldsymbol{x} \in \Omega$ and $f(\boldsymbol{x}) = \langle K(\cdot, \boldsymbol{x}), f \rangle_{\mathscr{H}}$. This second property, implies that K reproduces function values via the inner product. It can be verified that K is symmetric in its arguments and positive definite as in (5.6).

The integral $\mu = \int_{\Omega} f(\boldsymbol{x})\, \varrho(\boldsymbol{x})\, \mathrm{d}\boldsymbol{x}$, which was identified as $\mathbb{E}[f(\boldsymbol{X})]$ in (5.3), can be approximated by a sample mean:

$$
\widehat{\mu} = \frac{1}{N} \sum_{i=1}^{N} f(\boldsymbol{x}_i). \tag{5.10}
$$

The quality of this approximation to the integral, i.e., this cubature, depends in part on how well the empirical distribution of the design, $\mathscr{X} = \{\boldsymbol{x}_i\}_{i=1}^N$, matches the target distribution F associated with the density function ϱ.

Define the cubature error as

$$\begin{aligned}\mathrm{err}(f, \mathscr{X}) = \mu - \widehat{\mu} &= \int_\Omega f(\boldsymbol{x})\, \varrho(\boldsymbol{x}) \mathrm{d}\boldsymbol{x} - \frac{1}{N} \sum_{i=1}^N f(\boldsymbol{x}_i) \\ &= \int_\Omega f(\boldsymbol{x})\, \mathrm{d}[F(\boldsymbol{x}) - F_{\mathscr{X}}(\boldsymbol{x})]. \end{aligned} \tag{5.11}$$

Under modest assumptions on the reproducing kernel, $\mathrm{err}(\cdot, \mathscr{X})$ is a bounded, linear functional. By the Riesz representation theorem, there exists a unique representer, $\xi \in \mathscr{H}$, such that

$$\mathrm{err}(f, \mathscr{X}) = \langle \xi, f \rangle_{\mathscr{H}}, \quad \forall f \in \mathscr{H}.$$

The reproducing kernel allows us to write down an explicit formula for that representer, namely, $\xi(\boldsymbol{x}) = \langle K(\cdot, \boldsymbol{x}), \xi \rangle_{\mathscr{H}} = \langle \xi, K(\cdot, \boldsymbol{x}) \rangle_{\mathscr{H}} = \mathrm{err}(K(\cdot, \boldsymbol{x}), \mathscr{X})$. By the Cauchy-Schwarz inequality, there is a tight bound on the squared cubature error, namely

$$|\mathrm{err}(f, \mathscr{X})|^2 = \langle \xi, f \rangle_{\mathscr{H}}^2 \le \|\xi\|_{\mathscr{H}}^2 \, \|f\|_{\mathscr{H}}^2 . \tag{5.12}$$

The first term on the right describes the contribution made by the quality of the cubature rule, while the second term describes the contribution to the cubature error made by the nature of the integrand.

The square norm of the representer of the error functional is

$$\begin{aligned}\|\xi\|_{\mathscr{H}}^2 &= \langle \xi, \xi \rangle_{\mathscr{H}} = \mathrm{err}(\xi, \mathscr{X}) \quad \text{since } \xi \text{ represents the error functional} \\ &= \mathrm{err}(\mathrm{err}(K(\cdot, \cdot\cdot), \mathscr{X}), \mathscr{X}) \quad \text{since } \xi(\boldsymbol{x}) = \mathrm{err}(K(\cdot, \boldsymbol{x}), \mathscr{X}) \\ &= \int_{\Omega \times \Omega} K(\boldsymbol{t}, \boldsymbol{x})\, \mathrm{d}[F(\boldsymbol{t}) - F_{\mathscr{X}}(\boldsymbol{t})] \mathrm{d}[F(\boldsymbol{x}) - F_{\mathscr{X}}(\boldsymbol{x})]. \end{aligned}$$

We can equate this formula for $\|\xi\|_{\mathscr{H}}^2$ with the formula for $D^2(\mathscr{X}, F, K)$ in (5.8). Thus, the tight, worst-case cubature error bound in (5.12) can be written in terms of the discrepancy as

$$|\mathrm{err}(f, \mathscr{X})| \le \|f\|_{\mathscr{H}} \, D(\mathscr{X}, F, K).$$

This implies our second interpretation of the discrepancy in Table 5.1.

We now identify the RKHS for the kernel K defined in (5.9). Let $(\boldsymbol{a}, \boldsymbol{b})$ be some d dimensional box containing the origin in the interior or on the boundary. For any $\mathfrak{u} \subseteq \{1, \ldots, d\}$, define $\partial^{\mathfrak{u}} f(\boldsymbol{x}_{\mathfrak{u}}) := \partial^{|\mathfrak{u}|} f(\boldsymbol{x}_{\mathfrak{u}}, \boldsymbol{0}) / \partial \boldsymbol{x}_{\mathfrak{u}}$, the mixed first-order partial derivative of f with respect to the x_j for $j \in \mathfrak{u}$, while setting $x_j = 0$ for all $j \notin \mathfrak{u}$. Here, $\boldsymbol{x}_{\mathfrak{u}} = (x_j)_{j \in \mathfrak{u}}$, and $|\mathfrak{u}|$ denotes the cardinality of $\mathfrak{u}$. By convention, $\partial^{\emptyset} f := f(\boldsymbol{0})$. The inner product for the reproducing kernel K defined in (5.9) is defined as

$$\begin{aligned}\langle f, g\rangle_{\mathscr{H}} &:= \sum_{\mathfrak{u}\subseteq\{1,\ldots,d\}} \int_{(\boldsymbol{a},\boldsymbol{b})} \partial^{\mathfrak{u}} f(\boldsymbol{x}_{\mathfrak{u}})\partial^{\mathfrak{u}} g(\boldsymbol{x}_{\mathfrak{u}})\,\mathrm{d}\boldsymbol{x}_{\mathfrak{u}} \\ &= f(\mathbf{0})g(\mathbf{0}) + \int_{a_1}^{b_1} \partial^{\{1\}} f(x_1)\partial^{\{1\}} g(x_1)\,\mathrm{d}x_1 \\ &\quad + \int_{a_2}^{b_2} \partial^{\{2\}} f(x_2)\partial^{\{2\}} g(x_2)\,\mathrm{d}x_2 + \cdots \\ &\quad + \int_{a_2}^{b_2}\int_{a_1}^{b_1} \partial^{\{1,2\}} f(x_1, x_2)\partial^{\{1,2\}} g(x_1, x_2)\,\mathrm{d}x_1\mathrm{d}x_2 + \cdots \\ &\quad + \int_{(\boldsymbol{a},\boldsymbol{b})} \partial^{\{1,\ldots,d\}} f(\boldsymbol{x})\partial^{\{1,\ldots,d\}} g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}.\end{aligned} \tag{5.13}$$

To establish that the inner product defined above corresponds to the reproducing kernel K defined in (5.9), we note that

$$\begin{aligned}\partial^{\mathfrak{u}} K((\boldsymbol{x}_u, \mathbf{0}), \boldsymbol{t}) &= \prod_{j\in\mathfrak{u}} \frac{1}{2}\left[\operatorname{sign}(x_j) - \operatorname{sign}(x_j - t_j)\right] \\ &= \prod_{j\in\mathfrak{u}} \operatorname{sign}(t_j)\mathbb{1}_{(\min(0,t_j),\max(0,t_j))}(x_j).\end{aligned}$$

Thus, $K(\cdot, \boldsymbol{t})$ possesses sufficient regularity to have finite $\mathscr{H}$-norm. Furthermore, K exhibits the reproducing property for the above inner product because

$$\begin{aligned}&\langle K(\cdot, \boldsymbol{t}), f\rangle_{\mathscr{H}} \\ &\quad= \sum_{\mathfrak{u}\subseteq\{1,\ldots,d\}} \int_{(\boldsymbol{a},\boldsymbol{b})} \partial^{\mathfrak{u}} K((\boldsymbol{x}_u, \mathbf{0}), \boldsymbol{t})\partial^{\mathfrak{u}} f(\boldsymbol{x}_u, \mathbf{0})\,\mathrm{d}\boldsymbol{x}_u \\ &\quad= \sum_{\mathfrak{u}\subseteq\{1,\ldots,d\}} \int_{(\boldsymbol{a},\boldsymbol{b})} \prod_{j\in\mathfrak{u}} \operatorname{sign}(t_j)\mathbb{1}_{(\min(0,t_j),\max(0,t_j))}(x_j)\partial^{\mathfrak{u}} f(\boldsymbol{x}_u, \mathbf{0})\,\mathrm{d}\boldsymbol{x}_u \\ &\quad= \sum_{\mathfrak{u}\subseteq\{1,\ldots,d\}} \sum_{\mathfrak{v}\subseteq\mathfrak{u}} (-1)^{|\mathfrak{u}|-|\mathfrak{v}|} f(\boldsymbol{t}_{\mathfrak{v}}, \mathbf{0}) = f(\boldsymbol{t}).\end{aligned}$$

5.2.3 *Definition in Terms of the Root Mean Squared Cubature Error*

Assume Ω is a measurable subset in $\mathbb{R}^d$ and F is the target probability distribution defined on Ω as defined earlier. Now, let $f : \Omega \to \mathbb{R}$ be a stochastic process with a constant pointwise mean, i.e.,

$$\mathbb{E}_{f\in\mathscr{A}}[f(\boldsymbol{x})] = m, \qquad \forall \boldsymbol{x} \in \Omega,$$

where $\mathscr{A}$ is the sample space for this stochastic process. Now we interpret K as the *covariance kernel* for the stocastic process:

$$K(\boldsymbol{t}, \boldsymbol{x}) := \mathbb{E}_{f\in\mathscr{A}}\left([f(\boldsymbol{t}) - m][f(\boldsymbol{x}) - m]\right) = \operatorname{cov}(f(\boldsymbol{t}), f(\boldsymbol{x})), \qquad \forall \boldsymbol{t}, \boldsymbol{x} \in \Omega.$$

It is straightforward to show that the kernel function is symmetric and positive definite.

Define the error functional $\operatorname{err}(\cdot, \mathscr{X})$ in the same way as in (5.11). Now, the mean squared error is

$$\begin{aligned}
\mathbb{E}_{f\in\mathscr{A}}[(\operatorname{err}(f, \mathscr{X})]^2 &= \mathbb{E}_{f\in\mathscr{A}}\left\{\int_\Omega f(\boldsymbol{x})\,\mathrm{d}F(\boldsymbol{x}) - \frac{1}{N}\sum_{i=1}^N f(\boldsymbol{x}_i)\right\}^2 \\
&= \mathbb{E}_{f\in\mathscr{A}}\left\{\int_\Omega (f(\boldsymbol{x}) - m)\,\mathrm{d}F(\boldsymbol{x}) - \frac{1}{N}\sum_{i=1}^N (f(\boldsymbol{x}_i) - m)\right\}^2 \\
&= \int_{\Omega^2} \mathbb{E}_{f\in\mathscr{A}}[(f(\boldsymbol{t}) - m)(f(\boldsymbol{x}) - m)]\,\mathrm{d}F(\boldsymbol{t})\mathrm{d}F(\boldsymbol{x}) \\
&\quad - \frac{2}{N}\sum_{i=1}^N \int_\Omega \mathbb{E}_{f\in\mathscr{A}}[(f(\boldsymbol{x}) - m)(f(\boldsymbol{x}_i) - m)]\,\mathrm{d}F(\boldsymbol{x}) \\
&\quad + \frac{1}{N^2}\sum_{i,k=1}^N \mathbb{E}_{f\in\mathscr{A}}[(f(\boldsymbol{x}_i) - m)(f(\boldsymbol{x}_k) - m)] \\
&= \int_{\Omega^2} K(\boldsymbol{t}, \boldsymbol{x})\,\mathrm{d}F(\boldsymbol{t})\mathrm{d}F(\boldsymbol{x}) - \frac{2}{N}\sum_{i=1}^N \int_\Omega K(\boldsymbol{x}, \boldsymbol{x}_i)\,\mathrm{d}F(\boldsymbol{x}) \\
&\quad + \frac{1}{N^2}\sum_{i,k=1}^N K(\boldsymbol{x}_i, \boldsymbol{x}_k).
\end{aligned}$$

Therefore, we can equate the discrepancy $D(\mathscr{X}, F, K)$ defined in (5.8) as the root mean squared error:

$$D(\mathscr{X}, F, K) = \sqrt{\mathbb{E}_{f\in\mathscr{A}}[(\operatorname{err}(f, \mathscr{X})]^2} = \sqrt{\mathbb{E}\left|\int_\Omega f(\boldsymbol{x})\varrho(\boldsymbol{x})\mathrm{d}\boldsymbol{x} - \frac{1}{N}\sum_{i=1}^N f(\boldsymbol{x}_i)\right|^2}.$$

5.3 When a Transformed Low Discrepancy Design Also Has Low Discrepancy

Having motivated the definition of discrepancy in (5.8) from three perspectives, we now turn our attention to question (Q), namely, does a transformation of low discrepancy points with respect to the uniform distribution yield low discrepancy points with respect to the new target distribution. In this section, we show a positive result, yet recognize some qualifications.

Consider some symmetric, positive definite kernel, $K_{\text{unif}} : (0, 1)^d \times (0, 1)^d \to \mathbb{R}$, some uniform design $\mathscr{U}$, some other domain, Ω, some other target distribution, F, and some transformation $\boldsymbol{\Psi} : (0, 1)^d \to \Omega$ as defined in (5.1). Then the squared discrepancy of the uniform design can be expressed according to (5.8) as follows:

$$
\begin{aligned}
&D^2(\mathscr{U}, F_{\text{unif}}, K_{\text{unif}}) \\
&= \int_{(0,1)^d \times (0,1)^d} K_{\text{unif}}(\boldsymbol{u}, \boldsymbol{v})\, \mathrm{d}\boldsymbol{u}\mathrm{d}\boldsymbol{v} - \frac{2}{N} \sum_{i=1}^{N} \int_{\Omega} K_{\text{unif}}(\boldsymbol{u}, \boldsymbol{u}_i)\, \mathrm{d}\boldsymbol{u} \\
&\qquad + \frac{1}{N^2} \sum_{i,k=1}^{N} K_{\text{unif}}(\boldsymbol{u}_i, \boldsymbol{u}_k) \\
&= \int_{\Omega \times \Omega} K_{\text{unif}}(\boldsymbol{\Psi}^{-1}(\boldsymbol{t}), \boldsymbol{\Psi}^{-1}(\boldsymbol{x})) \left| \frac{\partial \boldsymbol{\Psi}^{-1}(\boldsymbol{t})}{\partial \boldsymbol{t}} \right| \left| \frac{\partial \boldsymbol{\Psi}^{-1}(\boldsymbol{x})}{\partial \boldsymbol{x}} \right| \mathrm{d}\boldsymbol{t}\mathrm{d}\boldsymbol{x} \\
&\qquad - \frac{2}{N} \sum_{i=1}^{N} \int_{\Omega} K_{\text{unif}}(\boldsymbol{\Psi}^{-1}(\boldsymbol{t}), \boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i)) \left| \frac{\partial \boldsymbol{\Psi}^{-1}(\boldsymbol{t})}{\partial \boldsymbol{t}} \right| \mathrm{d}\boldsymbol{t} \\
&\qquad + \frac{1}{N^2} \sum_{i,k=1}^{N} K_{\text{unif}}(\boldsymbol{\Psi}^{-1}(\boldsymbol{x}_i), \boldsymbol{\Psi}^{-1}(\boldsymbol{x}_k)) \\
&= D^2(\mathscr{X}, F, K)
\end{aligned}
$$

where the kernel K is defined as

$$
K(\boldsymbol{t}, \boldsymbol{x}) = K_{\text{unif}}(\boldsymbol{\Psi}^{-1}(\boldsymbol{t}), \boldsymbol{\Psi}^{-1}(\boldsymbol{x})), \tag{5.14a}
$$

and provided that the density, ϱ, corresponding to the target distribution, F, satisfies

$$
\varrho(\boldsymbol{x}) = \left| \frac{\partial \boldsymbol{\Psi}^{-1}(\boldsymbol{x})}{\partial \boldsymbol{x}} \right|. \tag{5.14b}
$$

The above argument is summarized in the following theorem.

Theorem 5.1 *Suppose that the design $\mathscr{X}$ is constructed by transforming the design $\mathscr{U}$ according to the transformation (5.1). Also suppose that conditions (5.14) are satisfied. Then $\mathscr{X}$ has the same discrepancy with respect to the target distribution,*

F, defined by the kernel K as does the original design $\mathscr{U}$ with respect to the uniform distribution and defined by the kernel K_{unif}. That is,

$$D(\mathscr{X}, F, K) = D(\mathscr{U}, F_{\text{unif}}, K_{\text{unif}}).$$

As a consequence, under conditions (5.14), question (Q) has a positive answer.

Condition (5.14b) may be easily satisfied. For example, it is automatically satisfied by the inverse cumulative distribution transform (5.2). Condition (5.14a) is simply a matter of definition of the kernel, K, but this definition has consequences. From the perspective of Sect. 5.2.1, changing the kernel from K_{unif} to K means changing the definition of the distance between two Dirac measures. From the perspective of Sect. 5.2.2, changing the kernel from K_{unif} to K means changing the definition of the Hilbert space of integrands, f, in (5.3). From the perspective of Sect. 5.2.3, changing the kernel from K_{unif} to K means changing the definition of the covariance kernel for the integrands, f, in (5.3).

To illustrate this point, consider a cousin of the kernel in (5.9), which places the reference point at $\mathbf{0.5} = (0.5, \ldots, 0.5)$, the center of the unit cube $(0, 1)^d$:

$$\begin{aligned} K_{\text{unif}}(\boldsymbol{u}, \boldsymbol{v}) &= \prod_{j=1}^{d} \left[1 + \frac{1}{2} \left(\left| u_j - 1/2 \right| + \left| v_j - 1/2 \right| - \left| u_j - v_j \right| \right) \right] \qquad (5.15) \\ &= K(\boldsymbol{u} - \mathbf{0.5}, \boldsymbol{v} - \mathbf{0.5}) \qquad \text{for } K \text{ defined in (5.9).} \end{aligned}$$

This kernel defines the centered L^2-discrepancy [13]. Consider the standard multivariate normal distribution, F_{normal}, and choose the inverse normal distribution,

$$\boldsymbol{\Psi}(\boldsymbol{u}) = (\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_d)), \qquad (5.16)$$

where Φ denotes the standard normal distribution function. Then condition (5.14b) is automatically satisfied, and condition (5.14a) is satisfied by defining

$$\begin{aligned} K(\boldsymbol{t}, \boldsymbol{x}) &= K_{\text{unif}}(\boldsymbol{\Psi}^{-1}(\boldsymbol{t}), \boldsymbol{\Psi}^{-1}(\boldsymbol{x})) \\ &= \prod_{j=1}^{d} \Big[1 + \frac{1}{2} \left(\left| \Phi(t_j) - 1/2 \right| + \left| \Phi(x_j) - 1/2 \right| \right. \\ &\qquad \left. - \left| \Phi(t_j) - \Phi(x_j) \right| \right) \Big]. \end{aligned}$$

In one dimension, the distance between two Dirac measures defined using the kernel K_{unif} above is $\|\delta_x - \delta_t\|_{\mathscr{M}} = \sqrt{|x - t|}$, whereas the distance defined using the kernel K above is $\|\delta_x - \delta_t\|_{\mathscr{M}} = \sqrt{|\Phi(x) - \Phi(t)|}$. Under kernel K, the distance between two Dirac measures is bounded, even though the domain of the distribution is unbounded. Such an assumption may be unpalatable.

5.4 Do Transformed Low Discrepancy Points Have Low Discrepancy More Generally

The discussion above indicates that condition (5.14a) can be too restrictive. We would like to compare the discrepancies of designs under kernels that do not satisfy that restriction. In particular, we consider the centered L^2-discrepancy for uniform designs on $(0, 1)^d$ defined by the kernel in (5.15):

$$
\begin{aligned}
&D^2(\mathscr{U}, F_{\text{unif}}, K_{\text{unif}}) \\
&\quad = \left(\frac{13}{12}\right)^d - \frac{2}{N}\sum_{i=1}^{N}\prod_{j=1}^{d}\left[1 + \frac{1}{2}\left(|u_{ij} - 1/2| - |u_{ij} - 1/2|^2\right)\right] \\
&\qquad + \frac{1}{N^2}\sum_{i,k=1}^{N}\prod_{j=1}^{d}\left[1 + \frac{1}{2}\left(|u_{ij} - 1/2| + |u_{kj} - 1/2| - |u_{ij} - u_{kj}|\right)\right],
\end{aligned}
$$

where again, F_{unif} denotes the uniform distribution on $(0, 1)^d$, and $\mathscr{U}$ denotes a design on $(0, 1)^d$

Changing perspectives slightly, if F'_{unif} denotes the uniform distribution on the cube of volume one centered at the origin, $(-0.5, 0.5)^d$, and the design $\mathscr{U}'$ is constructed by subtracting $\mathbf{0.5}$ from each point in the design $\mathscr{U}$:

$$
\mathscr{U}' = \{\boldsymbol{u} - \mathbf{0.5} : \boldsymbol{u} \in \mathscr{U}\}, \tag{5.17}
$$

then

$$
D(\mathscr{U}', F'_{\text{unif}}, K) = D(\mathscr{U}, F_{\text{unif}}, K_{\text{unif}}),
$$

where K is the kernel defined in (5.9).

Recall that the origin is a special point in the definition of the inner product for the Hilbert space with K as its reproducing kernel in (5.13). Therefore, this K from (5.9) is appropriate for defining the discrepancy for target distributions centered at the origin, such as the standard normal distribution, F_{normal}. Such a discrepancy is

$$
\begin{aligned}
&D^2(\mathscr{X}, F_{\text{normal}}, K) = \left(1 + \sqrt{\frac{2}{\pi}}\right)^d \\
&- \frac{2}{N}\sum_{i=1}^{N}\prod_{j=1}^{d}\left[1 + \frac{1}{\sqrt{2\pi}} + \frac{1}{2}|x_{ij}| - x_{ij}\left(\Phi(x_{ij}) - \frac{1}{2}\right) - \phi(x_{ij})\right] \\
&+ \frac{1}{N^2}\sum_{i,k=1}^{N}\prod_{j=1}^{d}\left[1 + \frac{1}{2}\left(|x_{ij}| + |x_{kj}| - |x_{ij} - x_{kj}|\right)\right].
\end{aligned} \tag{5.18}
$$

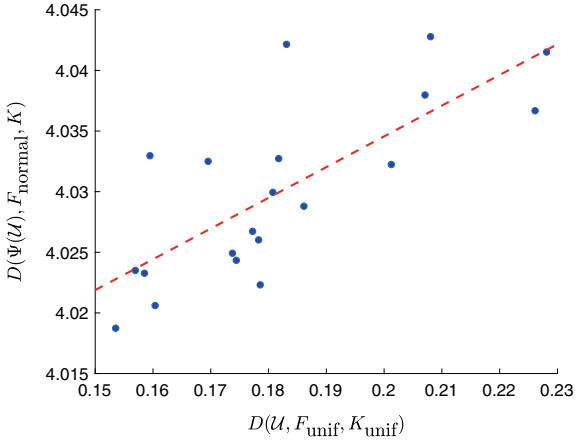


Fig. 5.2 Normal discrepancy versus uniform discrepancy for transformed designs

Here, ϕ is the standard normal probability density function. The derivation of (5.18) is given in the Appendix.

We numerically compare the discrepancy of a uniform design, $\mathscr{U}'$ given by (5.17) and the discrepancy of a design constructed by the inverse normal transformation, i.e., $\mathscr{X} = \boldsymbol{\Psi}(\mathscr{U})$ for $\boldsymbol{\Psi}$ in (5.16), where the $\mathscr{U}$ leading to both $\mathscr{U}'$ and $\mathscr{X}$ is identical. We do not expect the magnitudes of the discrepancies to be the same, but we ask

$$\text{Does } D(\mathscr{U}_1', F_{\text{unif}}', K) \le D(\mathscr{U}_2', F_{\text{unif}}', K) \qquad (\text{Q}')$$
$$\text{imply } D(\boldsymbol{\Psi}(\mathscr{U}_1), F_{\text{normal}}, K) \le D(\boldsymbol{\Psi}(\mathscr{U}_2), F_{\text{normal}}, K)?$$

Again, K is given by (5.9). So we are actually comparing discrepancies defined by the same kernels, but *not kernels that satisfy* (5.14a).

Let $d = 5$ and $N = 50$. We generate $B = 20$ independent and identically distributed (IID) uniform designs, $\mathscr{U}$ with $N = 50$ points on $(0, 1)^5$ and then use the inverse distribution transformation to obtain IID random $N(\mathbf{0}, \mathsf{I}_5)$ designs, $\mathscr{X} = \boldsymbol{\Psi}(\mathscr{U})$. Figure 5.2 plots the discrepancies for normal designs, $D(\boldsymbol{\Psi}(\mathscr{U}), F_{\text{normal}}, K)$, against the discrepancies for the uniform designs, $D(\mathscr{U}, F_{\text{unif}}, K_{\text{unif}}) = D(\mathscr{U}', F_{\text{unif}}', K)$ for each of the $B = 20$ designs. Question (Q′) has a positive answer if and only if the lines passing through any two points on this plot all have non-negative slopes. However, that is not the case. Thus (Q′) has a negative answer.

We further investigate the relationship between the discrepancy of a uniform design and the discrepancy of the same design after inverse normal transformation. Varying the dimension d from 1 to 10, we calculate the sample correlation between $D(\boldsymbol{\Psi}(\mathscr{U}), F_{\text{normal}}, K)$ and $D(\mathscr{U}, F_{\text{unif}}, K_{\text{unif}}) = D(\mathscr{U}', F_{\text{unif}}', K)$ for $B = 500$ IID designs of size $N = 50$. Figure 5.3 displays the correlation as a function of d. Although the correlation is positive, it degrades with increasing d.

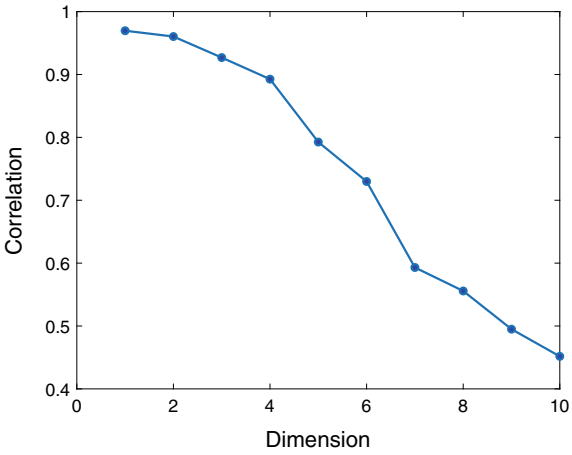


Fig. 5.3 Correlation between the uniform and normal discrepancies for different dimensions

Example 5.1 A simple cubature example illustrates that an inverse transformed low discrepancy design, $\mathscr{U}$, may yield a large $D(\boldsymbol{\Psi}(\mathscr{U}), F_{\text{normal}}, K)$ and also a large cubature error. Consider the integration problem in (5.3) with

$$\boldsymbol{X} \sim N(\mathbf{0}, \mathrm{I}_d), \qquad f(\boldsymbol{x}) = \frac{x_1^2 + \cdots + x_d^2}{1 + 10^{-8}(x_1^2 + \cdots + x_d^2)}, \qquad Y = f(\boldsymbol{X}), \tag{5.19a}$$

$$\mu = \mathbb{E}(Y) = \int_{\mathbb{R}^d} \frac{x_1^2 + \cdots + x_d^2}{1 + 10^{-8}(x_1^2 + \cdots + x_d^2)} \phi(\boldsymbol{x}) \, \mathrm{d}\boldsymbol{x}, \tag{5.19b}$$

where ϕ is the probability density function for the standard multivariate normal distribution. The function $f : \mathbb{R}^d \to \mathbb{R}$ is constructed to asymptote to a constant as $[\|2\|]\boldsymbol{x}$ tends to infinity to ensure that f lies inside the Hilbert space corresponding to the kernel K defined in (5.9). Since the integrand in (5.19) is a function of $[\|2\|]\boldsymbol{x}$, μ can be written as a one dimensional integral. For $d = 10$, $\mu = 10$ to at least 15 significant digits using quadrature.

We can also approximate the integral in (5.19) using a $d = 10$, $N = 512$ cubature (5.10). We compare cubatures using two designs. The design $\mathscr{X}_1$ is the inverse normal transformation of a scrambled Sobol' sequence, $\mathscr{U}_1$, which has a low discrepancy with respect to the uniform distribution on the d-dimensional unit cube. The design $\mathscr{U}_2$ takes the point in $\mathscr{U}_1$ that is closet to $\mathbf{0}$ and moves it to $\left(10^{-15}, \ldots, 10^{-15}\right)$, which is very close to $\mathbf{0}$. As seen in Table 5.2, the two uniform designs have quite similar, small discrepancies. However, the transformed designs, $\mathscr{X}_j = \boldsymbol{\Psi}(\mathscr{U}_j)$ for $j = 1, 2$, have much different discrepancies with respect to the normal distribution. This is due to the point in $\mathscr{X}_2$ that has large negative coordinates. Furthermore, the cubatures,

Table 5.2 Comparison of Integral Estimate

$\mathscr{U}$	$D(\mathscr{U}, F_{\text{unif}}, K)$	$D(\boldsymbol{\Psi}(\mathscr{U}), F_{\text{normal}}, K)$	$\widehat{\mu}$	Relative error
$\mathscr{U}_1$	0.0285	18.57	10.0182	0.0018
$\mathscr{U}_2$	0.0292	58.82	11.2238	0.1224

$\widehat{\mu}$, based on these two designs have significantly different errors. The first design has both a smaller discrepancy and a smaller cubature error than the second. This could not have been inferred by looking at the discrepancies of the original uniform designs.

5.5 Improvement by the Coordinate-Exchange Method

In this section, we propose an efficient algorithm that improves a design's quality in terms of the discrepancy for the target distribution. We start with a low discrepancy uniform design, such as a Sobol' sequence, and transform it into a design that approximates the target distribution. Following the optimal design approach, we then apply a coordinate-exchange algorithm to further improve the discrepancy of the design.

The coordinate-exchange algorithm was introduced in [18], and then applied widely to construct various kinds of optimal designs [16, 19, 21]. The coordinate-exchange algorithm is an iterative method. It finds the "worst" coordinate x_{ij} of the current design and replaces it to decrease loss function, in this case, the discrepancy. The most appealing advantage of the coordinate-exchange algorithm is that at each step one need only solve a univariate optimization problem.

First, we define the point deletion function, $\mathfrak{d}_p$, as the change in square discrepancy resulting from removing the a point from the design:

$$\mathfrak{d}_p(i) = D^2(\mathscr{X}) - \left(\frac{N-1}{N}\right)^2 D^2(\mathscr{X}\backslash\{\boldsymbol{x}_i\}). \tag{5.20}$$

Here, the design $\mathscr{X}\backslash\{\boldsymbol{x}_i\}$ is the $N-1$ point design with the point $\{\boldsymbol{x}_i\}$ removed. We suppress the choice of target distribution and kernel in the above discrepancy notation for simplicity. We then choose

$$i^* = \text{argmax}_{i=1,\ldots,N}\mathfrak{d}_p(i).$$

The definition of i^* means that removing $\boldsymbol{x}_{i^*}$ from the design $\mathscr{X}$ results in the smallest discrepancy among all possible deletions. Thus, $\boldsymbol{x}_{i^*}$ is helping the least, which makes it a prime candidate for modification.

Next, we define a coordinate deletion function, $\mathfrak{d}_c$, as the change in the square discrepancy resulting from removing a coordinate in our calculation of the discrepancy:

$$\mathfrak{d}_c(j) = D^2(\mathscr{X}) - D^2(\mathscr{X}_{-j}). \tag{5.21}$$

Here, the design $\mathscr{X}_{-j}$ still has N points but now only d dimensions, the jth coordinate having been removed. For this calculation to be feasible, the target distribution must have independent marginals. Also, the kernel must be of product form. To simplify the derivation, we assume a somewhat stronger condition, namely that the marginals are identical and that each term in the product defining the kernel is the same for every coordinate:

$$\Omega = \widetilde{\Omega} \times \cdots \times \widetilde{\Omega}, \qquad K(\boldsymbol{t}, \boldsymbol{x}) = \prod_{j=1}^{d} [1 + \widetilde{K}(t_j, x_j)], \qquad \widetilde{K} : \widetilde{\Omega} \times \widetilde{\Omega} \to \mathbb{R}. \tag{5.22}$$

We then choose

$$j^* = \text{argmax}_{j=1,\ldots,d} \mathfrak{d}_c(j).$$

For reasons analogous to those given above, the jth coordinate seems to be the best candidate for change.

Let $\mathscr{X}^*(x)$ denote the design that results from replacing $x_{i^*j^*}$ by x. We now define $\Delta(x)$ as improvement in the squared discrepancy resulting from replacing $\mathscr{X}$ by $\mathscr{X}^*(x)$:

$$\Delta(x) = D^2(\mathscr{X}) - D^2(\mathscr{X}^*(x)). \tag{5.23}$$

We can reduce the discrepancy by find an x such that $\Delta(x)$ is positive. The coordinate-exchange algorithm outlined in Algorithm 1 improves the design by maximizing $\Delta(x)$ for one chosen coordinate in one iteration. The algorithm terminates when it exhausts the maximum allowed number of iterations or the optimal improvement $\Delta(x^*)$ is so small that it becomes negligible ($\Delta(x^*) \leq \text{TOL}$). Algorithm 1 is a greedy algorithm, and thus it can stop at a local optimal design. We recommend multiple runs of the algorithm with different initial designs to obtain a design with the lowest discrepancy possible. Alternatively, users can include stochasticity in the choice of the coordinate that is to be exchanged, similarly to [16].

For kernels of product form, (5.22), and target distributions with independent and identical marginals, the formula for the squared discrepancy in (5.8) becomes

Algorithm 1 Coordinate Exchange Algorithm.

Input: An initial design $\mathscr{X}$ on the domain Ω, a target distribution, F, a kernel, K of the form (5.22), a small value TOL to determine the convergence of the algorithm, and the maximum allowed number of iterations, $M_{\max}$.

Output: Low discrepancy design $\mathscr{X}$.

1: **for** $m = 1, 2, \ldots, M_{\max}$ **do**
2: Compute the point deletion function $\mathfrak{d}_p(1), \ldots, \mathfrak{d}_p(N)$. Choose the i^*-th point which has the largest point deletion value, i.e. $i^* = \text{argmax}_i \mathfrak{d}_p(i)$.
3: Compute the coordinate deletion function $\mathfrak{d}_c(1), \ldots, \mathfrak{d}_c(d)$ and choose the j^*-th coordinate which has the largest coordinate deletion value, i.e., $j^* = \text{argmax}_j \mathfrak{d}_c(j)$.
4: Replace the coordinate $x_{i^*j^*}$ by x^* which is defined by the univariate optimization problem

$$x^* = \text{argmax}_{x \in \widetilde{\Omega}} \Delta(x).$$

5: **if** $\Delta(x^*) >$ TOL **then**
6: Replace $x_{i^*j^*}$ with x^* in the design $\mathscr{X}$, i.e., let $\mathscr{X}(x^*)$ replace the old $\mathscr{X}$.
7: **else**
8: Terminate the loop.
9: **end if**
10: **end for**
11: Return the design, $\mathscr{X}$, and the discrepancy, $D(\mathscr{X}, F, K)$.

$$D^2(\mathscr{X}, \rho, K) = (1+c)^d - \frac{2}{N} \sum_{i=1}^{N} H(\boldsymbol{x}_i) + \frac{1}{N^2} \sum_{i,k=1}^{N} K(\boldsymbol{x}_i, \boldsymbol{x}_k),$$

where

$$h(x) = \int_{\widetilde{\Omega}} \widetilde{K}(t, x)\, \widetilde{\varrho}(t)\, \mathrm{d}t, \tag{5.24a}$$

$$c = \int_{\widetilde{\Omega} \times \widetilde{\Omega}} \widetilde{K}(t_k, x_k)\, \widetilde{\varrho}(t) \widetilde{\varrho}(x)\, \mathrm{d}t \mathrm{d}x = \int_{\widetilde{\Omega}} h(x)\, \widetilde{\varrho}(x)\, \mathrm{d}x, \tag{5.24b}$$

$$H(\boldsymbol{x}) = \prod_{j=1}^{d} [1 + h(x_j)]. \tag{5.24c}$$

An evaluation of $h(x)$ and $\widetilde{K}(t, x)$ each require $\mathscr{O}(1)$ operations, while an evaluation of $H(\boldsymbol{x})$ and $K(\boldsymbol{t}, \boldsymbol{x})$ each require $\mathscr{O}(d)$ operations. The computation of $D(\mathscr{X}, \rho, K)$ requires $\mathscr{O}(dN^2)$ operations because of the double sum. For a standard multivariate normal target distribution and the kernel defined in (5.9), we have

$$c = \sqrt{\frac{2}{\pi}},$$
$$h(x) = \frac{1}{\sqrt{2\pi}} + \frac{1}{2}|x| - x[\Phi(x) - 1/2] - \phi(x),$$
$$\widetilde{K}(t, x) = \frac{1}{2}(|t| + |x| - |x - t|).$$

The point deletion function defined in (5.20) then can be expressed as

$$\mathfrak{d}_p(i) = \frac{(2N-1)(1+c)^d}{N^2} - \frac{2}{N}\left[\frac{1}{N}\sum_{k=1}^{N} H(\boldsymbol{x}_k) + \left(1 - \frac{1}{N}\right) H(\boldsymbol{x}_i)\right] + \frac{1}{N^2}\left[2\sum_{k=1}^{N} K(\boldsymbol{x}_i, \boldsymbol{x}_j) - K(\boldsymbol{x}_i, \boldsymbol{x}_i)\right].$$

The computational cost for $\mathfrak{d}_p(1), \ldots, \mathfrak{d}_p(N)$ is then $\mathcal{O}(dN^2)$, which is the same order as the cost of the discrepancy of a single design.

The coordinate deletion function defined in (5.21) can be be expressed as

$$\mathfrak{d}_c(j) = (c-1)c^{d-1} - \frac{2}{N}\sum_{i=1}^{N} \frac{h(x_{ij})H(\boldsymbol{x}_i)}{1 + h(x_{ij})} + \frac{1}{N^2}\sum_{i,k=1}^{N} \frac{\widetilde{K}(x_{ij}, x_{kj})K(\boldsymbol{x}_i, \boldsymbol{x}_j)}{1 + \widetilde{K}(x_{ij}, x_{kj})}.$$

The computational cost for $\mathfrak{d}_c(1), \ldots, \mathfrak{d}_p(d)$ is also $\mathcal{O}(dN^2)$, which is the same order as the cost of the discrepancy of a single design.

Finally, the function Δ defined in (5.23) is given by

$$\Delta(x) = -\frac{2\left[h(x_{i^*j^*}) - h(x)\right] H(\boldsymbol{x}_{i^*})}{N[1 + h(x_{i^*j^*})]} + \frac{1}{N^2}\left(2\sum_{\substack{i=1 \\ i\neq i^*}}^{N} \frac{[\widetilde{K}(x_{i^*j^*}, x_{ij^*}) - \widetilde{K}(x, x_{ij^*})]K(\boldsymbol{x}_{i^*}, \boldsymbol{x}_i)}{1 + \widetilde{K}(x_{i^*j^*}, x_{ij^*})} + \frac{[\widetilde{K}(x_{i^*j^*}, x_{i^*j^*}) - \widetilde{K}(x, x)]K(\boldsymbol{x}_{i^*}, \boldsymbol{x}_{i^*})}{1 + \widetilde{K}(x_{i^*j^*}, x_{i^*j^*})}\right)$$

If we drop the terms that are independent of x, then we can maximize the function

$$\Delta'(x) = Ah(x) - \frac{1}{N}\sum_{\substack{i=1 \\ i\neq i^*}}^{N} B_i \widetilde{K}(x, x_{ij^*}) - C\widetilde{K}(x, x)$$

where

$$A = \frac{2H(\boldsymbol{x}_{i^*})}{1 + h(x_{i^*j^*})}, \quad B_i = \frac{2K(\boldsymbol{x}_{i^*}, \boldsymbol{x}_i)}{1 + \widetilde{K}(x_{i^*j^*}, x_{ij^*})}, \quad C = \frac{K(\boldsymbol{x}_{i^*}, \boldsymbol{x}_{i^*})}{N[1 + \widetilde{K}(x_{i^*j^*}, x_{i^*j^*})]}.$$

Note that $A, B_1, \ldots, B_N, C$ only need to be computed once for each iteration of the coordinate exchange algorithm.

Note that the coordinate-exchange algorithm we have developed is a greedy and deterministic algorithm. The coordinate that we choose to make exchange is the one has the largest point and coordinate deletion function values, and we always make the exchange for new coordinate as long as the new optimal coordinate improves the objective function. It is true that such deterministic and greedy algorithm is likely to return a design of whose discrepancy attains a *local* minimum. To overcome this, we can either run the algorithm with multiple random initial designs, or we can combine the coordinate-exchange with stochastic optimization algorithms, such as simulated annealing (SA) [17] or threshold accepting (TA) [12]. For example, we can add a random selection scheme when choosing a coordinate to exchange, and when making the exchange of the coordinates, we can incorporate a random decision to accept the exchange or not. The random decision can follow the SA or TA method. Tuning parameters must be carefully chosen to make the SA or TA method effective. Interested readers can refer to [22] to see how TA can be applied to the minimization of discrepancy.

5.6 Simulation

To demonstrate the performance of the d-dimensional standard normal design proposed in Sect. 5.5, we compare three families of designs: (1) RAND: inverse transformed IID uniform random numbers; (2) SOBOL: inverse transformed Sobol' set; (3) E-SOBOL: inverse transformed scrambled Sobol' set where the one dimensional projections of the Sobol' set have been adjusted to be $\{1/(2N), 3/(2N), \ldots, (2N-1)/(2N)\}$; and (4) CE: improved E-SOBOL via Algorithm 1. We have tried different combinations of dimension, d, and sample size, N. For each (d, N) and each algorithm we generate 500 designs and compute their discrepancies (5.18).

Figure 5.4 contains the boxplots of normal discrepancies corresponding to the four generators with $d = 2$ and $N = 32$. It shows that SOBOL, E-SOBOL, and CE all outperform RAND by a large margin. To better present the comparison between the better generators, in Fig. 5.5 we generally exclude RAND.

We also report the average execution times for the four generators in Table 5.3. All codes were run on a MacBook Pro with 2.4 GHz Intel Core i5 processor. The maximum number of iterations allowed is $M_{\max} = 200$. Algorithm 1 converges within 20 iterations in all simulation examples.

We summarize the results of our simulation as follows.

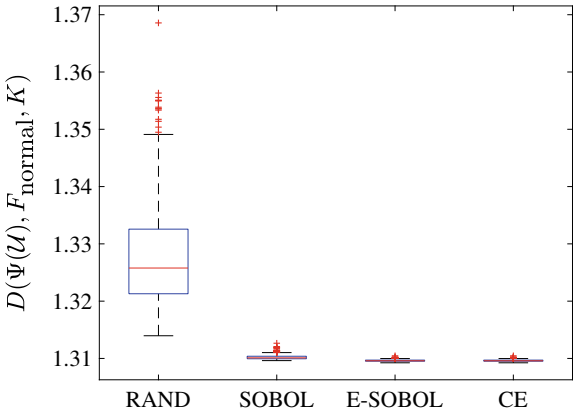


Fig. 5.4 Performance comparison of designs

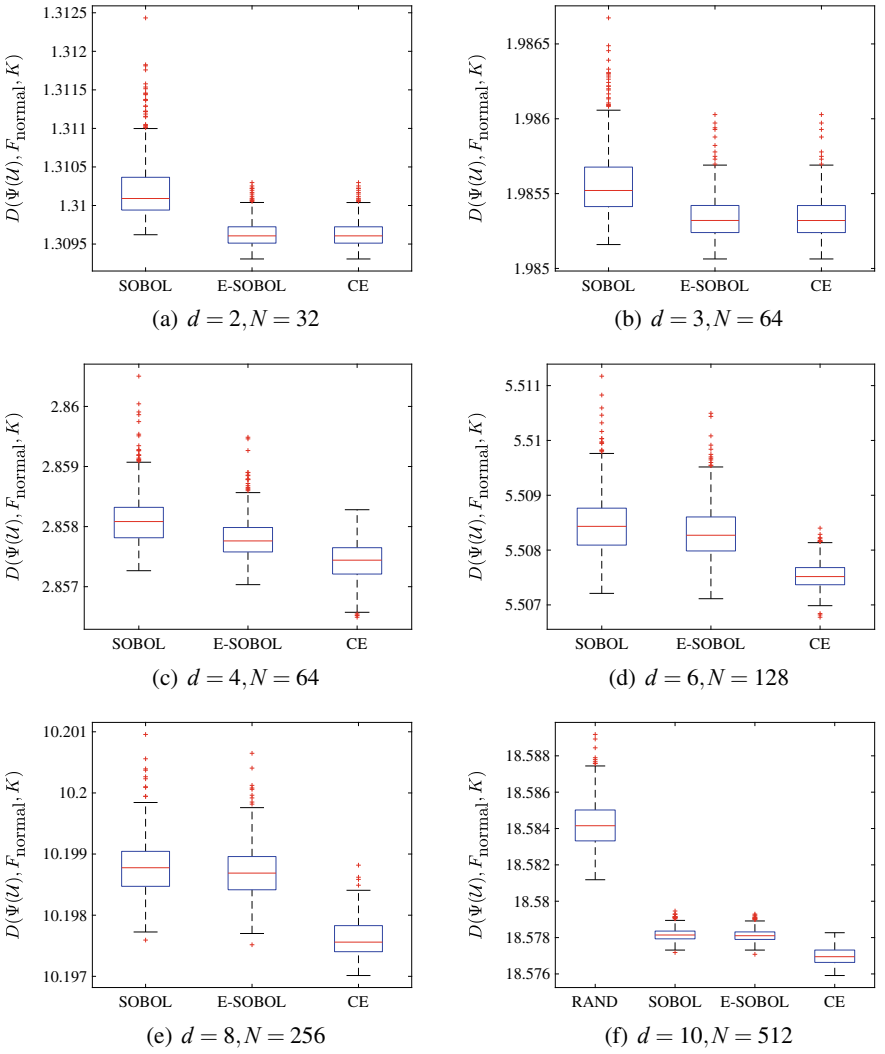
Table 5.3 Execution Time of Generators (in seconds)

d	2	3	4	6	8	10
N	32	64	64	128	256	512
RAND	3.22E−5	5.21E−5	5.27E−5	9.92E−5	2.48E−4	5.32E−4
SOBOL	8.60E−4	0.10E−2	0.11E−2	0.16E−2	0.21E−2	0.28E−2
E-SOBOL	8.71E−4	0.11E−2	0.12E−2	0.16E−2	0.23E−2	0.32E−2
CE	1.34E−2	2.73E−2	6.12E−2	0.24	1.04	3.84

1. Overall, CE produces the smallest discrepancy.
2. When the design is relatively dense, i.e., N/d is large, E-SOBOL and CE have similar performance.
3. When the design is more sparse, i.e., N/d is smaller, SOBOL and E-SOBOL have similar performance, but CE is superior to both of them in terms of the discrepancy. Not only in terms of the mean but also in terms of the *range* for the 500 designs generated.
4. CE requires the longest computational time to construct a design, but this is moderate. When the cost of obtaining function values is substantial, then the cost of constructing the design may be insignificant.

5.7 Discussion

This chapter summarizes the three interpretations of the discrepancy. We show that for kernels and variable transformations satisfying conditions (5.14), variable transformations of low discrepancy uniform designs yield low discrepancy designs with respect to the target distribution. However, for more practical choices of kernels, this correspondence may not hold. The coordinate-exchange algorithm can improve the



(a) $d = 2, N = 32$ (b) $d = 3, N = 64$

(c) $d = 4, N = 64$ (d) $d = 6, N = 128$

(e) $d = 8, N = 256$ (f) $d = 10, N = 512$

Fig. 5.5 Performance comparison of designs

discrepancies of candidate designs that may be constructed by variable transformations.

While discrepancies can be defined for arbitrary kernels, we believe that the choice of kernel can be important, especially for small sample sizes. If the distribution has a symmetry, e.g. $\varrho(\boldsymbol{T}(\boldsymbol{x})) = \varrho(\boldsymbol{x})$ for some probability preserving bijection $\boldsymbol{T} : \Omega \to \Omega$, then we would like our discrepancy to remain unchanged under such a bijection, i.e., $D(\boldsymbol{T}(\mathscr{X}), \varrho, K) = D(\mathscr{X}, \varrho, K)$. This can typically be ensured by choosing kernels satisfying $K(\boldsymbol{T}(\boldsymbol{t}), \boldsymbol{T}(\boldsymbol{x})) = K(\boldsymbol{t}, \boldsymbol{x})$. The kernel K_{unif} defined in

(5.15) satisfies this assumption for the standard uniform distribution and the transformation $\boldsymbol{T}(\boldsymbol{x}) = \mathbf{1} - \boldsymbol{x}$. The kernel K defined in (5.9) satisfies this assumption for the standard normal distribution and the transformation $\boldsymbol{T}(\boldsymbol{x}) = -\boldsymbol{x}$.

For target distributions with independent marginals and kernels of product form as in (5.22), *coordinate weights* [3] Sect. 4 are used to determine which projections of the design, denoted by $\mathfrak{u} \subseteq \{1, \ldots, d\}$, are more important. The product form of the kernel given in (5.22) can be generalized as

$$K_{\boldsymbol{\gamma}}(\boldsymbol{t}, \boldsymbol{x}) = \prod_{j=1}^{d} \left[1 + \gamma_j \widetilde{K}(t_j, x_j)\right].$$

Here, the positive coordinate weights are $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_d)$. The squared discrepancy corresponding to this kernel may then be written as

$$D^2(\mathscr{X}, F, K_{\boldsymbol{\gamma}}) = \sum_{\substack{\mathfrak{u} \subseteq \{1,\ldots,d\} \\ \mathfrak{u} \neq \emptyset}} \gamma_{\mathfrak{u}} D_{\mathfrak{u}}^2(\mathscr{X}, \rho, K), \qquad \gamma_{\mathfrak{u}} = \prod_{j \in \mathfrak{u}} \gamma_j$$

$$D_{\mathfrak{u}}^2(\mathscr{X}_{\mathfrak{u}}, F_{\mathfrak{u}}, K) = c^{|\mathfrak{u}|} - \frac{2}{N} \sum_{i=1}^{N} \prod_{j \in \mathfrak{u}} h(x_{ij}) + \frac{1}{N^2} \sum_{i,k=1}^{N} \prod_{j \in \mathfrak{u}} \widetilde{K}(x_{ij}, x_{kj}),$$

where c and h are defined in (5.24). Here, $\mathscr{X}_{\mathfrak{u}}$ denotes the projection of the design into the coordinates contained in $\mathfrak{u}$, and $F_{\mathfrak{u}} = \prod_{j \in \mathfrak{u}} F_j$ is the $\mathfrak{u}$-marginal distribution. Each discrepancy piece, $D_{\mathfrak{u}}(\mathscr{X}_{\mathfrak{u}}, F_{\mathfrak{u}}, K)$, measures how well the projected design $\mathscr{X}_{\mathfrak{u}}$ matches $F_{\mathfrak{u}}$.

The values of the coordinate weights can be chosen to reflect the user's belief as to the importance of the design matching the target for various coordinate projections. A large value of γ_j relative to the other $\gamma_{j'}$ places more importance on the $D_{\mathfrak{u}}(\mathscr{X}_{\mathfrak{u}}, F_{\mathfrak{u}}, K)$ with $j \in \mathfrak{u}$. Thus, γ_j is an indication of the importance of coordinate j in the definition of $D(\mathscr{X}, F, K_{\boldsymbol{\gamma}})$.

If $\boldsymbol{\gamma}$ is one choice of coordinate weights and $\boldsymbol{\gamma}' = C\boldsymbol{\gamma}$ is another choice of coordinate weights where $C > 1$, then $\gamma'_{\mathfrak{u}} = C^{|\mathfrak{u}|}\gamma_{\mathfrak{u}}$. Thus, $D(\mathscr{X}, F, K_{\boldsymbol{\gamma}'})$ emphasizes the projections corresponding to the $\mathfrak{u}$ with large $|\mathfrak{u}|$, i.e., the higher order effects. Likewise, $D(\mathscr{X}, F, K_{\boldsymbol{\gamma}'})$ places relatively more emphasis lower order effects. Again, the choice of coordinate weights reflects the user's belief as to the relative importance of the design matching the target distribution for lower order effects or higher order effects.

Appendix

We derive the formula in (5.18) for the discrepancy with respect to the standard normal distribution, Φ, using the kernel defined in (5.9). We first consider the case $d = 1$. We integrate the kernel once:

$$\begin{aligned}
&\int_{-\infty}^{\infty} K(t,x)\,\mathrm{d}\Phi(t)\\
&\qquad=\int_{-\infty}^{\infty}\left(1+\frac{1}{2}|x|+\frac{1}{2}|t|-\frac{1}{2}|x-t|\right)\phi(t)\,\mathrm{d}t\\
&\qquad=1+\frac{1}{\sqrt{2\pi}}+\frac{1}{2}|x|-\frac{1}{2}\left[\int_{-\infty}^{x}(x-t)\phi(t)\,\mathrm{d}t+\int_{x}^{\infty}(t-x)\phi(t)\,\mathrm{d}t\right]\\
&\qquad=1+\frac{1}{\sqrt{2\pi}}+\frac{1}{2}|x|-x[\Phi(x)-1/2]-\phi(x).
\end{aligned}$$

Then we integrate once more:

$$\begin{aligned}
&\int_{-\infty}^{\infty}\int_{-\infty}^{\infty} K(t,x)\,\mathrm{d}\Phi(t)\mathrm{d}\Phi(x)\\
&\qquad=\int_{-\infty}^{\infty}\left(1+\frac{1}{\sqrt{2\pi}}+\frac{1}{2}|x|-x[\Phi(x)-1/2]-\phi(x)\right)\phi(x)\,\mathrm{d}x\\
&\qquad=1+\sqrt{\frac{2}{\pi}}+\int_{-\infty}^{\infty}\{-x\Phi(x)\phi(x)+[\phi(x)]^2\}\,\mathrm{d}x\\
&\qquad=1+\sqrt{\frac{2}{\pi}}-\frac{1}{\sqrt{4\pi}}+\int_{-\infty}^{\infty}\frac{1}{2\pi}\mathrm{e}^{-x^2}\mathrm{d}x=1+\sqrt{\frac{2}{\pi}}.
\end{aligned}$$

Generalizing this to the d-dimensional case yields

$$\int_{\mathbb{R}^d\times\mathbb{R}^d} K(\boldsymbol{x},\boldsymbol{t})\,\mathrm{d}\Phi(\boldsymbol{x})\mathrm{d}\Phi(\boldsymbol{t})=\left(1+\sqrt{\frac{2}{\pi}}\right)^d,$$

$$\int_{\mathbb{R}^d} K(\boldsymbol{x},\boldsymbol{x}_n)\,\mathrm{d}\Phi(\boldsymbol{x})=\prod_{j=1}^{d}\left[1+\frac{1}{\sqrt{2\pi}}+\frac{1}{2}|x_j|-x_j[\Phi(x_j)-1/2]-\phi(x_j)\right].$$

Thus, the discrepancy for the normal distribution is

$$\begin{aligned}
&D^2(\mathscr{X},\Phi,K)\\
&=\left(1+\sqrt{\frac{2}{\pi}}\right)^d-\frac{2}{N}\sum_{\boldsymbol{x}\in P}\prod_{j=1}^{d}\left[1+\frac{1}{\sqrt{2\pi}}+\frac{1}{2}|x_j|-x_j[\Phi_1(x_j)-1/2]-\phi(x_j)\right]\\
&\quad+\frac{1}{N^2}\sum_{\boldsymbol{x},\boldsymbol{t}\in P}\prod_{j=1}^{d}\left[1+\frac{1}{2}|x_j|+\frac{1}{2}|t_j|-\frac{1}{2}|x_j-t_j|\right].
\end{aligned}$$

References

1. Aronszajn, N.: Theory of reproducing kernels. Trans. Amer. Math. Soc. **68**, 337–404 (1950)
2. Devroye, L.: Nonuniform Random Variate Generation. Handbooks in Operations Research and Management Science, pp. 83–121 (2006)

3. Dick, J., Kuo, F., Sloan, I.H.: High dimensional integration—the Quasi- Monte Carlo way. Acta Numer. **22**, 133–288 (2013)
4. Fang, K.T., Hickernell, F.J.: Uniform experimental design. Encyclopedia Stat. Qual. Reliab., 2037–2040 (2008)
5. Fang, K.T., Li, R., Sudjianto, A.: Design and Modeling for Computer Experiments. Computer Science and Data Analysis. Chapman & Hall, New York (2006)
6. Fang, K.T., Liu, M.-Q., Qin, H., Zhou, Y.-D.: Theory and Application of Uniform Experimental Designs. Springer Nature (Singapore) and Science Press (Bejing), Mathematics Monograph Series (2018)
7. Fang, K.T., Ma, C.X.: Wrap-around l2-discrepancy of random sampling, latin hypercube and uniform designs. J. Complexity **17**, 608–624 (2001)
8. Fang, K.T. , Ma, C.X.: Relationships Between Uniformity, Aberration and Correlation in Regular Fractions 3s-1. Monte Carlo and quasi-Monte Carlo methods 2000, pp. 213–231 (2002)
9. Fang, K.T., Ma, C.X., Winker, P.: Centered l2-discrepancy of random sampling and latin hypercube design, and construction of uniform designs. Math. Comp. **71**, 275–296 (2002)
10. Fang, K.T., Mukerjee, R.: A connection between uniformity and aberration in regular fractions of two-level factorials. Biometrika **87**, 193–198 (2000)
11. Fang, K.T., Wang, Y.: Number-Theoretic Methods in Statistics. Chapman and Hall, New York (1994)
12. Fang, K.T., Lu, X., Winker, P.: Lower bounds for centered and wrap- around l2-discrepancies and construction of uniform designs by threshold accepting. J. Complex. **19**(5), 692–711 (2003)
13. Hickernell, F.J.: A generalized discrepancy and quadrature error bound. Math. Comp. **67**, 299–322 (1998)
14. Hickernell, F.J.: Goodness-of-fit statistics, discrepancies and robust designs. Statist. Probab. Lett. **44**, 73–78 (1999)
15. Hickernell, F.J.: The trio identity for Quasi-Monte Carlo error. In: MCQMC: International conference on Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing, pp. 3–27 (2016)
16. Kang, L.: Stochastic coordinate-exchange optimal designs with complex constraints. Qual. Eng. **31**(3), 401–416 (2019). https://doi.org/10.1080/08982112.2018.1508695
17. Kirkpatrick, S., Gelatt, C.D., Vecchi, M.P.: Optimization by Simulated Annealing. Science **220**(4598), 671–680 (1983)
18. Meyer, R.K., Nachtsheim, C.J.: The coordinate-exchange algorithm for constructing exact optimal experimental designs. Technometrics **37**(1), 60–69 (1995)
19. Overstall, A.M., Woods, D.C.: Bayesian design of experiments using approximate coordinate exchange. Technometrics **59**(4), 458–470 (2017)
20. Sall, J., Lehman, A., Stephens, M.L., Creighton, L.: JMP Start Statistics: a Guide to Statistics and Data Analysis Using JMP, 6th edn. SAS Institute (2017)
21. Sambo, F., Borrotti, M., Mylona, K.: A coordinate-exchange two-phase local search algorithm for the D-and I-optimal designs of split-plot experiments. Comput. Statist. Data Anal. **71**, 1193–1207 (2014)
22. Winker, P., Fang, K.T.: Application of threshold-accepting to the evaluation of the discrepancy of a set of points. SIAM J. Numer. Anal. **34**(5), 2028–2042 (1997)

Chapter 6
The Construction of Optimal Design for Order-of-Addition Experiment via Threshold Accepting

Peter Winker, Jianbin Chen, and Dennis K. J. Lin

Abstract The objective of the order-of-addition (OofA) experiment is to find the optimal addition order by comparing all responses with different orders. Assuming that the OofA experiment involves $m(\geq 2)$ components, there are $m!$ different orders of adding sequence. When m is large, it is infeasible to compare all $m!$ possible solutions (for example, $10! \approx 3.6$ millions). Two potential construction methods are systematic combinatorial construction and computer algorithmic search. Computer search methods presented in the literature for constructing optimal fractional designs of OofA experiments appear rather simplistic. In this paper, based on the pairwise-order (PWO) model and the tapered PWO model, the threshold accepting algorithm is applied to construct the optimal design (D-efficiency for the present application) with subsets of size n among all possible size $m!$. In practical, the designs obtained by threshold accepting algorithm for $4 \leq m \leq 30$ with $n = m(m-1)/2+1, m(m-1)+1, 3m(m-1)/2+1$ respectively are provided for practical uses. This is apparently the most complete list of order-of-addition (OofA) designs via computer search for $4 \leq m \leq 30$ in the literature. Their efficiencies are illustrated by a scheduling problem.

Keywords D-optimal design · Pair-wise ordering (pwo) mode · Threshold accepting · Tapered pwo model

P. Winker
Justus-Liebig-University Giessen, Licher Str. 64, 35394 Giessen, Germany
e-mail: Peter.Winker@wi.jlug.de

J. Chen
School of Statistics and Data Science & LPMC, Nankai University, Tianjin 300071, China
e-mail: chenjianbinlzu@163.com

D. K. J. Lin (✉)
Department of Statistics, Purdue University, 250 N. University Street,
West Lafayette, IN 47907, USA
e-mail: dkjlin@purdue.edu

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_6

6.1 Introduction

The order-of-addition (OofA) experiment has been popularly used when the response of interest is affected by the addition sequence of materials or components. Considering the addition of m different materials or components into the system, the different responses depend on different adding orders. Each permutation of $\{1, \ldots, m\}$ is a possible adding order, hence there are $m!$ different orders of adding sequences into the system which yield different responses. The OofA experiments are prevalent in many scientific and industrial areas, such as chemistry-related areas, bio-chemistry, food science, nutritional science, and pharmaceutical science.

The purpose of OofA experiment is to find the optimal addition order by comparing all possible responses with different orders. However, it is often infeasible to test all the $m!$ possible orders when m is large (for example, 10! is about 3.6 millions). In practice, a number of randomly selected orders are tested, but the empirical experience indicates that randomly selected orders may not be most informative. Hence the design problem arises to choose a subset of orders for comparison. A good design for the OofA experiments will help experimenters to identify the important order effects, and to find out the optimal addition order with substantially fewer experimental runs. Such an important problem has received a great deal of attention in the past decades. For example [18] considered the design with pair-wise ordering (PWO) effects. Based on the PWO model [19] proposed a number of design criteria and found some OofA designs which have the same correlation structures as the full OofA designs, for small number of components (m). Peng et al. [17] considered different types of optimality criteria and discussed the properties of some fractional designs. Zhao et al. [25] considered the minimal-point OofA designs. Yang et al. [24] has obtained a number of OofA designs called component orthogonal arrays (COAs) that are optimal under their component-position model. [13] reviewed the latest work on the design and model of OofA experiments, and introduced some new thoughts. [1] proposed another type OofA design named pair-wise ordering distance (PWOD) arrays that can be used in any models in the literature. Chen et al. [2] introduced a statistical method to speculate solutions of NP-hard problem involving orders by making use of design for OofA experiment.

This paper makes use of the threshold accepting algorithm to find the best subset of size n which implies searching for the optimal value of the objective function among all $m!$. This threshold accepting?algorithm provides high quality approximations to the global optimum. Therefore, designs obtained by our algorithm, involving only a fraction of all $m!$ possible permutations of components, are powerful for fitting models in terms of the D-efficiency and are efficient for predicting the optimal order. An illustrative example is provided to show the advantages of the obtained designs.

The remaining part of this article is organized as follows. Section 6.2 introduces the PWO model, the tapered PWO model and some optimality criteria. The threshold accepting algorithm is proposed in Sect. 6.3. The optimal fractional OofA designs obtained by the threshold algorithm are provided in Sect. 6.4, and a scheduling example is discussed in Sect. 6.5. Section 6.6 gives some concluding remarks.

6.2 Preliminary

6.2.1 PWO Model

The order-of-addition (OofA) experiment involves $m(\geq 2)$ components, and there are $m!$ different orders of adding sequences into the system to yield different responses. For any pair of components i and j, if the impact of component i preceding j is different from the impact of component j preceding i, such a difference is called the effect of pair (i, j). To express the order effect [18] proposed "pseudo factor". [19] called it pair-wise ordering (PWO) factor. The PWO factor is defined as

$$I_{i,j} = \begin{cases} +1 & if \ \ i \ \ precedes \ \ j, \\ -1 & if \ \ j \ \ precedes \ \ i. \end{cases} \tag{6.1}$$

This indicates whether the component i precedes the component j or not, where i and j are the components. There are $q = \binom{m}{2}$ PWO factors, corresponding to all pairs of component orders. These factors are arranged according to the lexicographic ordering of the components' indices. For illustration, when $m = 4$ and a possible order $\pi = 2143$ is given, we have $I_{12}(\pi) = -1, I_{13}(\pi) = +1, I_{14}(\pi) = +1, I_{23}(\pi) = +1, I_{24}(\pi) = +1$ and $I_{34}(\pi) = -1$. Assuming β_{ij} is the effect to response caused by I_{ij}, the PWO model is the first-order model by summing the effects of all I_{ij}'s, namely:

$$y = \beta_0 + \sum_{i<j} \beta_{ij} I_{ij} + \varepsilon, \tag{6.2}$$

where y is the response of interest, ε is a random error assumed to be independent and to have a normal distribution $N(0, \sigma^2)$, and $p = q + 1$ parameters $\{\beta_0, \beta_{12}, \beta_{13}, \ldots, \beta_{(m-1)m}\}$ should be estimated.

In practice, it is not affordable to test all the $m!$ orders when m is large. Let $\pi = (\pi_1, \ldots, \pi_m)$ be a permutation of $\{1, \ldots, m\}$ which specifies the order. Denote Π as the subset of size n from all of $m!$ possible orders. Based on the best subset Π, the expected PWO model can be written as

$$E(y|\pi) = \beta_0 + \sum_{i<j} \beta_{ij} I_{ij}(\pi), \qquad \pi \in \Pi. \tag{6.3}$$

The PWO model has two merits. Firstly, it is easy to interpret: the effect of β_{ij} shows the difference between the impacts of all the possible orders in which i precedes j and the impacts of all the orders in which j precedes i. Secondly, the PWO model is an economic model which requires a small number of runs ($p = q + 1$) compared with the total number of candidate runs ($m!$). In order to fit PWO model (6.3) for using the smallest runs [25] proposed a special type of design with only $q + 1$ runs (out of

$m!$ runs), which is called a minimal-point OofA design, as long as its D-efficiency is nonzero.

6.2.2 Tapered PWO Model

Although the PWO model is an economical model, the PWO effect has some weaknesses, for example, the PWO effect I_{12} in the sequences "$1 \to 2 \to \ldots$", "$1 \to \ldots \to 2$", "$\ldots \to 1 \to 2 \ldots$" and "$\ldots \to 1 \to \ldots \to 2 \ldots$" is assumed to be the same ($I_{12} = +1$). Obviously, these sequences have different sense between the component 1 and the component 2. It is possible to assume that the impact of any such pairwise order changes with an increase in the distance between the components in the pair in practice. So another model of interest is "tapered PWO model" as considered as in [17].

For any components $i, j (i \neq j)$ and $\pi = (\pi_1, \ldots, \pi_m) \in \Pi$, let $h(ij, \pi)$ be the distance between i and j in π, i.e., $h(ij, \pi) = |k - l|$ if $\pi_k = i$ and $\pi_l = j$, so that $h(ij, \pi) \in \{1, \ldots, m-1\}$. Denote

$$z_{ij} = \begin{cases} c_{h(ij,\pi)} & \textit{if } i \textit{ precedes } j \textit{ in } \pi \\ -c_{h(ij,\pi)} & \textit{if } j \textit{ precedes } i \textit{ in } \pi \end{cases} \tag{6.4}$$

as the "tapered PWO factor", where $c_h = 1/h$ or $c_h = c^{h-1}$ with known $c\,(0 < c < 1)$ for $h \in \{1, \ldots, m-1\}$. Then, the tapered PWO model can be expressed as

$$y = \beta_0 + \sum_{i<j} \beta_{ij} z_{ij} + \varepsilon, \tag{6.5}$$

where y is the response of interest, ε is a random error assumed to be independent and to have a normal distribution $N(0, \sigma^2)$, and β_0 and β_{ij} are the unknown parameters. For any $\pi = (\pi_1, \ldots, \pi_m) \in \Pi$, the corresponding tapered PWO model can be represented as

$$E(y|\pi) = \beta_0 + \sum_{i<j} \beta_{ij} z_{ij}(\pi), \qquad \pi \in \Pi. \tag{6.6}$$

The tapered PWO model is a generalized PWO model, if one chooses $c_h = 1$ for all h, then the tapering PWO factor (6.4) and the tapered PWO model (6.5) respectively become the PWO factor (6.1) and the usual PWO model (6.2) of [18] and [19].

Under the PWO model or the tapered model, this paper makes use of the threshold accepting algorithm to construct the optimal design, for all subsets of size $n\,(\geq q+1)$ among all possible $m!$ order. For simplicity, three particular run sizes of OofA designs are of interest, even though the proposed algorithm is capable for any n. Recall that $q = \binom{m}{2}$.

(a) Minimal-point design: n = q + 1,
(b) Double type design: n = 2 ∗ q + 1,
(c) Triplicate type design: n = 3 ∗ q + 1.

6.2.3 *Some Optimality Criteria*

There are many criteria for constructing optimal designs in literature. Let X denotes the model matrix with respect to the pre-specified model, p is the number of columns of X and n be the run size of X, then, the D-value of a design D is defined by $D_e(D) = \frac{1}{n}|X^TX|^{\frac{1}{p}}$, which is proportional to the generalized variance of the parameter estimators $\widehat{\beta}$ to be minimized. That is, the volume of the confidence ellipsoid for $\widehat{\beta}$ is minimized by maximizing the determinant $det(X^TX)$. A-optimal designs are those designs which minimize the average variance of the estimators $\widehat{\beta}$ and thus the criterion $trace((X^TX)^{-1})$. The E-criterion focus on the minimum eigenvalue of X^TX. We will select all possible choices of designs, by considering which one(s) attain the optimum in terms of these criteria. There are more other optimal design criteria (see, for example [16]).

In this paper, we mainly focus on the D-efficiency under the pre-specified model (PWO model or tapered PWO model) for simplicity. Likewise, the $A-$, $E-$efficiency can be defined. The larger D-efficiency the better, an optimal design has a D-efficiency of 1. Throughout this paper, let D_{full} denote the D-efficiency of the full design. For all other designs, the relative D-efficiency $D_r = D_e/D_{full}$ is used here. For the tapered PWO model, let $q = \binom{m}{2} = m(m-1)/2$ and $p = q + 1$ [17] proposed the D-efficiency of the design D is

$$D_r(D) = \frac{D_e(D)}{\left[\{b_0 + (m-2)b_1\}^{m-1}(b_0 - 2b_1)^{(m-1)(m-2)/2}\right]^{1/p}}, \tag{6.7}$$

where

$$b_0 = 2\{(m-1)c_1^2 + \Lambda + c_{m-1}^2\}/\{m(m-1)\},$$

$$b_1 = 2\Sigma_h\{m - h(1) - h(2)\}c_{h_{(1)}}\{2c_{h_{(1)}+h_{(2)}} - c_{h_{(2)}}\}/\{m(m-1)(m-2)\},$$

and Σ_h denotes the sum over positive integers $h(1), h(2)$ such that $h(1) + h(2) \leq m - 1$. For the PWO model, $c_h = 1$ for all h, $b_0 = 1$ and $b_1 = 1/3$, then the D-efficiency of the design D (6.7) reduces to

$$D_r(D) = D_e(D)/\left[\frac{(m+1)^{m-1}}{3^q}\right]^{1/p}. \tag{6.8}$$

Our goal here is to maximize the objective functions given in (6.7) and (6.8), making use of the threshold accepting algorithm as described below.

6.3 The Threshold Accepting Algorithm

The problem of finding good OofA designs might be interpreted as a complex discrete optimization problem. For a given number of m components, the full design matrix comprises $m!$ rows. For a given objective function, selecting the best subset of size n implies searching for the largest value of the objective function among all subsets of size n of a set of size $m!$, i.e. in a discrete set of size

$$S = \binom{m!}{n} .$$

It is obvious, that a full enumeration of this set is beyond available computational ressources except for very modest values of m and n.

In related problems of finding optimal U-type designs, the use of stochastic local search heuristics turned out to provide high quality results, which for some instances with theoretical lower bounds could be shown to be globally optimal [8]. Therefore, it appears appropriate to follow a similar strategy for the problem of finding good OofA designs. Specifically, we use an implementation of the threshold accepting heuristic [3], which is a simplified version of simulated annealing by using a deterministic acceptance criterion for each local move. It also belongs to the class of local search methods sequentially moving through the search space by making small changes to a given design. When comparing a new design with the current one, it allows downhill moves, i.e. accepts solutions which are (slightly) worse than the previous one, in order to escape local maxima. By decreasing the threshold, up to which a worsening of the objective function is allowed in a search step, to zero over the run time of the algorithm, the algorithm provides high quality approximations to the global optimum.

A survey on different heuristic approaches which could be used for the present problem instance can be found in [23], and a detailed description of the threshold accepting algorithm with several applications including some in experimental design is provided by [20]. Previous applications in the context of experimental design include the calculation of lower bounds for the star-discrepancy [21], and the generation of low discrepancy U-type designs for the star-discrepancy [22], several modifications of the L_2-discrepancy [5], for CL_2 [4, 7], and for CL_2 and WL_2 [6, 9]. Furthermore, further details on low-discrepancy designs can be found in [10, 12, 14]

The pseudo code for the threshold accepting implementation used for the OofA design problem is exhibited in Algorithm 1. Thereby, D stands for the D-criterion to be maximized. The algorithm remains unchanged if instead of D another objective function has to be maximized. For an objective function to be minimized, the algorithm can be applied on minus this objective function. In the results section, we will report the values of the D-criterion relative to the theoretical maximum for the full PWO design, i.e. the relative D-efficiency D_r.

The threshold accepting algorithm conducts a local search strategy on the set of all OofA designs with m components and n runs denoted by $\mathscr{O}(m, n)$. The initial

Algorithm 1 Pseudo-code for Threshold Accepting

1: Initialize n_R, n_{S_r} and the sequence of thresholds τ_r, $r = 1, 2, \ldots, n_R$
2: Generate starting design $O^0 \in \mathcal{O}(m, n)$
3: **for** $r = 1$ to n_R **do**
4: **for** $i = 1$ to n_{S_r} **do**
5: Generate $O^1 \in \mathcal{N}(O^0)$ (neighbor of O^0)
6: **if** $D(O^1) > D(O^0) - \tau_r$ **then**
7: $O^0 = O^1$
8: **end if**
9: **end for**
10: **end for**

design O^0 is selected randomly (2:). It should be noted that selecting a "good" initial design, which might correspond to a local maximum of the objective function does not improve the performance of the algorithm. Instead, using the best out of some repeated runs of the algorithm for different randomly selected initial designs might result in an improved performance and higher robustness as compared to a single run with a corresponding larger number of iterations.

Starting from the initial design, a local search step is repeated a substantial number of times. In each search step, a new candidate design O^1 is selected randomly within a neighborhood of the current design $\mathcal{N}(O^0) \subset \mathcal{O}(m, n)$ (5:). The value of the objective function for the new candidate solution is calculated $D(O^1)$. If it turns out to be larger than the one of the current design O^0, it will be accepted and becomes the current design (7:). However, the new design will also be accepted if it is worse than the current one, though only up to a certain threshold defined by the current value of the threshold sequence (τ_r) (6:). This "threshold accepting" behavior of the algorithm avoids getting stuck in suboptimal local maxima of the objective function. Nevertheless, as the sequence of threshold values decreases to zero during the course of the algorithm, towards the end of a run, only improvements will be accepted. The current implementation uses an elitist approach, i.e., the best design obtained during the runtime of the algorithm is reported. For a properly tuned implementation of the algorithm, this should be equal to or at least quite close to the last design accepted by the algorithm.

While the overall layout of the algorithm is simple, and it proofed to be robust to minor modifications of neighbourhood structure and parameter settings, the actual performance still depends on some problem specific tuning. The three most important aspects are the choice of local neighbourhoods, the specification of the threshold sequence, and, for obvious reasons, the total number of iterations the local search step is repeated within the algorithm. With regard to the definition of neighbourhoods, we follow the experience from earlier applcations of threshold accepting in experimental design. Starting with a design O^0, a small number of rows (2 in our implementation) are randomly selected. These rows are exchanged with a row differing only in the ordering of few elements close to each other. In principle, this definition of local neighbourhood also allows for a fast updating of the some objective functions as described in [6] for the first time. However, for the current implementation this feature

is not implemented yet. Nevertheless, given the tremendous growth in computational resources available, it is feasible to conduct repeated runs (10) for each problem instance with up to 10 000 000 iterations. In the results section, we report the best result obtained over all these runs. The corresponding designs are provided in the appendix.

The sequence of threshold values τ_r, $r = 1, \ldots, n_R$ is generated according to a data driven procedure first described in [20]. The rational of the approach is the observations that when performing local search over a discrete and finite search space such as $\mathcal{O}(m, n)$, also local changes of the objective function can take on only a finite number of different values. For the threshold accepting steps, all values of τ_r falling between two actual occurring differences will result in the same decision. Therefore, the threshold sequence is obtained by an empirical approximation to the underlying distribution of local changes of the objective function. To this end, first, a large number of OofA designs are randomly generated. For each of these designs a random neighbor is obtained employing the definition of local neighborhoods introduced before. The absolute value of the difference of the objective function between the two designs is calculated. The values are sorted in decreasing order and—given that too large thresholds imply an almost random search process—a lower quantile of these sequence is used as the actual sequence of threshold values. For the current application, the 60% of lower values (including zeros if the neighbor selected happens to be identical to the original design) is used.

6.4 Main Results

The best designs obtained by the threshold accepting algorithm are presented in Tables 6.1 and 6.2 (for $m \leq 30$). Both tables report the number of components m, the number of runs n, the D-efficiency as compared to the full PWO or full tapered PWO design, and the relation of n to the number of runs of the full designs. The corresponding designs are provided online (www.jlug.de/optimaloofadesigns). Given that there does not exist a closed form expression for the D-criterion of tapered PWO designs, we report D-efficiency for tapered PWO designs only up to $m = 10$. For larger values of m, the straightforward calculation of the D-criterion fails due to memory constraints. Therefore, Table 6.2 provides only results for the PWO designs, although using the algorithm also tapered PWO designs can be obtained for $m > 10$. A further constraint is imposed due to the numerical precision when calculating the values of the objective function. Using standard double precision only values of m up to about 30 are feasible, but the algorithm could be adjusted to work with higher precision.

Table 6.1 *D*-efficiency of tapering OofA designs obtained by threshold accepting

Components (m)	Runs (n)	TA optimized design		Share of runs ($n/m!$)
		D-efficiency PWO	*D*-efficiency Tapered PWO	
4	7	0.89613	0.84433	0.2917
4	13	0.98571	0.98585	0.5417
4	19	0.98122	0.98097	0.7917
5	11	0.90267	0.91904	0.0917
5	21	0.97278	0.97848	0.1750
5	31	0.98733	0.98974	0.2583
6	16	0.88107	0.84169	0.0222
6	31	0.97039	0.96663	0.0431
6	46	0.98854	0.98629	0.0649
7	22	0.81196	0.77259	4.3651×10^{-3}
7	43	0.96517	0.95798	8.5317×10^{-3}
7	64	0.98285	0.98217	12.6984×10^{-3}
8	29	0.75717	0.73876	0.7192×10^{-3}
8	57	0.95166	0.94345	1.4137×10^{-3}
8	85	0.97750	0.97429	2.1081×10^{-3}
9	37	0.72626	0.69174	0.1020×10^{-3}
9	73	0.93923	0.93100	0.2012×10^{-3}
9	109	0.97339	0.96662	0.3004×10^{-3}
10	46	0.68087	0.65436	0.0127×10^{-3}
10	91	0.92463	0.91838	0.0251×10^{-3}
10	136	0.96336	0.95770	0.0375×10^{-3}

Note: For each m, the D-efficiency of the obtained design for three experimental runs n are displayed: (a) the minimal point n = q + 1; (b) double type n = 2q + 1; (c) triplicate type n = 3q + 1; where $q = \binom{m}{2}$

The results indicate that with a rather small number of runs, highly efficient designs can be obtained. For Case (a), the minimal-point design $n = q + 1$; all designs reach at least 80% of the efficiency of the full design, though with only a small fraction of runs, especially for large m, $n/m!$ becomes almost zero—a substantial saving. For example, for $m = 20$ and $n = 191$, we have $n/m! = 0.0785 \times 10^{-15}$. Hence, if the practitioner attempts to save resource and time, the minimal-point design is a good choice. For Case (b), the double type design $n = 2q + 1$; all designs reach at least 95% of the efficiency of the full design, while the corresponding $n/m!$ is also almost zero. For example, for $m = 20$ and $n = 381$, we have $n/m! = 0.1566 \times 10^{-15}$. We recommend the experimenter who seeks designs with high D-efficiency to use the double designs when the resource and time allow. For Case (c), the triplicate type design $n = 3q + 1$; the optimal designs attain at least 97% of the efficiency of the full design, while the run size $n/m!$ is near zero. For example, for $m = 20$ and

Table 6.2 *D*-efficiency of OofA designs obtained by threshold accepting for $m > 10$

Components (m)	Runs (n)	TA optimized design D-efficiency (PWO)	Share of runs ($n/m!$)
11	56	0.80170	1.4029×10^{-6}
11	111	0.95969	2.7808×10^{-6}
11	166	0.98228	14.1586×10^{-6}
12	67	0.78958	0.1399×10^{-6}
12	133	0.95646	0.2777×10^{-6}
12	199	0.98081	0.4154×10^{-6}
13	79	0.77952	0.0127×10^{-6}
13	157	0.95238	0.0252×10^{-6}
13	235	0.97934	0.0377×10^{-6}
14	92	0.76463	1.0553×10^{-9}
14	183	0.94925	2.0991×10^{-9}
14	274	0.97744	3.1430×10^{-9}
15	106	0.75398	0.0811×10^{-9}
15	211	0.94704	0.1614×10^{-9}
15	316	0.97637	10.2417×10^{-9}
16	121	0.74091	0.0058×10^{-9}
16	241	0.94420	0.0115×10^{-9}
16	361	0.97389	0.0173×10^{-9}
17	137	0.73361	0.3852×10^{-12}
17	273	0.94096	0.7675×10^{-12}
17	409	0.97229	1.1499×10^{-12}
18	154	0.72681	0.0241×10^{-12}
18	307	0.93764	0.0480×10^{-12}
18	460	0.97088	0.0718×10^{-12}
19	172	0.71426	1.4139×10^{-15}
19	343	0.93483	2.8197×10^{-15}
19	514	0.96900	14.2254×10^{-15}
20	191	0.70542	0.0785×10^{-15}
20	381	0.93160	0.1566×10^{-15}
20	571	0.96728	0.2347×10^{-15}
25	301	0.65850	1.9405×10^{-23}
25	601	0.91783	3.8746×10^{-23}
25	901	0.95955	5.8087×10^{-23}
30*	871	0.90459	3.2837×10^{-30}
30	1306	0.95064	4.9236×10^{-30}

Note 1: For each m, the D-efficiency of the obtained design for three experimental runs n are displayed: (a) the minimal point $n = q + 1$; (b) double type $n = 2q + 1$; (c) triplicate type $n = 3q + 1$; where $q = \binom{m}{2}$.

Note 2: Using standard double precision only values of m up to about 30 are feasible (for $m = 30$ it depends on n), but the algorithm could be adjusted to work with higher precision

$n = 571$, we have $n/m! = 0.2347 \times 10^{-15}$. The triplicate designs are recommended due to a high D-efficiency, if there are no restrictions on the experimental conditions. In general, all the obtained three experimental designs with a small number of runs ($n = q + 1, 2q + 1, 3q + 1$) achieve a high level of D-efficiency when the underlying model is the PWO or the tapered PWO model.

6.5 Example: Scheduling Problem

The purpose of the scheduling problem is to schedule the resources and tasks to be optimized with regard to one or more objectives. A popular class of scheduling problem is the "job scheduling" problem, which seeks an optimal order of these jobs. Consider m jobs requiring processing in a certain machine environment, the scheduler hopes to sequence these jobs under given constraints. Let p_i ($i = 1, \ldots, m$) represents the processing time of job i on a machine, ω_i being pre-specified weights, which reflects the importance of job i relative to the other jobs in the system. A schedule $\pi = (\pi_1, \ldots, \pi_m)$ is a permutation of $\{1, \ldots, m\}$ which specifies the order in which to process jobs. The completion time of the operation of job j is denoted as $C_j(\pi) = \sum_{i=1}^{j} p_i$, and the total cost function (total weighted completion time) is denoted by $W = \sum_{k=1}^{m} W_k = \sum_{k=1}^{m} \omega_k C_k^2$, where W_k denotes the cost function of job k. The objective is to find a optimal job-order such that the total cost function W is minimized.

For illustration, a simple example of job scheduling for a single machine model with $m = 3$ is discussed. Suppose the processing times are 5, 3 and 2 h for the 1st job, 2nd job and 3rd job, respectively, and the weights (cost coefficients) of these jobs are 6, 8, 7, respectively. Consider the job-order $1 \to 2 \to 3$, the completion time for job 1 is $C_1 = p_1 = 5hr$, the completion time for job 2 is $C_2 = p_1 + p_2 = 5 + 3 = 8hr$, the completion time for job 3 is $C_3 = p_1 + p_2 + p_3 = 5 + 3 + 2 = 10hr$. Thus the job-order $1 \to 2 \to 3$ has a total cost function: $W = 6 \times 5^2 + 8 \times 8^2 + 7 \times 10^2 = 1362$.

The purpose of the job scheduling problem is to find the optimal sequence from all possible solutions to minimize the total penalty. This is the same as the goal of the OofA problem. Hence we can consider the job scheduling problem as an OofA experiments problem. The PWO model can be used as the approximate model to deal with the job scheduling problem. If we have the prior information of the cost function, we can compute the costs of all possible job orders and compare all possible different orders to learn the dependence of the response on the order. For the illustrative example, since there is only $m! = 3! = 6$ potential orders, one can evaluate all possible order to find the optimal order. However, with m components to add, an exhaustive search of all permutations requires $m!$ runs of experiments, which is usually not affordable. So the design problem arises to choose a subset of orders for comparison.

Next, we discuss a case that 10 jobs are to be sequenced on a single machine with quadratic penalty function of its completion time. The pre-specified weights and processing times of these jobs are randomly generated from χ_1^2, such as $p = (0.1451700, 0.7428453, 7.1142859, 1.8774267, 7.1185982, 1.1431286, 10.5172882, 2.1484186, 2.4950454, 2.8989094)$, $\omega = (2.2094712, 7.1116628, 0.4190265, 6.7777317, 1.2965368, 1.5379331, 0.7221195, 1.7368003, 0.5205548, 0.3908880)$. With 10 components, it is infeasible to conduct all possible orders (tests) $10! \approx 3.6$ million. This is especially true for physical experiments and some expensive computer experiments. To fit the PWO model for using the smallest runs to save costs, we use the minimal-point OofA design $n = \binom{10}{2} + 1 = 46$ runs obtained from our algorithm in this paper. Under the quadratic penalty function, the 46 runs design and the corresponding total cost are shown in Table 6.3. Upon using the least squares approach, the parameters $\widehat{\beta}_{ij}$ in PWO model (6.2) are estimated. Ultimately, an OofA experiment is to find the optimal addition order. According to the method proposed by [2], the active variables are very critical for selecting the optimal orders. Considering all of degree of freedom for the minimal-point design are used to estimate the parameters in PWO model, there is no remaining degree of freedom to estimate the standard deviation. The Lenth ([11]) method is used here to identify the active variables, we take the pseudo standard error (PSE) to estimate the standard deviation. Upon calculating the Lenth statistics: $t_{PSE,i} = \widehat{\beta}_{ij}/PSE$, the active variables are tabulated in Table 6.4.

The favorable pattern "i precedes j" indicates an edge from i to j. From the favorable patterns of all active variables, one can always generate the corresponding directed graph by sequentially considering the significant parameters according to the absolute values of the active variables' estimated effects. In other words, we first consider the directed edge determined by the most significant parameter (the largest effect), then consider the directed edge from the second significant parameter (the second largest effect), and so on. In this procedure, we omit the active variables that are contrary to the current generated directed graph. Take the sequence '1628' as example, we sequently consider the active variables $\widehat{\beta}_{2,8}$ (966.425), $\widehat{\beta}_{2,6}$ (775.033), $\widehat{\beta}_{1,6}$ (700.148), $\widehat{\beta}_{1,8}$ (624.832). In this way, we omit the active variable $\widehat{\beta}_{1,8}$, because it is contrary to the generated sequence '1628'.

From the favorable patterns exhibited in Table 6.4, the directed graph of 10 jobs results in Fig. 6.1. According to the Fig. 6.1, the first five components of an optimal sequence are '16284', and the remaining optimal possible job orders can be obtained by permutating the last components '3, 5, 9, 10, 7', hence the possible numbers of order is $5!/2 = 120$. One of possible optimal orders is $1 \longrightarrow 6 \longrightarrow 2 \longrightarrow 8 \longrightarrow 4 \longrightarrow 5 \longrightarrow 9 \longrightarrow 10 \longrightarrow 7 \longrightarrow 3$, and the cost function is 1958.716.

We randomly select 100 runs from 10! and compute the costs of these 100 runs. As a comparison, we randomly select 54 (we already run 46 experiments to collect the data, for fairness, $100 - 46 = 54$ possible runs are selected) possible orders from our possible 120 orders, the results are showed in Fig. 6.2a. It can be seen that the costs of our method are smaller than the costs of randomly selected. This highlights the merit of our approach.

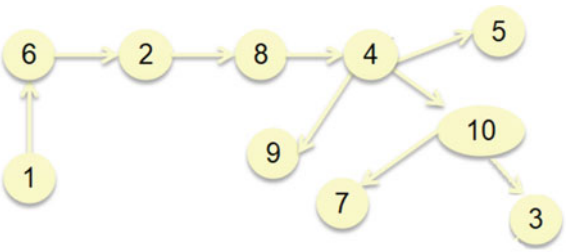
Table 6.3 The job-orders and total costs for a 10 job scheduling problem

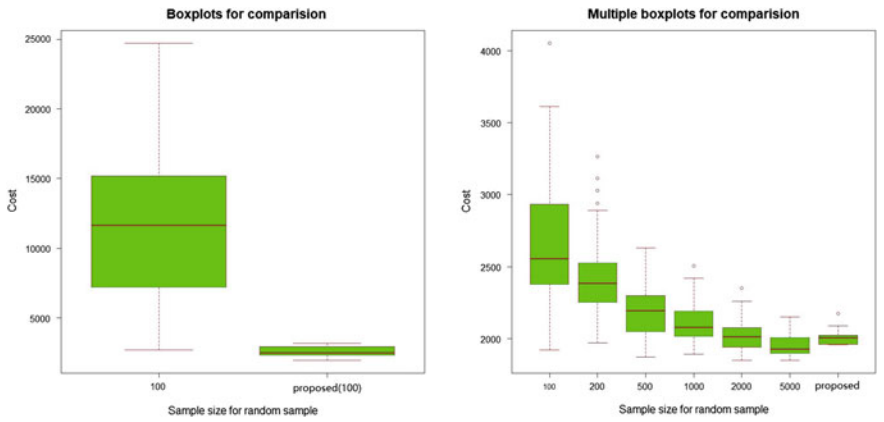
Run	Order										Cost (W)
1	4	5	10	2	3	8	9	6	7	1	7463.671
2	8	3	4	10	6	5	9	2	7	1	10754.214
3	3	1	8	10	6	2	9	5	7	4	12674.186
4	9	6	4	10	8	7	3	5	2	1	14862.981
5	8	2	10	6	4	7	9	1	3	5	4318.170
6	9	2	7	3	6	5	10	1	8	4	15859.299
7	3	9	1	7	8	5	4	10	2	6	20486.508
8	1	5	6	2	9	3	10	4	7	8	8041.683
9	6	10	2	4	3	9	1	8	7	5	4193.411
10	10	6	7	2	1	8	5	4	3	9	9654.549
11	8	9	7	5	1	10	3	4	2	6	21368.549
12	10	8	2	6	4	3	5	7	1	9	5686.498
13	10	2	1	7	8	5	9	3	6	4	12976.774
14	7	4	9	1	3	10	6	5	8	2	16184.010
15	5	6	8	2	1	4	10	9	3	7	4205.401
16	9	3	7	8	6	4	2	1	5	10	14663.133
17	7	1	6	4	2	8	3	5	10	9	6363.397
18	8	5	9	2	10	1	3	6	7	4	12750.318
19	8	6	4	9	3	10	1	2	7	5	5912.879
20	3	8	2	7	1	4	5	6	9	10	9265.334
21	10	4	1	3	2	6	9	7	5	8	6175.202
22	4	2	8	1	7	3	9	10	5	6	4791.054
23	4	6	7	8	2	9	5	3	10	1	7447.334
24	10	3	7	6	1	9	5	4	2	8	21755.562
25	5	2	9	8	4	3	1	6	10	7	5465.729
26	3	6	8	1	10	7	4	5	2	9	15743.304
27	3	5	8	9	4	10	1	7	6	2	17428.615
28	10	9	2	7	4	6	3	8	5	1	9745.608
29	7	10	5	8	3	9	1	2	6	4	23354.098
30	1	6	3	5	4	2	7	10	9	8	8598.878
31	10	3	7	4	1	8	9	5	6	2	19101.778
32	4	6	1	3	10	8	9	5	7	2	11780.016
33	3	9	4	1	2	5	10	6	8	7	5984.161
34	6	1	5	7	8	9	10	3	2	4	19423.341
35	1	3	10	7	2	4	9	5	8	6	13212.515
36	1	3	7	2	6	8	4	9	10	5	10300.234
37	8	2	9	3	7	4	10	6	1	5	9889.003
38	4	9	6	8	10	7	1	2	5	3	6620.735
39	7	10	2	8	3	1	6	4	9	5	11364.447
40	10	5	7	8	6	4	3	2	9	1	18716.247
41	9	5	6	3	8	7	10	4	2	1	22834.160
42	8	6	7	4	5	10	1	2	3	9	10426.000
43	1	4	8	6	9	7	2	10	3	5	5203.967
44	7	8	9	1	4	6	10	2	3	5	9175.293
45	9	3	8	5	6	7	10	1	2	4	22223.030
46	10	8	1	9	4	7	6	5	3	2	12630.479

Table 6.4 The active variables for 10 jobs scheduling

Active variables	Estimator of the effects	Signs of the effects	Favorable patterns
$I_{1,2}$	−442.110	−	1 precedes 2
$I_{1,6}$	−700.148	−	1 precedes 6
$I_{1,8}$	624.832	+	8 precedes 1
$I_{1,10}$	−492.119	−	1 precedes 10
$I_{2,3}$	−1110.545	−	2 precedes 3
$I_{2,4}$	−431.188	−	2 precedes 4
$I_{2,5}$	−1314.436	−	2 precedes 5
$I_{2,6}$	775.033	+	6 precedes 2
$I_{2,7}$	−1349.35	−	2 precedes 7
$I_{2,8}$	−966.425	−	2 precedes 8
$I_{2,10}$	−858.088	−	2 precedes 10
$I_{3,4}$	631.513	+	4 precedes 3
$I_{3,6}$	476.540	+	6 precedes 3
$I_{3,10}$	452.552	+	10 precedes 3
$I_{4,5}$	−1343.049	−	4 precedes 5
$I_{4,7}$	−562.938	−	4 precedes 7
$I_{4,8}$	785.005	+	8 precedes 4
$I_{4,9}$	−714.446	−	4 precedes 9
$I_{4,10}$	−1039.873	−	4 precedes 10
$I_{6,7}$	−674.048	−	6 precedes 7
$I_{7,10}$	641.382	+	10 precedes 7

Fig. 6.1 Directed Graph





(a) 100 from 10! and 120 possible orders (b) different sample sizes

Fig. 6.2 Boxplots of different sample sizes

As a comparison benchmark, a sample of $n = 100$ orders was randomly chosen (from all possible 10! orders) and their corresponding costs were evaluated. Among these 100 costs, the minimal cost was recorded. We repeat this process for 50 times. These 50 minimal costs were displayed as the first boxplot in Fig. 6.2b. Similarly, the same process is applied to $n = 200, 500, 1000, 2000$ and 5000 respectively. Their boxplots for minimal costs are also displayed in Fig. 6.2b. As expected, the larger n, the smaller minimal cost is found (with more consistency as well). As compared to our finding, we randomly chose 54 orders (out of the obtained result of 120 orders). Their corresponding costs were evaluated and the minimal cost is recorded. We also repeat this process for 50 times and the results is displayed as the last boxplot in Fig. 6.2b. It is clear that the performance of the proposed method with 100 ($= 46 + 54$) runs is as good as the random sample with 5000 runs. This more or less confirms the validity of the proposed method.

6.6 Conclusion

Order-of-addition (OofA) experiments have increasingly received a great deal of attention in scientific and industrial applications. This paper uses threshold accepting algorithm to provide a class of minimal-point designs with $n = q + 1$ runs, a class of double type design with $n = 2q + 1$ runs and a class of triplicate type design with $n = 3q + 1$ runs under the PWO model and tapered PWO model, respectively. The D-efficiency of these designs are shown in Tables 6.1 and 6.2. As a matter of fact, the threshold accepting algorithm can be used to construct the optimal design among any size n under any other models, such as, the triplet-order model [15] and the component-position model [24].

It is shown that a threshold accepting implementation can be used to generate highly efficient designs for OofA experiments. Here, the analysis was restricted to consider PWO designs. However, the framework is general enough to allow the construction of efficient designs taking into account higher orders. For improving the performance of the algorithm for larger values of m and n, the implementation has to be modified both with regard to the generation of random OofA designs and their mapping to the corresponding Z-matrices and with regard to the updating of the objective function for small modifications of a design in a neighbourhood. These improvements of the algorithm will allow tackling also larger problem instances and higher orders regarding the sequence of additions.

Acknowledgements All designs obtained can be found in website: www.jlug.de/optimaloofadesigns. This work was partially supported by the *National Science Foundation* via Grant DMS 18102925. The work of Jianbin Chen was supported by the National Natural Science Foundation of China (Grant Nos. 11771220). Professor Kai-Tai Fang has been a true leader in our society and has been a strong supporter of young fellows. His original work on uniform design had a significant impact on this work. It is our great privilege to contribute this work on computer experiments to this special issue in honor of his 80th birthday.

References

1. Chen, J.B., Han, X.X., Yang, L.Q., Ge, G.N., Zhou, Y.D.: Fractional designs for order of addition experiments. Submitt. Manuscr. (2019)
2. Chen, J.B., Peng, J.Y., Lin, D.K.J.: A statistical perspective on NP-Hard problems: making uses of design for order-of-addition experiment. Manuscript (2019)
3. Dueck, G., Scheuer, T.: Threshold accepting: a general purpose algorithm appearing superior to simulated annealing. J. Comput. Phys. **90**, 161–175 (1990)
4. Fang, K.-T., Lin, D.K.J.: Uniform experimental designs and their applications in industry. In: Khattree, R., Rao, C.R. (eds.) Handbook of Statistics, vol. 22, pp. 131–170. Elsevier, Amsterdam (2003)
5. Fang, K.-T., Lin, D.K.J., Winker, P., Zhang, Y.: Uniform design: theory and application. Technometrics **42**, 237–248 (2000)
6. Fang, K.-T., Lu, X., Winker, P.: Lower bounds for centered and wrap-around L_2-discrepancies and construction of uniform designs by Threshold Accepting. J. Complex **19**, 692–711 (2003)
7. Fang, K.-T., Ma, C.-X., Winker, P.: Centered L_2 discrepancy of random sampling and latin hypercube design, and construction of uniform designs. Math. Comput. **71**, 275–296 (2002)
8. Fang, K.-T., Maringer, D., Tang, Y., Winker, P.: Lower bounds and stochastic optimization algorithms for uniform designs with three or four levels. Math. Comput. **75**(254), 859–878 (2005)
9. Fang, K.-T., Tang, Y., Yin, J.: Lower bounds for wrap-around L_2-discrepancy and constructions of symmetrical uniform designs. Forthcomming (2004)
10. Fang, K.-T., Wang, Y.: Applications of Number Theoretic Methods in Statistics. Chapman and Hall, London (1994)
11. Lenth, R.V.: Quick and easy analysis of unreplicated factorials. Technometrics **31**, 469–473 (1989)
12. Lin, D.K.J., Sharpe, C., Winker, P.: Optimized U-type designs on flexible regions. Comput. Stat. Data Anal. **54**, 1505–1515 (2010)
13. Lin, D.K.J., Peng, J.Y.: The order-of-addition experiments: a review and some new thoughts (with discussion). Qual. Eng. **31**(1), 49–59 (2019)

14. Liu, M.Q., Hickernell, F.J.: $E(s^2)$-optimality and minimum discrepancy in 2-level superdatu-rated designs. Statistica Sinica **12**, 931–939 (2002)
15. Mee, R.W.: Order of addition modeling. Statistica Sinica. In press (2019)
16. Pukelsheim, F.: Optimal Design of Experiments. Wiley, New York (1993)
17. Peng, J.Y., Mukerjee, R., Lin, D.K.J.: Design of order-of-addition experiments. Biometrika. In press (2019)
18. Van Nostrand, R.C.: Design of experiments where the order of addition is important. In: ASA Proceedings of the Section on Physical and Engineering Sciences, pp. 155–160. American Statistical Association, Alexandria, Virginia (1995)
19. Volkel, J.G.: The design of order-of-addition experiments. J. Qual. Technol. (2019) https://doi.org/10.1080/00224065.2019.1569958
20. Winker, P.: Optimization Heuristics in Econometrics. Wiley, Chichester (2001)
21. Winker, P., Fang, K.-T.: Application of Threshold Accepting to the evaluation of the discrepancy of a set of points. SIAM J. Numer. Anal. **34**, 2028–2042 (1997)
22. Winker, P., Fang, K.-T.: Optimal U-type designs. In: Niederreiter, H., Hellekalek, P., Larcher, G., Zinterhof, P. (eds.) Monte Carlo and Quasi-Monte Carlo Methods, pp. 436–488. Springer, New York (1997)
23. Winker, P., Gilli, M.: Applications of optimization heuristics to estimation and modelling problems. Comput. Stat. Data Anal. **47**, 211–223 (2004)
24. Yang, J.F., Sun, F.S., Xu, H.: Component orthogonal arrays for order-of-addition experiments. Submitt. Manuscr. (2019)
25. Zhao, Y.N., Lin, D.K.J., Liu, M.Q.: Minimal-point design for order of addition experiment. Submitt. Manuscr. (2019)

Chapter 7
Construction of Uniform Designs on Arbitrary Domains by Inverse Rosenblatt Transformation

Mei Zhang, Aijun Zhang, and Yongdao Zhou

Abstract The uniform design proposed by Fang [6] and Wang and Fang [17] has become an important class of designs for both traditional industrial experiments and modern computer experiments. There exist established theory and methods for constructing uniform designs on hypercube domains, while the uniform design construction on arbitrary domains remains a challenging problem. In this paper, we propose a deterministic construction method through inverse Rosenblatt transformation, as a general approach to convert the uniformly designed points from the unit hypercubes to arbitrary domains. To evaluate the constructed designs, we employ the central composite discrepancy as a uniformity measure suitable for irregular domains. The proposed method is demonstrated with a class of flexible regions, constrained and manifold domains, and the geographical domain with very irregular boundary. The new construction results are shown competitive to traditional stochastic representation and acceptance-rejection methods.

7.1 Introduction

The uniform design of experiments has generated a great amount of research papers and impact cases ever since it was first proposed by Fang [6] and Wang and Fang [17]. It has been successfully used for both traditional industrial experiments and modern computer experiments; see the monographs [7, 9].

M. Zhang
College of Mathematics, Sichuan University, Chengdu 610064, China
e-mail: zhangmei150320@163.com

A. Zhang (✉)
Department of Statistics and Actuarial Science, The University of Hong Kong, Hong Kong, China
e-mail: ajzhang@hku.hk

Y. Zhou
School of Statistics and Data Science & LPMC, Nankai University, Tianjin 300071, China
e-mail: ydzhou@nankai.edu.cn

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_7

To construct uniform designs on the unit hypercube $C^s = [0, 1]^s$, there exist both theoretical approaches and numerical optimization methods; see the latest book of Fang et al. [8] for a complete treatment. Among these methods, the classical good lattice point (GLP) method based on number theory is simple yet effective, and it is also widely used in quasi-Monte Carlo sampling [14]. For the GLP method with respect to the classical star-discrepancy criterion, Fang and Wang [9] (Appendix A) provides a catalogue of generating vectors up to 18 dimensions. For two-dimensional uniform designs in particular, it is well-known that the GLPs generated through Fibonacci numbers enjoy the low star-discrepancy properties. However, it is not clear whether such Fibonacci designs also enjoy the low centered-ℓ_2 discrepancy (CD2) properties.

It is of our interest to construct the uniform designs on arbitrary experimental domains, including regular and irregular regions. For regular regions such as ball, sphere and simplex, Fang and Wang [9] suggested the inverse transformation method through establishing a non-trivial analytic stochastic representation (SR) for the random vector uniformly distributed on each regular region, and then generating the uniform designs by inversely mapping from the unit hypercubes. For mixture experiments with single-factor constraints, Fang and Yang [10] proposed a conditional distribution method that also takes a non-trivial SR form. This method is further applied by Tian et al. [16] for generating uniform designs on tetragon and convex polyhedrons. For an irregular region $\mathscr{X} \subset \mathbb{R}^s$ that does not have the explicit SR form, one usually resorts to the acceptance-rejection (AR) method that first generates uniform designs on a superset hypercube $\mathscr{C} \supseteq \mathscr{X}$, then retains only the design points within the region of interest. Such AR method was suggested by Borkowski and Piepel [2] for mixture experiments with complex multi-factor constraints. However, the AR method is less efficient especially when $\mathscr{X}$ is much smaller than $\mathscr{C}$, and the resulting design in $\mathscr{X}$ sometimes has poor uniformity.

The construction of uniform designs on arbitrary domains remains a challenging problem. Numerically, one may use the stochastic optimization methods to directly search for the design points according to a certain uniformity criterion. Chuang and Hung [5] proposed a central composite discrepancy (CCD) to measure uniformity with regard to an arbitrary domain $\mathscr{X}$, then used a switching algorithm to search for the best design over a set of pre-specified points. Also based on the CCD criterion, Lin et al. [13] applied the threshold accepting algorithm to optimize the U-type designs on flexible regions, and Chen et al. [3] developed the discrete particle swarm optimization algorithm with GPU acceleration. Other space-filling criteria can also be used to directly measure the uniformity on irregular regions, e.g. the maximin distance used by Chen et al. [4].

In this paper, we study a deterministic method based on the inverse Rosenblatt transformation (IRT) for constructing uniform designs on arbitrary domains. It can be viewed as a special kind of inverse method, as is remarked by Fang and Wang [9, p. 54]. To distinguish the IRT from the existing inverse method based on the specific analytical SR, we call the latter as the SR method in this paper. Unlike the SR methods reviewed above, the IRT method does not necessarily require the analytical forms of conditional distribution functions, which can be easily approximated for a

uniform experimental domain with irregular boundary. The IRT method includes the conditional distribution method by Fang and Yang [10] as a special case for restricted mixtures. We demonstrate how the proposed method can be used to construct uniform designs on a class of flexible regions, constrained and manifold domains, and the irregular domains such as geographical maps. Among the uniformity criteria, we employ the aforementioned CCD to evaluate the constructed designs on general domains. For regular regions, the construction results by the IRT method are compared with both SR and AR methods; while for irregular regions, they are compared with the AR method.

The rest of this paper is organized as follows. In Sect. 7.2 we propose the IRT method based on the marginal and conditional distributions subject to permutation, and illustrate it through a synthetic example. Section 7.3 presents the construction results on a variety of regular and irregular domains. Some concluding remarks are given in Sect. 7.4. In the Appendix, we provide a brief review of the GLP method for constructing uniform designs on hypercube domains, which are used by the proposed IRT method. We show that the leave-one-out Fibonacci designs achieve the minimum centered ℓ_2-discrepancy for the mixed GLP method.

7.2 Inverse Rosenblatt Transformation Method

The Rosenblatt transformation [15] is a general mapping of multivariate random variables with a continuous distribution to the uniform distribution on unit hypercubes. It can be used as a tool for construction of multivariate distributions and goodness-of-fit testing; see e.g. Justel et al. [12] and Arnold et al. [1].

Let $X \in \mathscr{X} \subseteq \mathbb{R}^s$ be a random vector with joint density

$$f(x_1, \ldots, x_s) = f_1(x_1) f_{2|1}(x_2|x_1) \cdots f_{s|1,\ldots,s-1}(x_s|x_1, \ldots, x_{s-1}). \tag{7.1}$$

Denote F_1 as the marginal cumulative distribution function (CDF) of the first component X_1, and by $F_{2|1}, \ldots, F_{s|1,\ldots,s-1}$ the consecutive conditional CDFs. Then, the Rosenblatt transformation (RT) is defined by

$$\begin{cases} U_1 = F_1(X_1), \\ U_j = F_{j|1,\ldots,j-1}(X_j|X_1, \ldots, X_{j-1}), \ \ j = 2, \ldots, s. \end{cases} \tag{7.2}$$

It is clear that (a) $U_1, \ldots, U_s$ are independent Uniform[0, 1] random variables, (b) $(X_1, \ldots, X_s) \to (U_1, \ldots, U_s)$ is one-to-one from $\mathscr{X}$ to C^s, and (c) the Jacobian of RT corresponds to the joint density function f. Note that RT is not permutation-invariant and there exist $s!$ kinds of different forms according to reordering $(X_1, \ldots, X_s)$. For each permutation $(i_1, \ldots, i_s)$, denote $T_{(i_1,\ldots,i_s)}$ as the RT from the permutated $(X_{i_1}, \ldots, X_{i_s})$ to $(U_1, \ldots, U_s)$.

Given the uniform distribution on C^s, the inverse $T^{-1}_{(i_1,\ldots,i_s)}$ maps $\boldsymbol{u} = (U_1, \ldots, U_s)$ back to $\boldsymbol{x} = (X_{i_1}, \ldots, X_{i_s}) \in \mathscr{X}$. More specifically, for a given

observation $(u_1, \ldots, u_s) \in C^s$, the inverse Rosenblatt transformation (IRT) can be expressed by

$$\begin{cases} x_{i_1} = Q_{i_1}(u_1), \\ x_{i_j} = Q_{i_j|i_1,\ldots,i_{j-1}}(u_j | X_{i_1} = x_{i_1}, \ldots, X_{i_{j-1}} = x_{i_{j-1}}), \ j = 2, \ldots, s \end{cases} \tag{7.3}$$

based on the quantile functions (i.e. inverse CDFs). To illustrate how the IRT works, we consider the following example with two disjoint rectangular regions, where the uniform distribution is assumed on each sub-region.

Example 7.1 (**Two-rectangle Domain**) Consider the uniform distribution on the domain $\mathscr{X}$ with two disjoint rectangles, as shown in Fig. 7.1. It is clear the marginal and conditional CDFs for either permutation (x_1, x_2) or (x_2, x_1) are of the piecewise linear forms, as depicted in Fig. 7.2. Suppose we are given a point $\boldsymbol{u} = (0.2, 0.75) \in C^2$, then it is straightforward to apply the IRT in (7.3), we can

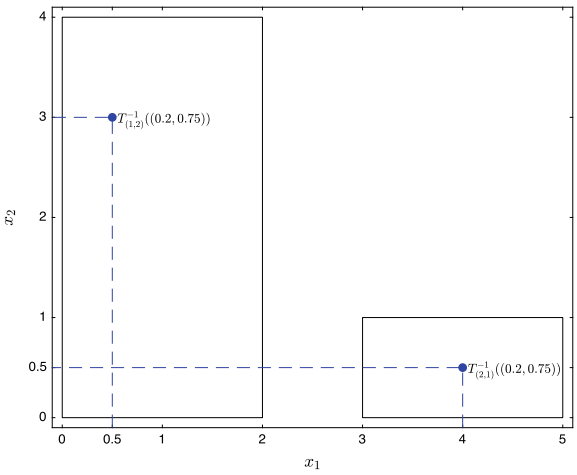


Fig. 7.1 Experimental domain $\mathscr{X}$ formed with two disjoint rectangles with the vertices $((0, 0), (2, 0), (2, 4), (0, 4))$ and $((3, 0), (5, 0), (5, 1), (0, 1))$, respectively. The two points at $(0.5, 3)$ and $(4, 0.5)$ are converted from $(0.2, 0.75) \in C^2$ through $T^{-1}_{(1,2)}$ and $T^{-1}_{(2,1)}$, respectively

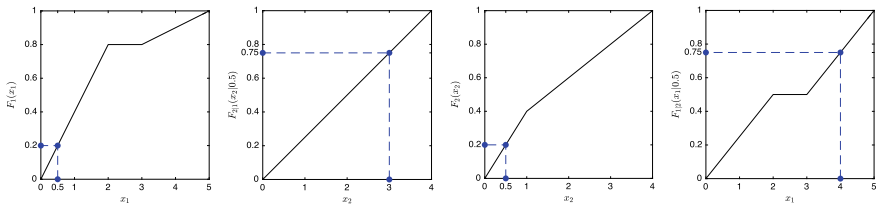


Fig. 7.2 The marginal and conditional CDFs used for obtaining the IRT points on the two-rectangle domain in Fig. 7.1. The point $(0.2, 0.75) \in C^2$ is taken as an example for performing the IRT

obtain the corresponding points in $\mathscr{X}$ through both $T_{(1,2)}^{-1}$ and $T_{(2,1)}^{-1}$, as shown in Fig. 7.1.

Algorithm 3 Inverse Rosenblatt Transformation Method

Input: An arbitrary domain $\mathscr{X}$ with closed boundary, and an n-run uniform design $\{\boldsymbol{u}_i\}_{i=1}^n$ on C^s.

1: Choose a permutation $(i_1, \ldots, i_s)$ of $(1, \ldots, s)$, then find the corresponding $T_{(i_1,\ldots,i_s)}$ based on uniform distribution within in the given boundary.

2: Use IRT to convert $\{\boldsymbol{u}_i\}_{i=1}^n$ to the domain $\mathscr{X}$:

$$\boldsymbol{x}_i = T_{(i_1,\ldots,i_s)}^{-1}(\boldsymbol{u}_i), \quad i = 1, \ldots, n.$$

3: Evaluate the CCD criterion (7.4) for the resulted $X = \{\boldsymbol{x}_i\}_{i=1}^n$.

4: Repeat Steps 1–3 for all $s!$ permutations. Output the best design X^* with the lowest CCD score.

The two-rectangle domain example motivates us to develop a practical IRT method for constructing uniform designs on arbitrary experimental domains, as presented in Algorithm 3. For any domain with uniform distribution and closed boundary, in the first step, we can find its RT's $T_{(i_1,\ldots,i_s)}$ subject to a different permutation. The marginal and conditional CDFs can be either derived analytically by (multiple) integration, or obtained through hyperrectangle approximation. In the latter situation, each CDF is approximated by a non-decreasing piecewise linear function.

In the second step, we can apply (7.3) to each point of a given n-run uniform design on C^s, so as to generate the corresponding points on $\mathscr{X}$. As we have reviewed, there are a rich collection of existing methods for constructing uniform designs on unit hypercubes. See the Appendix about the GLP method together with the elegant Fibonacci designs on unit squares.

Each uniform design on C^s leads to at maximum $s!$ different designs since there are $s!$ versions of IRT $T_{(i_1,\ldots,i_s)}^{-1}$ subject to different permutations. To determine the best design on $\mathscr{X}$, we employ the aforementioned CCD criterion as a measure of uniformity on arbitrary domains. According to Chuang and Hung [5], for any interior point $\mathbf{z}$ in $\mathscr{X}$, it can be treated as the Cartesian center to cut $\mathscr{X}$ into 2^s quadrants, then the ℓ_2 form of CCD is defined by

$$\mathrm{CCD}(X) = \left\{ \frac{1}{V(\mathscr{X})} \int_{\mathscr{X}} \frac{1}{2^s} \sum_{k=1}^{2^s} \left| \frac{N(\mathscr{X}_k(\mathbf{z}), X)}{n} - \frac{V(\mathscr{X}_k(\mathbf{z}))}{V(\mathscr{X})} \right|^2 d\mathbf{z} \right\}^{1/2}, \quad (7.4)$$

where $N(\mathscr{X}_k(\mathbf{z}), X)$ denotes the number of design points in $\mathscr{X}_k(\mathbf{z})$, and $V(\mathscr{X})$ and $V(\mathscr{X}_k(\mathbf{z}))$ denote the volumes of $\mathscr{X}$ and $\mathscr{X}_k(\mathbf{z})$, respectively. In practice, the integration over $\mathscr{X}$ can be approximated by Monte Carlo average over a large number of equal-spaced grid points (say, 100 grid points along each coordinate).

Note that in Algorithm 3, when the domain $\mathscr{X}$ is symmetric in two or more coordinates, some permutations can be relaxed and we only need consider all permutations for asymmetric coordinates. For the symmetric flexible regions in $\mathbb{R}^2$

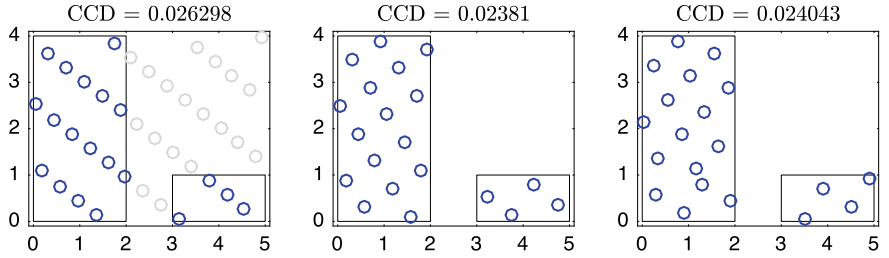


Fig. 7.3 Construction results of 20-run uniform designs on the two-rectangle domain in Fig. 7.1. Left panel: AR method; Center panel: IRT method with permutation (1, 2); Right panel: IRT method with permutation (2, 1)

to be discussed in next section, there is no need to consider permutations, so the evaluation of CCD criterion can be also skipped. Take the cylinder domain $\mathscr{X} = \{(x_1, x_2, x_3) : x_1^2 + x_2^2 \le r^2, a \le x_3 \le b\} \subset \mathbb{R}^3$ as another example. We only need consider three permutations (1, 2, 3), (1, 3, 2) and (3, 1, 2).

Let us test the proposed IRT algorithm to construct the uniform design on the two-rectangle domain in Example 7.1. Using the LOO-Fibonacci design with $n = 20$ runs (see Fig. 7.8), we obtain the IRT construction results shown in Fig. 7.3. It turns out the permutation (x_1, x_2) leads to smaller CCD score than the permutation (x_2, x_1). In contrast, the AR method is also tested with 39-run uniform design on the outer rectangle with vertices $((0, 0), (5, 0), (5, 4), (0, 5))$. The accepted 20-run sub-design within the domain of interest has a relatively worse CCD score.

7.3 Construction Results

In this section, we present the construction results by the IRT method for four kinds of experimental domains in $\mathbb{R}^2$. Numerical comparisons of effectiveness are conducted between the proposed method and the existing SR and AR methods.

7.3.1 Flexible Regions

The flexible regions on $\mathbb{R}^2$ controlled by a shape parameter $m > 0$ are defined by

$$\mathscr{X}_F^{(m)} = \left\{(x_1, x_2) \in [0, 1]^2 : \ |2x_1 - 1|^m + |2x_2 - 1|^m \le 1\right\}. \tag{7.5}$$

Figure 7.4 shows the boundaries of such flexible regions with $m = \infty$, 2, 1 and 0.5, respectively. The circled points within each flexible region represent the constructed 88-run uniform designs by the proposed IRT method derived below.

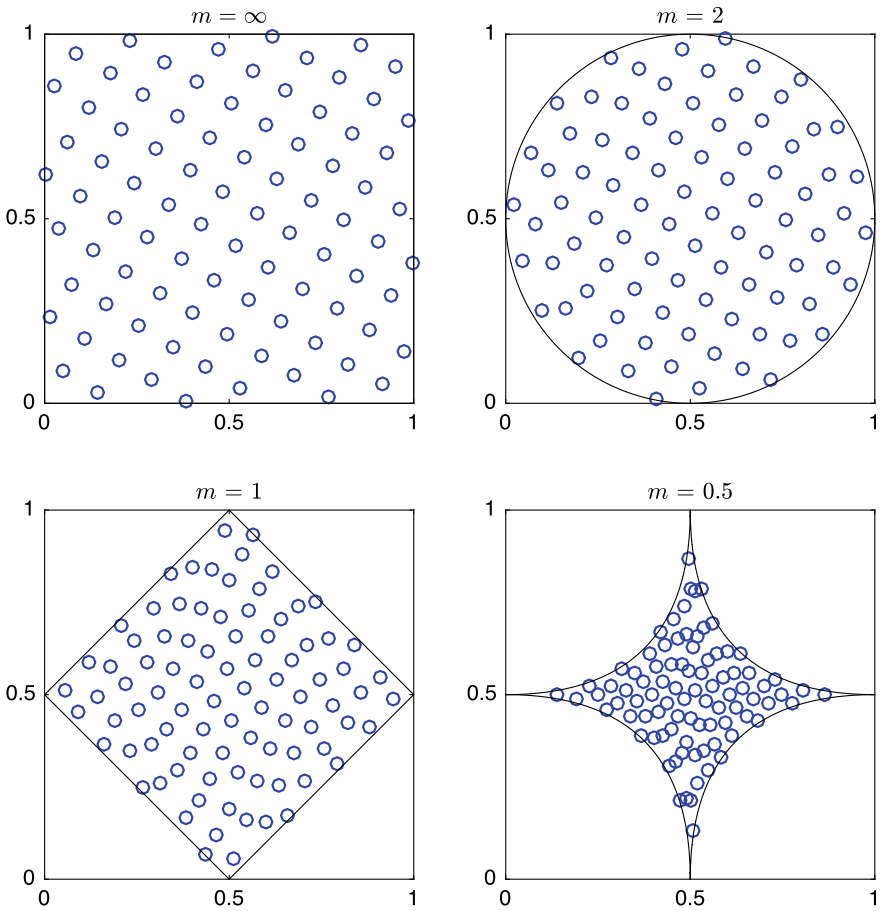


Fig. 7.4 Uniform designs with $n = 88$ runs on flexible regions with varying shape parameters. For $m = \infty$, the design is obtained by the GLP method (LOO-Fibonacci design) on the unit square (see Fig. 7.8). The design points in the $m = \infty$ case are then converted by the IRT method to the flexible regions for $m = 2$, 1 and 0.5

The flexible regions are symmetric in (x_1, x_2), so there is no need to consider permutations. For the random vector $X = (X_1, X_2) \sim \text{Uniform}(\mathscr{X}_F^{(m)})$, the marginal CDF $F_1(x)$ of the first component X_1 is

$$F_1(x) = \int_0^x \int_{0.5-(1-|2x_1-1|^m)^{1/m}/2}^{0.5+(1-|2x_1-1|^m)^{1/m}/2} \frac{1}{V(\mathscr{X}_F^{(m)})} dx_2 dx_1$$

where $V(\mathscr{X}_F^{(m)}) = \int_0^1 (1 - |2x_1 - 1|^m)^{1/m} dx_1$. Note that

$$\int_0^x (1-|2x_1-1|^m)^{1/m} dx_1$$
$$= \begin{cases} x, & m=\infty; \\ \frac{B\left(\frac{1}{m},\frac{1}{m}+1\right)}{2m}\left[1+\text{sign}(x-0.5)I_{|2x-1|^m}\left(\frac{1}{m},\frac{1}{m}+1\right)\right], & 0<m<\infty, \end{cases}$$

where $B(a,b)$ and $I_c(a,b)$ (with $a,b>0, c\in[0,1]$) are the values of Beta function and Incomplete Beta ratio with the following forms

$$B(a,b)=\int_0^1 t^{a-1}(1-t)^{b-1}dt, \quad I_c(a,b)=\frac{B_c(a,b)}{B(a,b)}=\frac{\int_0^c t^{a-1}(1-t)^{b-1}dt}{\int_0^1 t^{a-1}(1-t)^{b-1}dt}.$$

Therefore, the marginal CDF is given by

$$F_1(x)=\begin{cases} x, & m=\infty; \\ \frac{1}{2}+\frac{\text{sign}(x-0.5)}{2}\cdot I_{|2x-1|^m}\left(\frac{1}{m},\frac{1}{m}+1\right), & 0<m<\infty. \end{cases} \tag{7.6}$$

and its inverse is given by

$$F_1^{-1}(u)=\begin{cases} u, & m=\infty; \\ \frac{1}{2}+\frac{\text{sign}(u-0.5)}{2}\left(I^{-1}_{|2u-1|}\left(\frac{1}{m},\frac{1}{m}+1\right)\right)^{1/m}, & 0<m<\infty. \end{cases}$$

Moreover, the conditional CDF $F_{2|1}(x|x_1)$ of the second component given the value of the first component being x_1 is given by

$$F_{2|1}(x|x_1)=\begin{cases} x, & m=\infty; \\ \frac{x-\left(0.5-(1-|2x_1-1|^m)^{1/m}/2\right)}{(1-|2x_1-1|^m)^{1/m}}, & 0<m<\infty \end{cases} \tag{7.7}$$

and its inverse is given by

$$F^{-1}_{2|1}(u|x_1)=\begin{cases} u, & m=\infty; \\ 0.5+(u-0.5)(1-|2x_1-1|^m)^{1/m}, & 0<m<\infty. \end{cases}$$

Thus we obtain the analytical IRT $T^{-1}_{(1,2)}$ for $[0,1]^2\to\mathscr{X}_F^{(m)}$ as follows:

$$T^{-1}_{(1,2)}((u_1,u_2))=\left(F_1^{-1}(u_1),\ F^{-1}_{2|1}\left(u_2|F_1^{-1}(u_1)\right)\right). \tag{7.8}$$

The effectiveness of the IRT method can be compared with traditional AR and SR methods. We use the CCD criterion to evaluate the uniform designs for $n=10,20,\ldots,100$ runs based on different construction methods. For the AR method, for each target n, we search the uniform designs on C^2 with sizes $n+1,n+2,\ldots$ in order to find such a design with exactly n runs falling into $\mathscr{X}_F^{(m)}$. Note that such AR method has the chance to find no appropriate design with the target number of

runs. For the SR method, for $m = 2$, the method by Fang and Wang [9] is employed; for $m = 1$, the method by Tian et al. [16] is employed; for $m = 0.5$, there exists no SR method in the literature. All the needed uniform designs in C^2 are generated by the mixed GLP method (see Appendix). The numerical results for flexible regions with $m = 2, 1, 0.5$ are listed in Table 7.1. It can be found that IRT shows competitive performances in most cases.

Table 7.1 CCD scores for uniform designs constructed on the flexible regions with $m = 2, 1, 0.5$

m	n	AR	SR	IRT
2	10	–	0.054329	0.048877
	20	0.027465	0.030055	0.025930
	30	0.020286	0.024471	0.020323
	40	0.014839	0.021816	0.017132
	50	0.013721	0.018935	0.013708
	60	0.012722	0.015926	0.013522
	70	0.012529	0.013691	0.012670
	80	0.011813	0.013876	0.011595
	90	–	0.013461	0.011708
	100	0.012388	0.012752	0.011157
1	10	0.041046	0.047865	0.044335
	20	0.033795	0.045915	0.026233
	30	–	0.018906	0.021509
	40	0.018492	0.016612	0.018882
	50	0.012644	0.025909	0.015558
	60	0.015144	0.014044	0.015588
	70	0.015771	0.013879	0.014514
	80	0.015241	0.013489	0.014223
	90	0.017004	0.021852	0.013704
	100	0.012856	0.014312	0.013199
0.5	10	0.048221	–	0.048607
	20	0.035220	–	0.032163
	30	0.034672	–	0.027973
	40	0.028218	–	0.028656
	50	0.033651	–	0.023997
	60	0.025663	–	0.025844
	70	0.029249	–	0.025079
	80	0.024936	–	0.025046
	90	0.024674	–	0.024543
	100	0.020609	–	0.025098

7.3.2 Constrained Domain

Tian et al. [16] studied a tetragon shape of constrained domain for drug combination experiment, as defined by

$$\mathscr{X}_T = \left\{(x_1, x_2) \in \mathbb{R}_+^2 : 20 < 101.91 - 31.17x_1 - 9.56x_2 < 80\right\}. \tag{7.9}$$

They constructed a 19-run uniform design on this domain by the SR method, as shown in the left panel of Fig. 7.5. In this section, we apply the proposed IRT method for constructing a competitive uniform design on this specific constrained domain with the same number of runs.

First we can convert the domain $\mathscr{X}_T$ to the following symmetric domain

$$\mathscr{X}_{\tilde{T}} = \left\{(\tilde{x}_1, \tilde{x}_2) \in \mathbb{R}_+^2 : 21.91 < \tilde{x}_1 + \tilde{x}_2 < 81.91\right\},$$

where $\tilde{x}_1 = 31.17x_1$ and $\tilde{x}_2 = 9.56x_2$. Write $c_1 \equiv 21.91$ and $c_2 \equiv 81.91$, then it is easy to get the marginal CDF:

$$F_1(\tilde{x}_1) = \begin{cases} \frac{2\tilde{x}_1}{c_1 + c_2}, & \text{if } 0 \le \tilde{x}_1 \le c_1; \\ 1 - \frac{(c_2 - \tilde{x}_1)^2}{c_2^2 - c_1^2}, & \text{if } c_1 < \tilde{x}_1 \le c_2. \end{cases} \tag{7.10}$$

For each fixed $\tilde{x}_1$, the conditional CDF $F_{2|1}(\tilde{x}_2|\tilde{x}_1)$ is given by the uniform distribution with the range $[c_1 - \tilde{x}_1, c_2 - \tilde{x}_1]$ if $\tilde{x}_1 \in [0, c_1]$ and the range $[0, c_2 - \tilde{x}_1]$ if $\tilde{x}_1 \in [c_1, c_2]$.

Using the IRT method based on a 19-run uniform design on C^2, we first obtain the transformed design on $\mathscr{X}_{\tilde{T}}$, then convert the design points back to $\mathscr{X}_T$ as plotted in the right panel of Fig. 7.5. This new design is more uniform than Tian et al. [16]'s result according to the CCD criterion.

7.3.3 Manifold Domain

Other than the flexible regions discussed in the previous section, we consider another special manifold domain defined by the ring constraint:

$$\mathscr{X}_R = \left\{(x_1, x_2) \in \mathbb{R}^2 : \frac{1}{4} \le x_1^2 + x_2^2 \le 1\right\}. \tag{7.11}$$

To get the marginal and conditional CDFs, instead of striving to derive the analytical forms, we adopt the following approximation method:

(1) Partition the x_1-coordinate from $[-1, 1]$ into $N = 1000$ equal-spaced intervals, each interval with mid-point $z_k = (2k - 1)/N - 1$ for $k = 1, \ldots, N$.

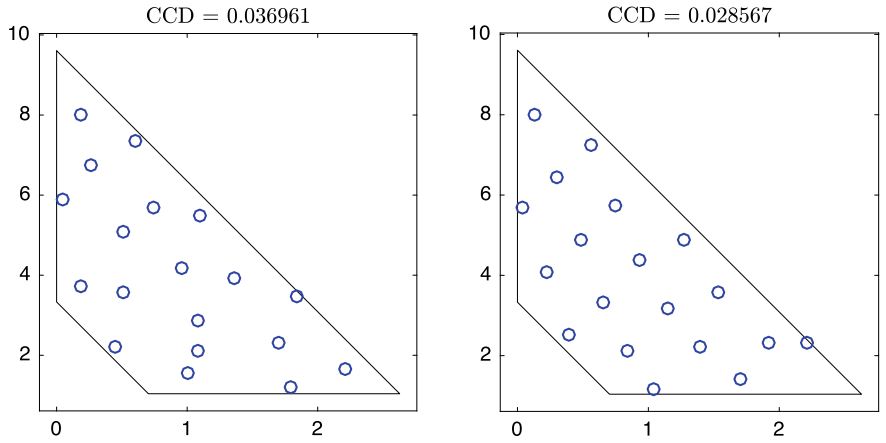


Fig. 7.5 Uniform designs with 19 runs on a constrained domain (with x_2 shifted 1.04 units upward). Left panel: SR method by Tian et al. [16]; Right panel: IRT method

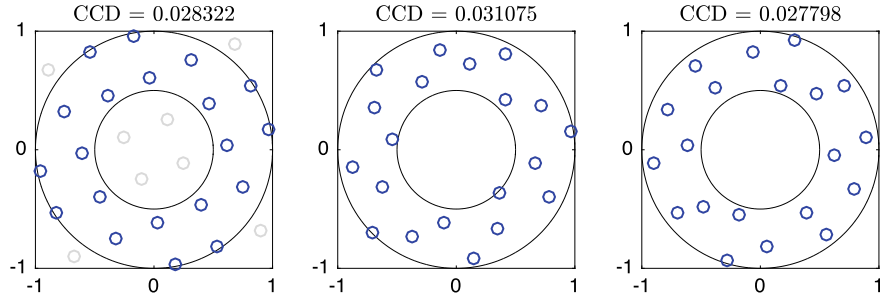


Fig. 7.6 Uniform designs with 20 runs on the ring domain. Left panel: AR method; Center panel: SR method by Zhang [19]; Right panel: IRT method

(2) Obtain the approximate marginal CDF for x_1 based on the midpoints:

$$\widehat{F}_1(z_k) = \frac{\sqrt{1-z_k^2} - \sqrt{1/4 - z_k^2} \cdot I\left(|z_k| \le \frac{1}{2}\right)}{\sum_{k=1}^{N} \left(\sqrt{1-z_k^2} - \sqrt{1/4 - z_k^2} \cdot I\left(|z_k| \le \frac{1}{2}\right)\right)}, \quad k = 1, \ldots, N. \tag{7.12}$$

(3) When x_1 takes discretized z_k-values, obtain the conditional CDF for $x_2|x_1$ by the uniform distribution with the range $\left[-\sqrt{1-z_k^2}, \sqrt{1-z_k^2}\right]$ if $|z_k| \ge 1/2$ or the range $\left[-\sqrt{1-z_k^2}, -\sqrt{1/4-z_k^2}\right] \cup \left[\sqrt{1/4-z_k^2}, \sqrt{1-z_k^2}\right]$ if $|z_k| \le 1/2$.

Figure 7.6 (right panel) shows the IRT constructed 20-run uniform design on the ring domain. In contrast, on the left panel is the result by the AR method based on

Table 7.2 CCD scores for uniform designs constructed on the ring-shaped domain

n	AR	SR	IRT
10	0.051980	0.048791	0.047818
20	0.028322	0.031075	0.027798
30	0.016596	0.021944	0.017866
40	0.013043	0.017600	0.015126
50	0.013418	0.015682	0.014486
60	0.011472	0.012808	0.011033
70	0.009177	0.011868	0.010474
80	0.011507	0.011086	0.009580
90	0.009958	0.009751	0.009025
100	0.012203	0.009791	0.007863

28-run uniform design on the unit cube. On the center panel is the result by the SR method through $x_{i1} = \sqrt{u_{i1}}\sin(2\pi u_{i2})$, $x_{i2} = \sqrt{u_{i1}}\cos(2\pi u_{i2})$; see Zhang [19]. It is found that in this case the IRT outperforms AR and SR methods. Moreover, we run through $n = 10, 20, \ldots, 100$ to compare the three methods, with numerical results presented in Table 7.2. We can see that the IRT method always outperforms the SR method, and sometimes have better performance than the AR method.

7.3.4 Geographical Domain

Lastly, we consider the geographical domain that is usually rather irregular. In this section we consider the Land Map of China as the experimental domain $\mathscr{X}_{\text{map}}$. The entire domain consists of several closed subregions. It is difficult to determine the exact form of Rosenblatt transformation, so we use the approximated CDFs.

To approximate the marginal and conditional CDFs on the map domain, we find a rectangle to completely cover $\mathscr{X}_{\text{map}}$ and establish a cartesian coordinate system. The rectangle has the resolution of 1297×1083 pixels, and the contour of $\mathscr{X}_{\text{map}}$ contains $N = 675328$ pixels in total. The marginal CDF of x_1 is approximated by

$$\widehat{F}_1(x_1) = \frac{1}{N}\sum_{i=1}^{N} I(x_{i1} \le x_1), \quad x_1 = z_1, \ldots, z_{1297}. \tag{7.13}$$

For $x_1 = z_1, \ldots, z_{1297}$, the conditional CDF of $x_2|x_1$ is approximated by

$$\widehat{F}_{2|1}(x_2|x_1 = z_k) = \frac{1}{|\Omega_k|}\sum_{j\in\Omega_k} I(x_{j2} \le x_2), \tag{7.14}$$

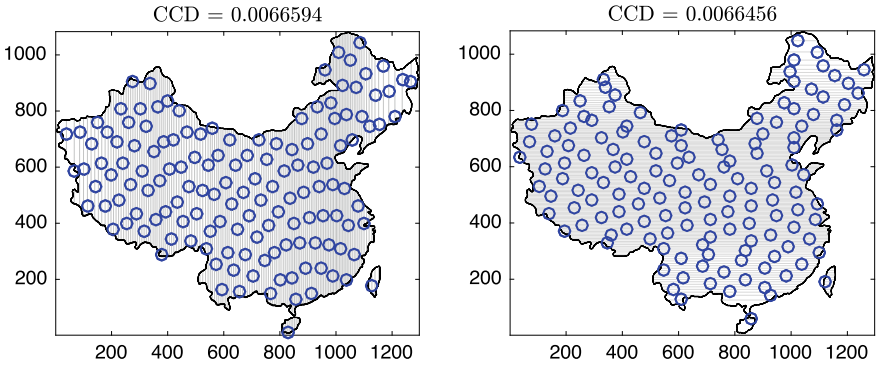


Fig. 7.7 Uniform designs with 143 runs on China map domain. Left panel: IRT method with permutation (1, 2); Right panel: IRT method with permutation (2, 1). The vertical and horizontal gray lines represent the partitions of the map for approximating the conditional CDFs

where Ω_k denotes the subset of pixels $\{j : x_{j1} = z_k\}$. Similarly for permutation (x_2, x_1), we can obtain $\widehat{F}_2(x_2)$ and $\widehat{F}_{1|2}(x_1|x_2)$.

Suppose we are given a 143-run LOO-Fibonacci design (see Fig. 7.8). We may use the IRT method with respect to permutations (x_1, x_2) and (x_2, x_1) to construct the corresponding uniform designs on $\mathscr{X}_{\text{map}}$. The construction results are visualized in Fig. 7.7, with the permutation (x_2, x_1) leading to a slightly lower CCD score. In each permutation, there are two points to represent the Hainan and Taiwan islands on the map, respectively.

7.4 Conclusion

The construction of uniform designs on irregular regions has been a relatively challenging task as compared with the case on regular regions. Inspired by the stochastic representation method in Fang and Wang [9], we propose to construct uniform designs on arbitrary domains by the inverse Rosenblatt transformation (IRT) based on marginal and conditional distributions. We have demonstrated how to use this method in multiple kinds of experimental domains in two-dimensional space, and the construction results are rather competitive and promising.

There are several interesting problems that are worth further study. First, the IRT method is proposed for not only two-dimensional space, but also higher dimensional space. In the latter case it is however computationally demanding. It is important to develop a highly efficient algorithm for approximating the marginal and conditional distribution functions. Second, there exist other manifold domains than the ring shape and flexible regions, e.g. the sphere and donut kinds of surfaces in high-dimensional space. It is interesting to extend the IRT method to these manifold cases. Third, in this study we find that the central composite discrepancy is an imperfect measure

of uniformity on arbitrary domains. For example, it lacks the property of invariance under rotation. It is among our research plans to develop a better kind of discrepancy measure for space-filling designs on arbitrary domains with general distributional assumption.

Acknowledgements This work was supported by National Natural Science Foundation of China (11871288) and Natural Science Foundation of Tianjin (19JCZDJC31100).

Appendix: Good Lattice Point Method

The uniform designs constructed on the unit hypercubes by the GLP method are also known as the NT-nets [9], which uses the classical star-discrepancy for evaluating the uniformity of the candidate designs. In Algorithm 4 we write the GLP method using the centered-ℓ_2 discrepancy (CD2), a more popular criterion proposed by Hickernell [11]. Meanwhile, it is easy to check that the GLP designs (7.15) always include a point $\boldsymbol{x}_n = (1 - 1/2n, \ldots, 1 - 1/2n) \in C^s$. The leave-one-out (LOO) GLP method is to remove such a dummy point, then scale the remaining points by $n/(n-1)$ in all coordinates. Thus, in order to construct an n-run uniform design, we can use a mixed GLP method by selecting the lower-CD2 design between the GLP (with input n) and LOO-GLP (with input $n+1$) outputs.

It is well-known that for $s = 2$ and $n = F_k$ (Fibonacci numbers 5, 8, 13, 21, ...), the lattice designs generated by $h_1 = 1$ and $h_2 = F_{k-1}$ enjoy the remarkable low star-discrepancy property [18]. It is of our interest to investigate whether such Fibonacci designs may also attain low discrepancy with respect to the CD2 criterion. As a key difference, the star-discrepancy is anchored at the origin of the unit hypercube, while the CD2 is anchored at the center. It turns out the Fibonacci designs are sub-optimal under CD2. Nevertheless, we find that the LOO-Fibonacci designs with $n = F_k - 1$ ($F_k \leq 1597$) runs remarkably minimize the CD2 criterion among all the generating vectors for the mixed GLP method. See Fig. 7.8 about the LOO-Fibonacci designs with 20, 88 and 143 runs. See Table 7.3 for the numerical results based on

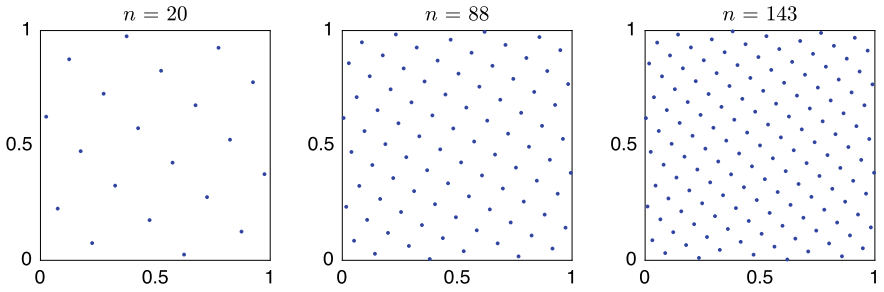


Fig. 7.8 Scatter plots of LOO-Fibonacci designs for $n = 20$, 88 and 143 runs

Algorithm 4 Good Lattice Point Method

Input: The number of factors s, and the number of runs n.

1: Form a generating vector $(h_1, h_2, \ldots, h_s)$ by choosing distinct positive integers that are less than and relatively prime to n.

2: Form the n-run lattice design $\boldsymbol{X} = [x_{ij}]_{n\times s}$ with entries

$$x_{ij} = \left\{\frac{2ih_j - 1}{2n}\right\},\ i = 1, \ldots, n,\ \ j = 1, \ldots, s \tag{7.15}$$

where $\{z\}$ is the factorial part of z.

3: Evaluate the criterion of the centered-ℓ_2 discrepancy

$$\mathrm{CD}_2(\boldsymbol{X}) = \left\{\left(\frac{13}{12}\right)^s - \frac{2}{n}\sum_{i=1}^{n}\prod_{j=1}^{s}\left(1 + \frac{1}{2}\left|x_{ij} - \frac{1}{2}\right| - \frac{1}{2}\left|x_{ij} - \frac{1}{2}\right|^2\right)\right.$$
$$\left. + \frac{1}{n^2}\sum_{i,k=1}^{n}\prod_{j=1}^{s}\left(1 + \frac{1}{2}\left|x_{ij} - \frac{1}{2}\right| + \frac{1}{2}\left|x_{kj} - \frac{1}{2}\right| - \frac{1}{2}|x_{ij} - x_{kj}|\right)\right\}^{1/2} \tag{7.16}$$

4: Repeat Steps 1–3 for all distinct generating vectors. Output $\boldsymbol{X}^*$ with the lowest CD_2 value.

Table 7.3 LOO-Fibonacci designs with $h_1 = 1$ and $h_2 = F_{k-1}$ minimize the CD2 criterion for the mixed GLP method, where h_2^* represents the best found generating vectors per each method

$n = F_k - 1$	$h_2 = F_{k-1}$	h_2^* (LOO-GLP)	min-CD2	h_2^* (GLP)	min-CD2
4	3	2, 3	1.275E–01	3	1.350E–01
7	5	3, 5	7.631E–02	3, 5	8.122E–02
12	8	5, 8	4.557E–02	5	5.058E–02
20	13	8, 13	2.843E–02	9	3.133E–02
33	21	13, 21	1.764E–02	14, 26	1.947E–02
54	34	21, 34	1.117E–02	35	1.288E–02
88	55	34, 55	7.010E–03	37	7.661E–03
143	89	55, 89	4.456E–03	63	4.823E–03
232	144	89, 144	2.806E–03	147	3.115E–03
376	233	144, 233	1.784E–03	165	1.916E–03
609	377	233, 377	1.123E–03	256	1.224E–03
986	610	377, 610	7.128E–04	579	7.600E–04
1596	987	610, 987	4.484E–04	617	4.859E–04

exhaustive search up to $F_k = 1597$. From Table 7.3, it can be found that the LOO-Fibonacci designs with $n = F_k - 1$ also include $h_1 = 1$ and $h_2 = F_{k-2}$ as the optimal generating vector. This can be actually justified by the reflection-invariant property of the CD2 criterion.

References

1. Arnold, B.C., Castillo, E., Sarabia, J.M.: Families of multivariate distributions involving the Rosenblatt construction. J. Am. Stat. Assoc. **101**(476), 1652–1662 (2006)
2. Borkowski, J.J., Piepel, G.F.: Uniform designs for highly constrained mixture experiments. J. Qual. Technol. **41**, 35–47 (2009)
3. Chen, R.B., Shu, Y.H., Hung, Y., Wang, W.: Central composite discrepancy-based uniform designs for irregular experimental regions. Comput. Stat. Data Anal. **72**, 282–297 (2014)
4. Chen, R.B., Li, C.H., Hung, Y., Wang, W.: Optimal noncollapsing space-filling designs for irregular experimental regions. J. Comput. Graph. Stat. **28**(1), 74–91 (2019)
5. Chuang, S.C., Hung, Y.: Uniform design over general input domains with applications to target region estimation in computer experiments. Comput. Stat. Data Anal. **54**, 219–232 (2010)
6. Fang, K.T.: The uniform design: application of number theoretic methods in experimental design. Acta Math. Appl. Sin. **3**, 363–372 (1980)
7. Fang, K.T., Li, R.Z., Sudjianto, A.: Design and Modeling for Computer Experiments. Chapman and Hall/CRC, Boca Raton, FL (2006)
8. Fang, K.T., Liu, M.Q., Qin, H. and Zhou, Y.D.: Theory and Application of Uniform Experimental Designs. Springer, Singapore (2018)
9. Fang, K.T., Wang, Y.: Number-Theoretic Methods in Statistics. Chapman and Hall, London (1994)
10. Fang, K.T., Yang, Z.H.: On uniform design of experiments with restricted mixtures and generation of uniform distribution on some domains. Stat. Probab. Lett. **46**, 113–120 (2000)
11. Hickernell, F.: A generalized discrepancy and quadrature error bound. Math. Comput. **67**(221), 299–322 (1998)
12. Justel, A., Pena, D., Zamar, R.: A multivariate Kolmogorov-Smirnov test of goodness of fit. Stat. Probab. Lett. **35**(3), 251–259 (1997)
13. Lin, D.K.J., Sharpe, C., Winker, P.: Optimized U-type designs on flexible regions. Comput. Stat. Data Anal. **54**, 1505–1515 (2010)
14. Niederreiter, H.: Random Number Generation and Quasi-Monte Carlo Methods. SIAM, Philadelphia (1992)
15. Rosenblatt, M.: Remarks on a multivariate transformation. Ann. Math. Stat. **29**, 470–472 (1952)
16. Tian, G.L., Fang, H.B., Tan, M., Qin, H., Tang, M.L.: Uniform distributions in a class of convex polyhedrons with applications to drug combination studies. J. Multivar. Anal. **100**, 1854–1865 (2009)
17. Wang, Y., Fang, K.T.: A note on uniform distribution and experimental design. Kexue Tongbao **26**, 485–489 (1981)
18. Zaremba, S.K.: Good lattice points, discrepancy, and numerical integration. Annali di matematica pura ed applicata **73**(1), 293–317 (1966)
19. Zhang, R.C.: On a transformation method in constructing multivariate uniform designs. Stat. Sinica **6**, 455–469 (1996)

Chapter 8
Drug Combination Studies, Uniform Experimental Design and Extensions

Ming T. Tan and Hong-Bin Fang

Abstract Drug combination has been an important therapeutic development approach for cancer, viral or microbial infections, and other diseases involving complex biological networks. Synergistic drug combinations, which are more effective than predicted from summing effects of individual drugs, often achieve improved therapeutic index. Because drug-effect is dose-dependent, multiple doses of an individual drug need to be evaluated, giving rapidly escalating number of combinations and a challenging high dimensional statistical modeling problem. The lack of proper design and analysis methods for multi-drug combination studies have resulted in many missed therapeutic opportunities. It is known that, in the presence of model uncertainties, uniform measures that scatter the design points (the dose levels) uniformly in the experiment domain is the best strategy to yield maximum information on the dose response relation. This chapter will review some efficient experimental designs for drug combination studies especially those related to uniform measures and extensions using maximum entropy design.

8.1 Introduction

Drug combination has played an important role in developmental therapeutics for cancer, viral or microbial infections, and other diseases involving complex biological pathways since most mono-chemotherapy results in drug resistance and toxicity with high dose levels. Synergistic drug combinations, which are more effective than predicted from summing the effects of individual drugs, often achieve greater efficacy at lower doses with less toxicity [21]. The joint action of two drugs is a fundamental problem in drug discovery and has a long history in pharmacology and

M. T. Tan (✉) · H.-B. Fang
Department of Biostatistics, Bioinformatics and Biomathematics,
Georgetown University Medical Center, Washington, DC 20057, USA
e-mail: mtt34@georgetown.edu

H.-B. Fang
e-mail: hf183@georgetown.edu

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_8

biostatistics [2, 20, 35]. Increasing the number of constituent agents in a combination has been another strategy for increasing the level or type of interaction produced [3, 23, 42]. There has been growing interest in developing quantitative methods for the fundamental problem of detecting drug synergy [6, 18, 30]. The past several decades have seen significant progresses in developing proper design and analysis methods for multi-drug combination studies, which have provided a rational approach to increase the chance of identifying optimized combinations of two or more drugs for further evaluations. The approach utilizes optimized designs and systems biology (see, e.g., [4, 10–13, 37, 38]) as well as adaptive phase I clinical trial designs that attempt to identify the best possible maximum tolerated doses through modeling the joint dose-toxicity relationship (see, e.g., [43–46]).

Because dose effect is known to be variable among virtually genetically identical animals (or even different aliquots) receiving precisely the same dose, the design for drug combination study becomes very important. Finney [17] proposed that the regression lines for the mixtures should be equally spaced between those for the drugs under the assumption of additive joint action. This design is modified by Tallarida et al. [36] by fixing the mixture ratio for a selected dose effect level to reduce the variance of estimated dose response. An optimal design is proposed by Abdelbasit and Plackett [1] by fixing the total dose in a specific model and optimizing the selection of the other parameter: the mixing proportions. Laska et al. [25] provided a design and analysis for a combination study with fixed individual doses. To assess the joint action of two drugs at different ratios, the ray design has been proposed where a ray corresponds to one fixed ratio [19, 33]. Based on parametric models, Meadows et al. [28] and Casey et al. [5] proposed a ray design for multiple drug combination studies. Often the ray design is repeated at multiple fixed ratios, however, it is not clear how many and what rays ought to be chosen. Thus, the design may require a large sample size but still be under-powered in detecting departures from additivity. Assuming the dose-response curves of individual drugs can be characterized by Hill models, Syracuse and Greco [34] proposed an equation to characterize interactions of two drugs [20, 34]. Carter et al. [7] proposed the dose-response surface for the assessment of combination drug synergy.

Since the model of the joint action is typically not well specified before experiment, a space-filling-type design is preferred [8]. However, the constituent doses of the combination are to be found instead of being given. Thus, a design that allows dose selection would be desired. When two drugs are applied in combination to a biological system, their joint effect can be either additive, synergistic, or antagonistic as compared to what is to be expected from the biological activity of the single drugs. Recognizing these unique features, we proposed a general statistical framework with semiparametric models based on the shapes of the single drug dose-response curves; proved mathematically that if the design points are chosen according to uniform measures, even with moderate sample sizes, the dose response can be estimated adequately; and obtained the asymptotic properties of the interaction index function [10, 40]. The design is derived by the uniform measures that maximize the power to detect any overall possible departures of a given magnitude from additivity while minimizing the lack of fit of the model for joint action [37, 38]. It was a pleasant surprise that

the number of combinations and replications generated by such design is moderate and highly feasible [10, 12, 13]. However, the extension of the method to multi-drug turns out to be a challenge (shown later in this article) because of the increased complexity in the additive model and in obtaining uniform scattered points in high-dimensional dose regions. We have proposed a novel method to screen the large number of combinations and identify the functional structure of the dose response relationship by using the dose response data of single drugs and pathway/network knowledge and the corresponding designs based on maximum entropy.

We give an overall review of the experimental designs for drug combination studies in this chapter. Section 8.2 introduces the general statistical model for the joint action of drugs and proposes an F-statistic to test if additive action. The designs for two- or three-drug combination studies are described in Sect. 8.3 according to various individual dose-response curves and utilizing uniform measures. The design procedure for high dimensional multidrug studies utilizing systems biology and maximum entropy are introduced in Sect. 8.4. Conclusion and further research are given in Sect. 8.5.

8.2 Statistical Modeling for Drug Combinations

Experimental approaches to characterizing combination therapy typically involve determining dose-response curves for inhibitors individually and in combination. When experimental dose-response data match the predictions of Loewe additivity, the inhibitors are said to be additive (corresponding to the zero-interaction case); greater than predicted potency indicates synergism (positive interaction); and lower potency argues for antagonism (negative interaction). The Loewe additivity is embodied in the isobologram method for characterizing departures from additivity. To describe the joint action of two drugs A and B at a specific dose level, the additivity of Loewe [27] is based on single drug dose-effect and is defined by the following isobole equation

$$\frac{x_A}{X_A} + \frac{x_B}{X_B} = \tau \tag{8.1}$$

where x_A and x_B are doses of the constituent drugs A and B of the combination needed to yield a given level of effect, e.g., 50% inhibition (ED50), or 50% death in experiment animals (LD50), where X_A and X_B are the doses needed for each drug alone to yield the level of effect. The τ is called the interaction index of the drugs A and B at the combination (x_A, x_B). When $\tau = 1$, the drugs A and B is additive (zero-interaction) at the combination (x_A, x_B); when $\tau < 1$, they are synergistic, namely, the combination (x_A, x_B) is more effective than expected from their single drug dose-response curves, otherwise ($\tau > 1$), they are antagonistic. According to Loewe's definition, the isoboles (isoeffect equation) of the k drugs $A_1, \ldots, A_k$ at combination $\mathbf{x} = (x_1, x_2, \ldots, x_k)^T$ is defined as (see (1a) in Berenbaum [2]),

$$\frac{x_1}{X_1} + \frac{x_2}{X_2} + \cdots + \frac{x_k}{X_K} = \tau, \tag{8.2}$$

where X_i is the dose of drug A_i $(i = 1, \ldots, k)$ alone that yields the same response as the combination $\mathbf{x} = (x_1, x_2, \ldots, x_k)^T$. Denote the combination dose-effect (response) by $y = f_{com}(x_1, \ldots, x_k)$ and dose-response relationships for individual drugs

$$y = f_i(X_i), \quad i = 1, \ldots, k, \tag{8.3}$$

where y is the dose-effect scaled to be a viability (proportion of cells surviving) or a tumor volume (with some transformation) and $f_i(X_i)$ is assumed to be an decreasing function of X_i in the dose range of interest. Then,

$$f_{com}(x_1, \ldots, x_k) = f_1(X_1) = \cdots = f_k(X_k). \tag{8.4}$$

The potency of drug A_i relative to drug A_1 is the ratio of isoeffective doses of A_1 and A_i, $\rho_i(X_i) = X_1/X_i$ where $f_i(X_i) = f_1(X_1)$, that is,

$$\rho_i(X_i) = f_1^{-1}\left(f_i(X_i)\right)/X_i, \quad i = 1, \ldots, k. \tag{8.5}$$

From (8.2) and (8.4), we have that

$$f_{com}(x_1, \ldots, x_k) = f_1\left(x_1 + \frac{X_1}{X_2}x_2 + \cdots + \frac{X_1}{X_k}x_k\right) + [f_1(X_1) - f_1(\tau X_1)]. \tag{8.6}$$

and the term $[f_1(X_1) - f_1(\tau X_1)] = 0$ if the joint action of $A_1, \ldots, A_k$ is additive. Then, the regression line for the combination with additive action of k drugs is

$$\begin{aligned} y &= f_1\left(x_1 + \frac{X_1}{X_2}x_2 + \cdots + \frac{X_1}{X_k}x_k\right) \\ &= f_1\left(x_1 + \rho_2(X_2)x_2 + \cdots + \rho_k(X_k)x_k\right), \end{aligned} \tag{8.7}$$

and $\rho_i(Xi)$ is a function of $(x_1, \ldots, x_k)$ determined by (8.2)–(8.5). If the potency ρ_i in (8.5) is not a constant, the additive model (8.7) has no closed forms.

Since we generally know little about the joint effect of the combinations before experiments, we consider a general semiparametric model for the joint effect of the k drugs in the experimental domain $\mathscr{D}_0$,

$$y = f_1\left(x_1 + \rho_2(X_2)x_2 + \cdots + \rho_k(X_k)x_k\right) + f(x_1, \ldots, x_k) + \varepsilon \tag{8.8}$$

where the function f is unspecified, ε is the error term due to variation in experiments and is assumed to be normally distributed with mean 0 and variance σ^2. Then, testing the additive action of the k compounds is equivalent to testing the hypothesis H_0: $f = 0$.

Assume that the additive model (8.7) can be expressed in a generalized additive structure and model (8.8) becomes

$$y = (\approx)\alpha_1 g_1(z_1) + \cdots + \alpha_k g_k(z_k) + g(z_1, \ldots, z_k) + \varepsilon, \tag{8.9}$$

with an one-to-one invertible transformation: $(x_1, \ldots, x_k)^T \in \mathscr{D}_0 \mapsto (z_1, \ldots, z_k)^T \in \mathscr{D}$ by $z_i = \phi_i(x_1, \ldots, x_k)$ $(i = 1, 2, \ldots, k)$ such that the functions $g_1, \ldots, g_k$ are linearly independent, $g(z_1, \ldots, z_k) = f(x_1, \ldots, x_k)$ and satisfies the following orthogonality condition:

$$\int_{\mathscr{D}} G(z_1, \ldots, z_k) g(z_1, \ldots, z_k) dz_1 \cdots dz_k = \mathbf{0}, \tag{8.10}$$

where $G(z_1, \ldots, z_k) = (g_1(z_1), \ldots, g_k(z_k))^T$.

As shown in Wiens [41], the test statistic for H_0: $g = 0$ can be derived using the lack-of-fit test involving least square error estimates under the full model (8.9) and the additive model (8.7). Assume that the m points in the experimental domain $\mathscr{D}$ are $\mathbf{z}^{(1)}, \ldots, \mathbf{z}^{(m)}$, and there are n_i experiments at the dose-level $\mathbf{z}^{(i)} = (z_1^{(i)}, \ldots, z_k^{(i)})^T \in \mathscr{D}$ with corresponding responses y_{ij}, $j = 1, \ldots, n_i$, $i = 1, \ldots, m$. Denote $n = n_1 + \ldots + n_m$. Let $\mathbf{y}$ be the $n \times 1$ vector with elements y_{ij} ordered lexicographically and $\mathbf{1}_k$ be the $k \times 1$ vector of one. Let Z be the $m \times k$ matrix with ith row $(g_1(z_1^{(i)}), \ldots, g_k(z_k^{(i)}))$, where g_i is given by (8.9). Denote $V = UZ(Z^T U^T U Z)^{?} Z^T U^T$, $J = U(U^T U)^{?} UT$ and $U = diag(\mathbf{1}_{n_1}, \ldots, \mathbf{1}_{n_m})$. Then, if hypothesis H_0 is true (i.e, the joint action of the k drugs is additive), the statistic for the test of lack of fit

$$F = \frac{\mathbf{y}^T (J - V)\mathbf{y}/(m - k)}{\mathbf{y}^T (I - J)\mathbf{y}/(n - m)} \tag{8.11}$$

has a central F-distribution with degrees of freedom $m - k$ and $n - m$ (see [41]). When the alternative hypothesis $H_1 : g \neq 0$ holds, the statistic (8.11) has a noncentral F-distribution with degrees of freedom $m - k$ and $n - m$, and the noncentrality parameter

$$\delta = \frac{n}{\sigma^2} \left[\int_{\mathscr{D}} g^2(\mathbf{z}) d\xi(\mathbf{z}) - \int_{\mathscr{D}} G^T(\mathbf{z}) g(\mathbf{z}) d\xi(\mathbf{z}) \left(\int_{\mathscr{D}} G(\mathbf{z}) G^T(\mathbf{z}) d\xi(\mathbf{z}) \right)^{-1} \int_{\mathscr{D}} G(\mathbf{z}) g(\mathbf{z}) d\xi(\mathbf{z}) \right] \tag{8.12}$$

where ξ is the design measure, which is a probability distribution function with mass $p_i = n_i/n$ at $\mathbf{z}^{(i)}$, $i = 1, \ldots, m$. Furthermore, the noncentrality parameter δ is maximized if the design measure ξ is uniform on $\mathscr{D}$ [38].

8.3 Experimental Design Based on Uniform Measures

The uniform design measure ξ maximizes the minimum power of the F-test (8.11) and minimizes the maximum bias in the estimation of σ^2 in (8.12). Suppose that the experimental points are uniformly scattered on the domain $\mathscr{D}$ and the number of runs (experiment units) at each point is the same. The sample sizes (number of experimental units) to detect a given meaningful synergism or antagonism (η) can be calculated at a given significance level (α) and a given power level ($1-\beta$) based on a noncentral F-distribution with degrees of freedom $m-k$ and $n-m$ and the noncentral parameter $\delta = n\int_{\mathscr{D}} g^2(\mathbf{z})d\mathbf{z}/\sigma^2 = n\eta^2 Vol(\mathscr{D})/\sigma^2$.

An important aspect is that when we plan experiments on drug combinations, we already have single agent dose response data and these data need to be fully utilized in the combination studies. Since the additivity effect in Eq. (8.7) is dependent on the single drug dose response curves, it is critical to the uniform design method [14, 16] to be able to derive an approximation of the additive model in Eq. (8.9) and to obtain uniform scattered points in the experimental domain. Different classes of drugs may have different dose response curves. According to Berenbaum [2] the single drug dose response for a given dose interval (for example, $ED_{20} - ED_{80}$) can be fitted by log-linear or linear curve. Then, experimental designs for combination studies of three drugs can be divided into four classes (three classes for two-drug combinations). In this section, we give experimental designs for the four classes.

8.3.1 Design for Log-Linear Dose-Responses

Let the single dose response curves of drugs A, B, and C be

$$\begin{aligned} y(X_A) &= \alpha_A + \beta_A \log(X_A), \\ y(X_B) &= \alpha_B + \beta_B \log(X_B), \\ y(X_C) &= \alpha_C + \beta_C \log(X_C), \end{aligned} \tag{8.13}$$

respectively. Without loss of generality, we assume that $\beta_C \le \beta_B \le \beta_A$. The potencies $\rho(X_B)$ and $\rho(X_C)$ of B and C relative to A are

$$\rho(X_B) = \rho_0 X_B^{\beta_B/\beta_A - 1}, \quad \rho(X_C) = \rho_1 X_C^{\beta_C/\beta_A - 1},$$

respectively, where $\rho_0 = \exp[(\alpha_B - \alpha_A)/\beta_A]$ and $\rho_1 = \exp[(\alpha_C - \alpha_A)/\beta_A]$.

When $\beta_C = \beta_A$, the potencies $\rho(X_B)$ and $\rho(X_C)$ are constant and equal to ρ_0 and ρ_1, respectively. In this case, the additive model at combination dose (x_A, x_B, x_C) is

$$\begin{aligned} y(x_A, x_B, x_C) = \alpha_A &+ \beta_A \log(z_1) + \beta_A \log[(1-\rho_0)z_2 + \rho_0] \\ &+ \beta_A \log\left[\left(1 - \frac{\rho_1}{\rho_0}\right)(1 - z_3) + \frac{\rho_1}{\rho_0}\right], \end{aligned} \tag{8.14}$$

where

$$\begin{cases} z_1 = x_A + x_B + x_C, \\ z_2 = \dfrac{x_A}{x_A + x_B + \rho_1 x_C/\rho_0}, \\ z_3 = \dfrac{x_C}{x_A + x_B + x_C}. \end{cases} \tag{8.15}$$

According to Sect. 8.2, the m experimental points $\{(z_1^{(i)}, z_2^{(i)}, z_3^{(i)}), i = 1, \ldots, m\}$ which maximize the statistical power in detecting synergy should be uniformly scattered in the experimental domain $\mathscr{D} = \{(z_1, z_2, z_3) : Z_L < z_1 < Z_H, (z_2, z_3) \in \mathbf{V}_2\}$, where Z_L and Z_H are the lower and upper limits of the total dose according to drug A, respectively, and the q-dimensional simplex $\mathbf{V}_q$ is defined as

$$\mathbf{V}_q = \left\{ (w_1, \ldots, w_q) : w_j > 0, j = 1, \ldots, q; \sum_{j=1}^{q} w_j < 1 \right\} \tag{8.16}$$

The m combinations $\{(x_A^{(i)}, x_B^{(i)}, x_C^{(i)}), i = 1, \ldots, m\}$ can be obtained by the inverse transformation of (8.15).

When $\beta_C < \beta_A$, the potency $\rho(X_C)$ depends on the dose-level X_C. Then, the additive model at combination dose (x_A, x_B, x_C) is

$$\begin{aligned} y(x_A, x_B, x_C) &= \alpha_A + \beta_A \log\left[x_A + \rho_0^{\beta_A/\beta_B} \rho_1^{1-\beta_A/\beta_B} \psi^{\frac{\beta_C(\beta_B-\beta_A)}{\beta_B(\beta_C-\beta_A)}} x_B + \rho_1 \psi x_C \right] \\ &\approx \alpha_A + \beta_A \log(z_1) + \beta_A \log[(1-\rho_0)z_2 + \rho_0] \\ &\quad + \beta_A \log\left[\left(1 - \frac{\rho_1}{\rho_0}\right)(1 - z_3) + \frac{\rho_1}{\rho_0} \right], \end{aligned} \tag{8.17}$$

where ψ is a function of (x_A, x_B, x_C) and can be obtained by solving the following equation

$$\psi = \left[\frac{x_A}{\rho_1} + \left(\frac{\rho_0}{\rho_1}\right)^{\beta_A/\beta_B} \psi^{\frac{\beta_C(\beta_B-\beta_A)}{\beta_B(\beta_C-\beta_A)}} x_B + \psi x_C \right]^{1-\beta_A/\beta_C},$$

and

$$\begin{cases} z_1 = x_A + \left(\dfrac{\rho_0}{\rho_1}\right)^{\beta_A/\beta_B - 1} \psi^{\frac{\beta_C(\beta_B-\beta_A)}{\beta_B(\beta_C-\beta_A)}} x_B + \psi x_C, \\ z_2 = \dfrac{x_A}{x_A + \left(\frac{\rho_0}{\rho_1}\right)^{\beta_A/\beta_B - 1} \psi^{\frac{\beta_C(\beta_B-\beta_A)}{\beta_B(\beta_C-\beta_A)}} x_B + \frac{\rho_1}{\rho_0}\psi x_C}, \\ z_3 = \dfrac{\psi x_C}{z_1}. \end{cases} \tag{8.18}$$

Similarly, the m experimental points $\{(z_1^{(i)}, z_2^{(i)}, z_3^{(i)}), i = 1, \ldots, m\}$ which maximize the statistical power in detecting synergy should be uniformly scattered in the experimental domain $\mathscr{D} = \{(z_1, z_2, z_3) : Z_L < z_1 < Z_H, (z_2, z_3) \in \mathbf{V}_2\}$, and the m combinations $\{(x_A^{(i)}, x_B^{(i)}, x_C^{(i)}), i = 1, \ldots, m\}$ can be obtained by the inverse transformation of (8.18) (see, [38]).

If we consider the combination experiments of only two drugs A and B, the additive model at combination dose (x_A, x_B) is

$$\begin{aligned} y(x_A, x_B) &= \alpha_A + \beta_A \log\left[x_A + \phi(x_A, x_B)x_B\right] \\ &\approx \alpha_A + \beta_A \log(z_1) + \beta_A \log[(1 - \rho_0)z_2 + \rho_0], \end{aligned} \tag{8.19}$$

where ϕ is a function of (x_A, x_B) and can be obtained by solving the following equation

$$\phi(x_A, x_B) = \rho_0 \left(\phi^{-1}(x_A, x_B)x_A + x_B\right)^{(\beta_B - \beta_A)/\beta_A},$$

and

$$\begin{cases} z_1 = x_A + \dfrac{1}{2} x_B^{\beta_B/\beta_A} \left[1 + \left(1 + \dfrac{4(\beta_B - \beta_A)x_A}{\beta_A x_B^{\beta_B/\beta_A} \rho_0}\right)^{1/2}\right], \\ z_2 = \dfrac{x_A}{z_1}. \end{cases} \tag{8.20}$$

The m experimental points $\{(z_1^{(i)}, z_2^{(i)}), i = 1, \ldots, m\}$ which maximize the statistical power in detecting synergy should be uniformly scattered in the experimental domain $\mathscr{D} = \{(z_1, z_2) : Z_L < z_1 < Z_H, 0 < z_2 < 1\}$, and the m combinations $\{(x_A^{(i)}, x_B^{(i)}), i = 1, \ldots, m\}$ can be obtained by the inverse transformation of (8.20) (see [12, 37]).

8.3.2 *Design for Linear Dose-Responses*

Let the single dose response curves of drugs A, B, and C be

$$\begin{aligned} y(X_A) &= \alpha_A + \beta_A X_A \\ y(X_B) &= \alpha_B + \beta_B X_B = \alpha_A + \beta_B \left(X_B - \frac{\alpha_A - \alpha_B}{\beta_B}\right) \\ y(X_C) &= \alpha_C + \beta_C X_C = \alpha_A + \beta_C \left(X_C - \frac{\alpha_A - \alpha_C}{\beta_C}\right), \end{aligned} \tag{8.21}$$

respectively. Then, the additive model at combination dose (x_A, x_B, x_C) is

$$y(x_A, x_B, x_C) = \alpha_A + \beta_A x_A + \beta_B \left(x_B - \frac{\alpha_A - \alpha_B}{\beta_B} \right) + \beta_C \left(x_C - \frac{\alpha_A - \alpha_C}{\beta_C} \right). \tag{8.22}$$

In the combination study of drugs A, B, and C, for detecting departures from additivity of three drugs, the dose ranges of interest for combination experiments are usually considered to be a triangular prism in $\mathbf{R}^3$

$$\mathscr{D}_0 = \left\{ (x_A, x_B, x_C) : a < y(x_A, x_B, x_C) < b; x_A > 0, x_b > \frac{\alpha_A - \alpha_B}{\beta_B}, x_C > \frac{\alpha_A - \alpha_C}{\beta_C} \right\}. \tag{8.23}$$

where a and b are chosen in collaboration with pharmacologists. For example, if the doses from ED_{20} to ED_{80} based on the additive model are of interest, then $a = 20\%$ and $b = 80\%$ [10]. An algorithm for generating uniformly scattered points in $\mathscr{D}_0$ is proposed in Tian [39].

If we consider the combination experiments of only two drugs A and B, the m experimental points $\{(x_A^{(i)}, x_B^{(i)}), i = 1, \ldots, m\}$ which maximize the statistical power in detecting synergy should be uniformly scattered in a tetragon $\mathscr{D}_0 = \{(x_A, x_B) : a < \alpha_A + \beta_A x_A + \beta_B \left(x_B - \frac{\alpha_A - \alpha_B}{\beta_B} \right) < b; x_A > 0, x_b > \frac{\alpha_A - \alpha_B}{\beta_B} \}$ [13].

8.3.3 *Design for Two Linear Dose-Responses and One Log-Linear Dose-Response*

Let the single dose response curves of drugs A, B, and C be

$$\begin{aligned} y(X_A) &= \alpha_A + \beta_A X_A \\ y(X_B) &= \alpha_B + \beta_B X_B \\ y(X_C) &= \alpha_C + \beta_C \log(X_C), \end{aligned} \tag{8.24}$$

respectively. Then, the additive model at combination dose (x_A, x_B, x_C) is

$$y(x_A, x_B, x_C) = \alpha_A + \beta_A z_1 + \beta_B z_2 + z_3 \tag{8.25}$$

where

$$\begin{cases} z_1 = x_A, \\ z_2 = x_B - \dfrac{\alpha_A - \alpha_B}{\beta_B}, \\ z_3 = \left(\dfrac{\alpha_C - \alpha_A}{\zeta(x_A, x_B, x_C)} + \dfrac{\beta_C \log \zeta(x_A, x_B, x_C)}{\zeta(x_A, x_B, x_C)} \right) x_C. \end{cases} \tag{8.26}$$

and $\zeta(x_A, x_B, x_C)$ can be obtained by solving the following equation

$$\frac{\alpha_C - \alpha_A}{\beta_A} + \frac{\beta_C}{\beta_A}\log\zeta = x_A + \frac{\beta_B}{\beta_A}\left(x_B - \frac{\alpha_A - \alpha_B}{\beta_B}\right) + \left(\frac{\alpha_C - \alpha_A}{\beta_A\zeta} + \frac{\beta_C \log\zeta}{\beta_A\zeta}\right)x_C.$$

Similarly, the m experimental points $\{(z_1^{(i)}, z_2^{(i)}, z_3^{(i)}), i = 1, \ldots, m\}$ which maximize the statistical power in detecting synergy should be uniformly scattered in the triangular prism $\mathscr{D} = \{(z_1, z_2, z_3) : a < \alpha_A + \beta_A z_1 + \beta_B z_2 + z_3 < b, z_1 > 0, z_2 > 0, z_3 > 0\}$, and the m combinations $\{(x_A^{(i)}, x_B^{(i)}, x_C^{(i)}), i = 1, \ldots, m\}$ can be obtained by the inverse transformation [10]

$$\begin{cases} x_A = z_1, \\ x_B = z_2 + \dfrac{\alpha_A - \alpha_B}{\beta_B}, \\ x_C = \dfrac{z_3 \exp[(\alpha_A + \beta_A z_1 + \beta_B z_2 + z_3 - \alpha_C)/\beta_C]}{\beta_A z_1 + \beta_B z_2 + z_3}. \end{cases} \tag{8.27}$$

8.3.4 Design for One Linear Dose-Response and Two Log-Linear Dose-Responses

Let the single dose response curves of drugs A, B, and C be

$$\begin{aligned} y(X_A) &= \alpha_A + \beta_A X_A \\ y(X_B) &= \alpha_B + \beta_B \log(X_B) \\ y(X_C) &= \alpha_C + \beta_C \log(X_C), \end{aligned} \tag{8.28}$$

respectively. Then, the additive model at combination dose (x_A, x_B, x_C) is

$$y(x_A, x_B, x_C) = \alpha_A + \beta_A z_1 + z_2 + z_3 \tag{8.29}$$

where

$$\begin{cases} z_1 = x_A, \\ z_2 = \exp\left(\dfrac{\alpha_B - \alpha_C}{\beta_B}\right)\left(\dfrac{\alpha_C - \alpha_A}{\varphi^{\beta_C/\beta_B}(x_A, x_B, x_C)} + \dfrac{\beta_C \log\varphi(x_A, x_B, x_C)}{\varphi^{\beta_C/\beta_B}(x_A, x_B, x_C)}\right)x_B, \\ z_3 = \left(\dfrac{\alpha_C - \alpha_A}{\beta_A\varphi(x_A, x_B, x_C)} + \dfrac{\beta_C \log\varphi(x_A, x_B, x_C)}{\varphi(x_A, x_B, x_C)}\right)x_C. \end{cases} \tag{8.30}$$

and $\varphi(x_A, x_B, x_C)$ can be obtained by solving the following equation

$$\alpha_C - \alpha_A + \beta_C \log\varphi = \alpha_A + \beta_A x_A + \exp\left(\frac{\alpha_B - \alpha_C}{\beta_B}\right)\left(\frac{\alpha_C - \alpha_A}{\varphi^{\beta_C/\beta_B}} + \frac{\beta_C \log\varphi}{\varphi^{\beta_C/\beta_B}}\right) x_B + \left(\frac{\alpha_C - \alpha_A}{\beta_A \varphi} + \frac{\beta_C \log\varphi}{\varphi}\right) x_C.$$

Similarly, the m experimental points $\{(z_1^{(i)}, z_2^{(i)}, z_3^{(i)}), i = 1, \ldots, m\}$ which maximize the statistical power in detecting synergy should be uniformly scattered in the triangular prism $\mathscr{D} = \{(z_1, z_2, z_3) : a < \alpha_A + \beta_A z_1 + z_2 + z_3 < b, z_1 > 0, z_2 > 0, z_3 > 0\}$, and the m combinations $\{(x_A^{(i)}, x_B^{(i)}, x_C^{(i)}), i = 1, \ldots, m\}$ can be obtained by the inverse transformation [10]

$$\begin{cases} x_A = z_1, \\ x_B = \dfrac{z_2 \exp((\alpha_B - \alpha_C)/\beta_B) \exp[(\alpha_A + \beta_A z_1 + z_2 + z_3 - \alpha_C)/\beta_C]^{\beta_C/\beta_B}}{\beta_A z_1 + z_2 + z_3}, \\ x_C = \dfrac{z_3 \exp[(\alpha_A + \beta_A z_1 + z_2 + z_3 - \alpha_C)/\beta_C]}{\beta_A z_1 + z_2 + z_3}. \end{cases} \tag{8.31}$$

If we consider the combination experiments of only two drugs A and B with linear and log-Linear single dose-responses, the additive model at combination dose (x_A, x_B) is

$$y(x_A, x_B) = \alpha_A + \beta_A z_1 + \beta_B z_2, \tag{8.32}$$

where

$$\begin{cases} z_1 = x_A, \\ z_2 = \dfrac{\beta_A \varphi(x_A, x_B)}{\beta_B \exp\left(\frac{\beta_A \varphi(x_A, x_B) - \alpha_B + \alpha_A}{\beta_B}\right)} x_B, \end{cases} \tag{8.33}$$

and $\varphi(x_A, x_B)$ can be obtained by solving the following equation

$$(x_A - \varphi(x_A, x_B)) \exp\left(\frac{\beta_A \varphi(x_A, x_B) - \alpha_B + \alpha_A}{\beta_B}\right) + x_B \varphi(x_A, x_B) = 0.$$

The m experimental points $\{(z_1^{(i)}, z_2^{(i)}), i = 1, \ldots, m\}$ which maximize the statistical power in detecting synergy should be uniformly scattered in a tetragon $\mathscr{D} = \{(z_1, z_2) : a < \alpha_A + \beta_A z_1 + \beta_B z_2 < b; z_1 > 0, z_2 > 0\}$ and the m combinations $\{(x_A^{(i)}, x_B^{(i)}), i = 1, \ldots, m\}$ can be obtained by the inverse transformation [13]

$$\begin{cases} x_A = z_1, \\ x_B = \dfrac{\beta_B z_2 \exp(\alpha_A + \beta_A z_1 + \beta_B z_2 - \alpha_B)/\beta_B}{\beta_A z_1 + \beta_B z_2}. \end{cases} \tag{8.34}$$

8.4 Experimental Design for Multi-drug Combinations

In the past several decades, the identification of a variety of novel signal transduction targets amenable to therapeutic intervention has revolutionized the approach to cancer therapy. These targets were identified based on improved understanding of the molecular mechanisms of action of second messengers, other components of signal transduction pathways and system biology. These advances have suggested many potential agents and call for new quantitative approaches to explore the possibilities of combination therapy [11, 18, 23, 42]. In principle, the methods in Sect. 8.2 could be applied to multi-drug combination studies. However, the rapidly rising large number of combinations with multiple drugs is an inevitable bottleneck in the emerging approach of drug combination discovery and evaluation. Because drug-effect is dose-dependent, for example, with 8 drugs, each with 6 doses, the number of potential combinations reaches 1,679,616 multiplied by number of replications, making testing all of them prohibitive. Such complexity, further complicated by non-ignorable variation in dose-effect, behooves a new statistical approach and innovative algorithms for optimal drug combination selection, study design and analysis. Indeed challenges presented by multi-drug combinations are exceptional.

The design of multi-drug combination experiment presents exceptional challenges and a high dimensional statistical problem. Since the number of combinations grows exponentially with numbers of drugs, it quickly precludes laboratory testing. To determine the interaction among multi-drugs, the dose response surface provides a comprehensive description on dose effects. For estimating the high-dimensional dose response surface, experimental designs are required that provide selected concentrations or dose-levels of combinations, which allow exploration of the dose effect surface with high accuracy at reasonable sample sizes. Recently, we developed a novel method to screen the large number of combinations and identify the functional structure of the dose response relationship by using the dose response data of single drugs and pathway/network knowledge [11]. That is, data from experiments of single drugs and existing signaling network knowledge are utilized to develop a statistical re-scaling model to describe the effects of drugs on network topology. The system comprises a series of statistical models with biological considerations, such as Hill equations, generic enzymatic rate equations, and a regression model, to represent the cumulative effect of genes implicated in activation of the cell death machinery.

Consider a combination study of s drugs $A_1, A_2, \ldots, A_s$ inhibiting some cell line or against some cancer tumor. Assume the dose response surface to be

$$y(\mathbf{x}) = g(x_1, \cdots, x_s), \text{ for } \mathbf{x} = (x_1, \cdots, x_s)^T \in \mathscr{D} \tag{8.35}$$

where x_i is the dose-level of drug A_i, y is the dose effect scaled to be a viability (proportion of cells surviving) or a tumor volume (with some transformation), and $\mathscr{D}$ is the dose region. Without loss of generality, we assume that $\mathscr{D} = C^s = [0, 1]^s$. Using the functional ANOVA decomposition (see, [31]), the dose response $y(\mathbf{x})$ has the following unique decomposition,

$$y(x_1, \ldots, x_s) = g_0 + \sum_{i=1}^{s} g_i(x_i) + \sum_{1 \le i < j \le s} g_{ij}(x_i, x_j) + \cdots + g_{1,2,\ldots,s}(x_1, \cdots, x_s), \tag{8.36}$$

where $g_0 = \int_{C^s} y(\mathbf{x}) d\mathbf{x}$ is the overall mean of $y(\mathbf{x})$, $\int_0^1 g_{i_1,\ldots,i_u}(x_{i_1}, \ldots, x_{i_u}) dx_{i_k} = 0$ for any $1 \le u \le s$ and $1 \le k \le u$, and the orthogonality

$$\int_{[0,1]^s} g_{i_1,\ldots,i_u}(x_{i_1}, \ldots, x_{i_u}) g_{j_1,\ldots,j_v}(x_{j_1}, \ldots, x_{j_v}) dx_1 \cdots dx_s = 0, \tag{8.37}$$

if $(i_1, \ldots, i_u) \neq (j_1, \ldots, j_v)$. The total and partial variances can be defined by

$$D = \int_{C^s} [y(\mathbf{x})]^2 d\mathbf{x} - g_0, \text{ and } D_I = \int_{C^s} [g_I(\mathbf{x}_I)]^2 d\mathbf{x} \text{ for } I \subset \{1, \ldots, s\}, \tag{8.38}$$

respectively. Obviously, $D = \sum_{I \subset \{1,\ldots,s\}} D_I$. Denote the ratio $R_I = D_I / D$ which is called *global sensitivity index* [31, 32]. All R_I are non-negative and their sum $\sum_{I \subset \{1,\ldots,s\}} R_I = 1$. The variances D and D_I's and, hence, the global sensitivity indices can be approximated by the quasi Monte Carlo method [15].

The global sensitivity indices are often used to rank the importance of the $g_I(\mathbf{x}_I)$'s appearing on the right-hand side of Eq. (8.36). The larger the index R_I is, the more significant the effect of $g_I(\mathbf{x}_I)$ in the dose response is. Thus, the functional structure of $y(\mathbf{x})$ can be studied by calculating the indices. Fang et al. [11] proposed a novel procedure to identify the most significant $g_I(\mathbf{x}_I)$'s by utilizing data from experiments of single drugs and existing signaling network knowledge.

The simulation studies showed that most contributions of single drugs and drug-interactions in the dose response yielding a total of global sensitivity indices over 85%, are consistent with those from the true dose response. Then, the dose-response surface (8.35) can be reduced into

$$y(\mathbf{x}) = \mathbf{z}(\mathbf{x})^T \theta + f(\mathbf{x}), \quad \text{for } \mathbf{x} = (x_1, \ldots, x_s)^T \in \mathscr{D} \tag{8.39}$$

with

$$\int_{\mathscr{D}} f(\mathbf{x}) d\mathbf{x} = 0, \quad \text{and} \quad \int_{\mathscr{D}} \mathbf{z}(\mathbf{x}) f(\mathbf{x}) d\mathbf{x} = 0, \tag{8.40}$$

where $\mathbf{z}(\mathbf{x}) = (1, z_1(\mathbf{x}), \ldots, z_p(\mathbf{x}))^T$ is the specific functional vector of the dominating terms (e.g., those terms with their total global sensitivity indices more than 80%) and $\theta = (\theta_0, \theta_1, \ldots, \theta_p)^T$ is the corresponding vector of regression coefficients. $f(\mathbf{x})$ is an unknown function and its global sensitivity index should be less than 20%.

Since the purpose of a drug combination study is to discover the promising dose-level combinations among the agents (e.g., identify the synergistic dose region), a prediction-based design appears to be more desirable. Huang et al. [24] proposed a maximum entropy design for the combination experiments based on (8.39). Entropy is a measure of unpredictability of a random vector Z, i.e., the larger the value of

entropy of Z, the more uniform the distribution of the random vector Z which in turn implies that the more unpredictable Z is likely to be.

Let $\mathscr{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_k\}$ be the candidate set of design points in the experimental domain $\mathscr{D}$, for example, $\mathscr{X}$ is typically chosen to be a set of lattice points or uniformly measurement points over the experimental domain. The aim is to choose n points (n) from $\mathscr{X}$ as the experimental points such that the prediction variability at un-experimental points, conditionally on the experimental points, is minimized. Based on Eq. (8.39), the dose response can be formulated as

$$Y_j(\mathbf{x}_i) = \mathbf{z}(\mathbf{x}_i)^T\theta + f(\mathbf{x}_i) + \varepsilon_{ij}, i = 1, \ldots, n; j = 1, \ldots, n_i, \tag{8.41}$$

where $Y_j(\mathbf{x}_i)$ is the response value of the jth replication at the point $\mathbf{x}_i$, $\varepsilon_{ij} \sim N(0, \sigma_\varepsilon^2)$ is the measurement error and n_i is the number of replications at $\mathbf{x}_i$. The unknown function $f(\mathbf{x})$ is modeled as a Gaussian random function with zero-mean and global covariance matrix $Cov[(f(\mathbf{x}_1), \ldots, f(\mathbf{x}_k))^T] = \sigma_f^2\mathbf{V}_f^k$. In other words, $\mathbf{F}_\mathscr{X} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_k)]^T$ is regarded as a realization of $f(\mathbf{x})$.

Without loss of generality, let e be a n-run experiment selected from $\mathscr{X} = \{x_1, \ldots, x_L\}$, that is, e has n distinct support points selected from $\mathscr{X}$. Denote Y_e are the vector of response values at e and $Y_{\bar{e}}$ are the vector of response values at $\bar{e} = \mathscr{X} - e$ where $\bar{e}$ is the set of un-experimental points such that $\bar{e} \cup e = \mathscr{X}$ and $e \cap \bar{e} = \emptyset$. Let $p_\mathbf{Z}(\cdot)$ be the probability density function of the random vector $\mathbf{Z}$, the entropy of $\mathbf{Z}$ is then defined by

$$Ent(\mathbf{Z}) = \int p_\mathbf{Z}(z) log p_\mathbf{Z}(z) dz,$$

and the standard formula from information theory suggests that (cf., [26])

$$Ent(\mathbf{Y}_\mathscr{X}) = Ent(\mathbf{Y}_e) + E_{\mathbf{Y}_e}\{Ent(\mathbf{Y}_{\bar{e}}|\mathbf{Y}_e)\}, \tag{8.42}$$

where $\mathbf{Y}_\mathscr{X} = (\mathbf{Y}_e, \mathbf{Y}_{\bar{e}})$ and the expectation is with respect to the marginal distribution of $\mathbf{Y}_e$. Obviously, it is desirable for a combination experiment to minimize the second term on the right-hand side of Eq. (8.42) because this term represents the average prediction variability of the unsampled vector given the experimental design. With the assumptions of that $\theta \sim N_{p+1}(\beta_\theta, \sigma_\theta^2 I_{p+1})$ and $\mathbf{F}_\mathscr{X} \sim N_k(0, \sigma_f^2\mathbf{V}_f^k)$, $\mathbf{Y}_\mathscr{X}$ is a k-dimensional Gaussian vector and $Ent(\mathbf{Y}_\mathscr{X})$ is a constant. Therefore, minimizing the value of $E_{\mathbf{Y}_e}\{Ent(\mathbf{Y}_{\bar{e}}|\mathbf{Y}_e)\}$ is equivalent to maximizing the value of $Ent(\mathbf{Y}_e)$. The optimal design, denoted by e^*, obtained by solving the following optimization problem

$$\begin{aligned} e^* &= \arg\max_{e \subset \mathscr{X}} Ent(\mathbf{Y}_e) \\ &= \arg\max_{e \subset \mathscr{X}} \det\left[(\sigma_\theta^2/\sigma_f^2)\mathbf{Z}_e\mathbf{Z}_e^T + \mathbf{V}_f^e + (\sigma_\varepsilon^2/\sigma_f^2)\mathbf{W}_e\right] \end{aligned} \tag{8.43}$$

where $\mathbf{Z}_e = [z(\mathbf{x}_1^e), \ldots, z(\mathbf{x}_n^e)]^T$, $\mathbf{W}_e$ and $\mathbf{V}_f^e$ are the submatrices of $\mathbf{W} = diag\{n_1, \ldots, n_k\}$ and $\mathbf{V}_f^k$ respectively, determined by the experiment e. This is referred to as the *maximum entropy design* in the literature [15, 29].

The maximum entropy design criterion (8.43) is relative to the variance ratios $\sigma_\theta^2/\sigma_f^2$, $\sigma_\varepsilon^2/\sigma_f^2$, and the correlation matrix of the random function $\mathbf{V}_f^e$. As mentioned in the functional ANOVA decomposition (8.39), the total global sensitivity indices of the dominating terms is usually more than 80%, whereas the global sensitivity index of $f(\mathbf{x})$ is less than 20%. This suggests that the variance ratio $\sigma_\theta^2/\sigma_f^2 \approx (\geq)4$. The measurement error variance σ_ε^2 can be estimated by the pooled variance from the single drug experimental data (Tan et al. [37, 38]; Fang et al. [10, 12]). The idea to estimate $\mathbf{V}_f^e$ and σ_f^2 is to use the single drug dose effect curves which are estimated from the experimental data of single drugs. Let the covariance function between $f(\mathbf{x}_i)$ and $f(\mathbf{x}_j)$ be defined by $cov[f(\mathbf{x}), f(\mathbf{x}_j)] = \sigma_f^2 R(\mathbf{x}_i, \mathbf{x}_j)$, where $\mathbf{x}_i$ and $\mathbf{x}_j$ are two design points and $R(\mathbf{x}_i, \mathbf{x}_j)$ is the correlation function. The most commonly used correlation function is the power exponential correlation [9, 29]. The corresponding computational algorithm for design construction is given by Huang et al. [24].

However, the maximum entropy design is to choose the experimental points from a large candidate set of points in the experimental domain such that the posterior information on the dose-response is maximized. We have to suffer from heavy computation if the candidate set of points is too large, otherwise we may have a suboptimal design if the candidate set of points is not large enough. Another way is utilizing the combination of D- optimal and the designs. To determine the best combinations of a given multi-drugs, based on the predicting dose response model (8.41) the s-dimensional space $\mathscr{R}^s$ can be divided into two orthogonal subspaces $\mathscr{R}^s = \mathscr{H}_0 \oplus \mathscr{H}_1$, where $\mathscr{H}_0$ is the space with the basis of $z_1(\mathbf{x}), z_2(\mathbf{x}), \ldots, z_p(\mathbf{x})$ and $\mathscr{H}_1$ is the orthogonal complementary of $\mathscr{H}_0$ since $f(\mathbf{x})$ is assumed to be orthogonal with the functions $z_1(\mathbf{x}), z_2(\mathbf{x}), \ldots, z_p(\mathbf{x})$. Then, the experimental domain $\mathscr{D}$ is divided into two parts $\mathscr{D} = \mathscr{D}_1 \oplus \mathscr{D}_2$, where $\mathscr{D}_1 \subset \mathscr{H}_0$ and $\mathscr{D}_2 \subset \mathscr{H}_1$. Because the predicting dose response model has two parts, one is a linear combination of $z_1(\mathbf{x}), z_2(\mathbf{x}), \ldots, z_p(\mathbf{x})$ in $\mathscr{D}_1$ and another is an unspecified function $f(\mathbf{x})$ in $\mathscr{D}_2$, the D- or A-optimal design will be used in $\mathscr{D}_1$ and the maximal power design [38] will be used in $\mathscr{D}_2$. According to the global sensitivity indexes of the linear combination part and the unspecified function $f(\mathbf{x})$, the number of experimental points can be obtained based on the proportion of their global sensitivity indexes in the two domains and the sensitivity analysis for the variance ratio of the linear combination part and the unspecified function $f(\mathbf{x})$. The optimal design are obtained by maximizing the entropy of $Ent(\mathbf{Y}_e)$ [15, 29]. The simulation studies show that the proposed experimental design (dose-level selection and sample size determination) is efficient for combination studies and statistical procedures to fit the high-dimensional dose response surface.

8.5 Discussion and Further Research

In drug combination studies, constituent doses selected so that they are uniformly scattered in the experimental domain maximizes the minimum power of the F-test for detecting departure from additivity. The power optimality is derived from the properties of uniform measures and by minimizing the variability in modeling the dose-effect while allocating the combinations reasonably to obtain best possible estimate of the dose-response surface of the joint action. In fact, the uniform design for generating experimental combinations (the doses of each drug) using the quasi-Monte Carlo methods is an optimal fractional factorial design under a general majorization framework with exponential kernels [47, 48]. Hickernell [22] showed using quasi-Monte Carlo methods instead of the Monte Carlo method usually improves accuracy of computing the integral of a function. More importantly, the number of experimental units and replicates (sample size) in the proposed design is feasible for both in vitro and in vivo experiments.

For multi-drug combination, we proposed statistical models to describe the drug effects on the network using data from experiments with single drugs and the existing network information. Through these statistical models, we conducted computer experiments (in silico) to derive a global sensitivity index of each term in the functional ANOVA of dose response model by generating doses of the drugs with the Quasi Monte-Carlo method. Then, we can predict the main effects that occur with combinations of multiple drugs. It is highly beneficial in bringing forth a framework for selecting drug interactions, and developing experimental designs and statistical procedures to estimate the high dimensional dose-response surface.

Cancer cells perform their functions following proper responses to the extracellular and intracellular inputs to their complex network of signaling pathways. Many protein-coding genes in these pathways are controlled by regulatory proteins that up-regulate or downregulate these genes depending on the inputs to the signaling network. Though significant progress has been made in extracting networks using a range of experimental data, signaling networks remain in large part at the level of topology rather than details on the rate constants and nonlinear message passing that occur within the networks. Models to distinguish between members of a population of cells, for example, different cancer cells from different normal tissue types, require differences in message passing parameters and/or expression levels of the genes in the network. Clearly more sophisticated models to capture the network complexity are needed to accelerate the pace of drug development and increase the chance for success.

Acknowledgements This research is partially supported by the National Cancer Institute (NCI) grant R01CA164717.

References

1. Abdelbasit, K.M., Plackett, R.L.: Experimental design for joint action. Biometrics **38**, 171–179 (1982)
2. Berenbaum, M.C.: What is synergy? Pharmacol. Rev. **41**, 93–141 (1989)
3. Berenbaum, M.C., Yu, V.L., Felegie, T.P.: Synergy with double and triple antibiotic combinations compared. J. Antimicrob. Chemother. **12**, 555–563 (1983)
4. Calzolari, D., et al.: Search algorithms as a framework for the optimization of drug combinations. PLoS Comput. Biol. **4**(12), e1000249 (2008)
5. Casey, M., Gennings, C., Carter Jr., W.H., et al.: D_s-Optimal designs for studying combinations of chemicals using multiple fixed-ratio ray experiments. Environmetrics **16**, 129–147 (2005)
6. Chen, H.X., Dancey, J.E.: Combinations of moleculartargeted therapies: opportunities and challenges. In: Kaufman, H.L., Wadler, S., Antman, K. (eds.) Molecular targeting in oncology, pp. 693–705. Humana Press, New Jersey (2008)
7. Carter Jr., W.H., Gennings, C., Staniswalis, J.G., et al.: A statistical approach to the construction and analysis of isobolograms. J. Am. Coll. Toxicol. **7**, 963–973 (1988)
8. Cox, D.R., Reid, N.: The Theory of the Design of Experiments. Chapman and Hall/CRC, London (2000)
9. Cressie, N.A.C.: Statistics for Spatial Data. Wiley, New York (1993)
10. Fang, H.B., Chen, X., Pei, X. Y. et al.: Experimental design and statistical analysis for three-drug combination studies. Stat. Methods Med. Res. **26**, 1261–1280 (2017)
11. Fang, H.B., Huang, H., Clarke, R., Tan, M.: Predicting multi-drug inhibition interactions based on signaling networks and single drug dose-response information. J. Comput. Syst. Biol. **2**, 1–9 (2016)
12. Fang, H.B., Ross, D.D., Sausville, E., Tan, M.: Experimental design and interaction analysis of combination studies of drugs with log-linear dose responses. Stat. Med. **27**, 3071–3083 (2008)
13. Fang, H.B., Tian, G.L., Li, W., et al.: Design and sample size for evaluating combinations of drugs of linear and loglinear dose response curves. J. Biopharm. Stat. **19**, 625–640 (2009)
14. Fang, K.T.: Uniform Design and Uniform Design Tables. Science Press, Beijing (1994)
15. Fang, K.T., Li, R., Sudjianto, A.: Design and Modeling for Computer Experiments. Chapman and Hall/CRC, New York (2006)
16. Fang, K.T., Lin, D.K.J., Winker, P., Zhang, Y.: Uniform design: theory and application. Technometrics **42**, 237–248 (2000)
17. Finney, D.J.: Probit Analysis, 3rd edn. Cambridge University Press, London (1971)
18. Fitzgerald, J.B., Schoeberl, B., Nielsen, U.B., et al.: Systems biology and combination therapy in the quest for clinical efficacy. Nat. Chem. Biol. **2**, 458–466 (2006)
19. Gennings, C., Carter Jr., W.H., Carney, E.W., et al.: A novel flexible approach for evaluating fixed ratio mixtures of full and partial agonists. Toxicol. Sci. **80**, 134–150 (2004)
20. Greco, W.R., Bravo, G., Parsons, J.C.: The search for synergy: a critical review from a response surface perspective. Pharmacol. Rev. **47**, 331–385 (1995)
21. Hait, W.N.: Targeted cancer therapeutics. Cancer Res. **69**, 1263–1267 (2009)
22. Hickernell, F.J.: A generalized discrepancy and quadrature error bound. Math. Comput. **67**, 299–322 (1998)
23. Hopkins, A.L.: Network pharmacology: the next paradigm in drug discovery. Nat. Chem. Biol. **4**, 682–690 (2008)
24. Huang, H., Fang, H.B., Tan, M.T.: Experimental design for multi-drug combination studies using signaling networks. Biometrics **74**, 538–547 (2018)
25. Laska, E.M., Meisner, M., Siegel, C.: Simple designs and model-free tests for synergy. Biometrics **50**, 834–841 (1994)
26. Lindley, D.V.: On a measure of information provided by an experiment. Ann. Math. Stat. **27**, 986–1005 (1956)
27. Loewe, S.: Isobols of dose-effect relations in the combination of pentylenetetrazole and phenobarbital. J. Pharmacol. Exp. Ther. **114**, 185–191 (1955)

28. Meadows, S.L., Gennings, C., Carter Jr., W.H., Bae, D.S.: Experimental design for mixtures of chemicals along fixed ratio rays. Environ. Health Perspect. **110**, 979–983 (2002)
29. Santner, T.J., Williams, B.J., Notz, W.I.: The Design and Analysis of Computer Experiments. Springer, New York (2003)
30. Shiozawa, K., Nakanishi, T., Tan, M., et al.: Preclinical studies of vorinostat (suberoylanilide hydroxamic acid, saha) combined with cytosine arabinoside (ara-c) and etoposide for treatment of acute leukemias. Clin. Cancer Res. **15**, 1698–1707 (2009)
31. Sobol', I.M.: Global sensitivity indices for nonlinear mathematical models and their Monte Carlo estimates. Math. Comput. Simul. **55**, 271–280 (2001)
32. Sobol', I.M.: Theorems and examples on high dimensional model representation. Reliab. Eng. Syst. Safety **79**, 187–193 (2003)
33. Straetemans, R., O'Brien, T., Wouters, L. et al.: Design and analysis of drug combination experiments. Biom. J. **47**, 299–308 (2005)
34. Syracuse, K.C., Greco, W.R.: Comparison between the method of Chou and Talalay and a new method for the assessment of the combined effects of drugs: a Monte-Carlo simulation study. In: American Statistical Association Proceedings of the Biopharmaceutical Section, pp. 127–132 (1986)
35. Tallarida, R.J.: Drug Synergism and Dose-effect Data Analysis. Chapman and Hall/CRC, New York (2000)
36. Tallarida, R.J., Stone, D.J., Raffa, R.B.: Efficient designs for studying synergistic drug combinations. Life Sci. **61**, 417–425 (1997)
37. Tan, M., Fang, H.B., Tian, G.L., Houghton, P.J.: Experimental design and sample size determination for drug combination studies based on uniform measures. Stat. Med. **22**, 2091–2100 (2003)
38. Tan, M., Fang, H.B., Tian, G.L.: Dose and sample size determination for multi-drug combination studies. Stat. Biopharm. Res. **1**, 301–316 (2009)
39. Tian, G.L., Fang, H.B., Tan, M., et al.: Uniform distributions in a class of convex polyhedrons with applications to drug combination studies. J. Multi. Anal. **100**, 1854–1865 (2009)
40. Wan, W., Pei, X.Y., Grant, S.: Nonlinear response surface in the study of interaction analysis of three combination drugs. Biom. J. **59**, 9–24 (2017)
41. Wiens, D.P.: Designs for approximately linear regression: two optimality properties of uniform designs. Stat. Probab. Lett. **12**, 217–221 (1991)
42. Xavier, J.B., Sander, C.: Principle of system balance for drug interactions. New Engl. J. Med. **362**, 1339–1340 (2010)
43. Yang, Y., Fang, H.B., Roy, A., Tan, M.: Adaptive oncology phase I trial design of drug combinations with drug-drug interaction modeling. Stat. Interface **11**, 109–127 (2018)
44. Yin, G., Yuan, Y.: A latent contingency table approach to dose finding for combinations of two agents. Biometrics **65**, 866–875 (2009)
45. Yin, G., Yuan, Y.: Bayesian dose finding in oncology for drug combinations by copula regression. J. R. Stat. Soc. Ser. C (Appl. Stat.) **58**, 211–224 (2009)
46. Yuan, Y., Yin, G.: Sequential continual reassessment method for two-dimensional dose finding. Stat. Med. **27**, 5664–5678 (2008)
47. Zhang, A.: Schur-convex discrimination of designs using power and exponential kernels. In: Fan, J., Li, G. (eds.) Contemporary Multivariate Analysis and Experimental Design, pp. 293–311. World Scientific Publisher, Singapore (2005)
48. Zhang, A., Fang, K.T., Li, R., Sudjianto, A.: Majorization framework for balanced lattice designs. Ann. Stat. **33**, 2837–2853 (2005)

Chapter 9
Modified Robust Design Criteria for Poisson Mixed Models

Hongyan Jiang and Rongxian Yue

Abstract The maximin D-optimal design (MMD-optimal design) and hypercube design (HCD-optimal design) are two robust designs which overcome the problem of design dependence on the unknown parameters. This article considers the robust designs for Poisson mixed models. Given the prior knowledge of the fixed effects parameters, a modification of the two robust design criteria is proposed by applying the number-theoretic methods. The simulated annealing algorithm is used to find the optimal exact designs. The results show that the modified optimal designs perform better in the relative *D*-efficiency and programming time.

9.1 Introduction

In the fields of optimal experimental design, the Fisher information matrix plays an important role. For nonlinear models or generalized linear models, the Fisher information matrix depends on the unknown values of the parameters, which means that the optimal design will depend on the parameters. Researchers can fix the value based on their knowledge, or just guess, then the design will be locally optimal.

Robust design criterion is a good choice to overcome the problem of dependence of a design on the unknown parameters, such as the maximin criterion and Bayesian criterion [4]. The Bayesian approach maximizes the expected Shannon information considering the prior information about the parameters of the model, while the maximin approach optimises over a specific domain of parameter values by maximizing the minimal value of a measure of the information matrix, in which the parameters are assumed to belong to a known domain, without any hypothesis on their underlying distribution [2].

H. Jiang
Huaiyin Institute of Technology, Jiangsu 223003, China
e-mail: hyitjhy@163.com

R. Yue (✉)
College of Mathematics and Science, Shanghai Normal University, Shanghai 200234, China
e-mail: yue2@shnu.edu.cn

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_9

Aside from classical robust design criteria, the product design criterion, first suggested by Atkinson and Cox [1], maximized the product of the determinants of Fisher information matrices of the models of interest, scaled to the number of parameters in each model. McGree et al. [11] applied the product design criterion to optimise the product of the normalised determinants of Fisher information over eight different mixed effects bio-impedance models, which was combined by the 2.5th and 97.5th percentiles of all four fixed effect parameters in the model. Foo and Duffull [9] proposed a hypercube D-optimality (HCD) criterion and a hypercube maximin D-optimality (HCMMD) criterion, by setting the domain Θ_{HC} of the fixed effect parameters as various combinations of the 2.5th and 97.5th percentiles from the known prior distribution of them in nonlinear mixed models. The HCD method is a particular case of the product design criterion, and the result shows that this method performs better at some combination of the extrema values of the parameters. What's more, a 100-fold improvement in the speed of this method compared to the Bayesian optimal design is particularly attractive.

However, the percentiles of the prior distribution do not scatter as 'uniform' as possible, and the underlying assumption of the HCD and HCMMD is that the efficiency of any locally D-optimal design of the 97.5% percentiles is more or as efficient to design of the parameter values located within the 97.5% interval [9]. We want to generate a set of the parameter values which are uniformly scattered in a given multi-dimensional prior distribution. Number-theoretic methods (NTMs) are used in experimental design by Fang and Wang [5]. The set of the representative points (RPs) based on NTMs is uniformly scattered under the notation of discrepancy. The aim of this paper is to provide a robust method of obtaining optimal designs based on the RPs. In what follows, given the prior distribution of the fixed effect parameters, a D-optimality criterion based on the set of RPs, denoted by RPD-optimality criterion, and a maximin optimality criterion based on the set of RPs, denoted by RPMMD, are proposed.

The rest of the paper is organized as follows. The Poisson mixed models are introduced in Sect. 9.2. Section 9.3 gives a brief review on the existing criteria, and presents a modification of the robust criteria by using the NTMs. Section 9.4 evaluates the new robust criteria via an one-variable first-order and second-order Poisson mixed models by comparing among several designs. Section 9.5 is the conclusion of the paper.

9.2 The Poisson Mixed Model

In this section, a Poisson mixed model is introduced, and the quasi-likelihood method [12, 13, 16] is applied to Poisson mixed model to obtain the quasi-information matrix.

9.2.1 Poisson Mixed Models

Suppose there are N independent individuals taken part in an experiment, and the responses y_{ij} at the experimental settings x_{ij} of an explanatory variable x for individual i follows a Poisson distribution, conditioned on an r-dimensional random effects vector b_i [12, 13]. It is assumed that y_{ij}'s are related to the fixed and random effects via a log link, that is $\log(\lambda_{ij}) = f_{ij}^T \beta + z_{ij}^T b_i$, and

$$p(y_{ij}|b_i) = \frac{\lambda_{ij}^{y_{ij}}}{y_{ij}!} \exp(-\lambda_{ij}), \quad y_{ij} = 0, 1, 2, \cdots, \quad i = 1, 2, ..., N, \quad j = 1, 2, \cdots, m_i, \tag{9.1}$$

where $p(y_{ij}|b_i)$ denotes the conditional probability density function of y_{ij} given b_i. Moreover, given the individual random effects b_i, the observations y_{ij} are assumed to be conditionally independent. The $p \times 1$ vector f_{ij} is the design vector of the explanatory variable at the jth measurement for individual i, β is the corresponding $p \times 1$ vector of unknown fixed effect parameters, z_{ij} is the $r \times 1$ ($r \leq p$) design vector for the random effects which is usually a subset of vector f_{ij}, and b_i, $i = 1, 2, \ldots, N$ is the corresponding $r \times 1$ vector of unknown random effects which are drawn independently from a multivariate normal distribution with mean zero and covariance matrix G.

Let the vector $y_i = (y_{i1}, \cdots, y_{im_i})^T$ be the count responses of individual i, and $y = (y_1^T, y_2^T, \cdots, y_N^T)^T$ be the response vector of the experiment for the N individuals.

9.2.2 Fisher Information Matrix of the Model

Our interest lies in measuring the responses under a reasonable experimental design to estimate the fixed effect parameter β as accurately as possible. For simplicity we will assume throughout that the covariance matrix G of b_i is known. Note that the covariance matrix G need not be of full rank, which allows for some or most of the parameters to be fixed across the individuals.

The likelihood function of β is

$$L(\beta) = \prod_{i=1}^{N} \int \prod_{j=1}^{m_i} p(y_{ij}|b_i) p(b_i) db_i, \tag{9.2}$$

where $p(b_i)$ is the probability density function of b_i. The maximum likelihood estimator of β cannot be written down in closed form due to the random effects in model (9.1). As mentioned in [12–14, 16], quasi-likelihood method is employed to construct the quasi-likelihood function $QL(\beta; y)$. See [8, 10, 16] for details. The quasi-information matrix for the experiment is

$$M(\beta) = D^T V^{-1}(\mu(\beta))D = \sum_{i=1}^{N} M_i(\beta), \tag{9.3}$$

where $\mu(\beta)$ is the marginal mean of y, $V(\mu(\beta))$ is the marginal covariance matrix of y, $D = \partial\mu(\beta)/\partial\beta$, and $M_i(\beta)$ is the quasi-information matrix of individual i.

According to the technique of variance correction in [14], we define a variance correction term

$$c(z_{ij}, z_{ij'}) = \exp(z_{ij}^T G z_{ij'}) - 1,$$

and let $C_i = (c(z_{ij}, z_{ij'}))$ be the $m_i \times m_i$ matrix of the correction terms. Then the quasi-information matrix of individual i is given by

$$M_i(\beta) = F_i^T A_i^T (A_i + A_i C_i A_i)^{-1} A_i F_i = F_i^T (A_i^{-1} + C_i)^{-1} F_i, \tag{9.4}$$

where A_i is a diagnose matrix with the individual mean vector $E(Y_i)$ on its diagonal. Note that

$$D_i = \frac{\partial \mu_i(\beta)}{\partial \beta} = A_i F_i,$$

where $F_i^T = (f_{i1}, \cdots, f_{im_i})$ is the design matrix of individual i.

In what follows we mainly consider the one-variable first-order Poisson mixed model

$$\lambda_{ij} = \exp(\beta_0 + b_{i0} + (\beta_1 + b_{i1})x_j), \tag{9.5}$$

and the one-variable second-order Poisson mixed model

$$\lambda_{ij} = \exp(\beta_0 + b_{i0} + (\beta_1 + b_{i1})x_j + (\beta_2 + b_{i2})x_j^2). \tag{9.6}$$

In these models the design vectors for the fixed effects and the random effects are equal, i.e., $f_{ij} = z_{ij}$ in model (9.1).

9.3 Robust Optimal Designs

9.3.1 Locally D-Optimal Designs

In most practical situations, exact design with a given total number of design points is required. The objective of this paper is to determine an optimal m-exact design of the following form

$$\xi_m = \begin{Bmatrix} x_1 & x_2 & \cdots & x_s \\ n_1 & n_2 & \cdots & n_s \end{Bmatrix},$$

where x_k, $k = 1, 2, \cdots, s$, are the s different settings for each individual, and n_k denotes the corresponding repetition times of observations at $x_k, k = 1, 2, \cdots, s$, and $\sum_{k=1}^{s} n_k = m$. The individual design with fixed m is considered, which is reasonable in practice. Each exact design can be considered as a design measure over the design region, which can be written as a probability measure with supports x_k's:

$$\xi = \begin{Bmatrix} x_1 & x_2 & \cdots & x_s \\ p_1 & p_2 & \cdots & p_s \end{Bmatrix}, \quad p_k = \frac{n_k}{m}, \quad \sum_{k=1}^{s} p_k = 1.$$

A design ξ that makes the estimation of the unknown parameters in a model, β, as effectively as possible, dominates over all other designs in the set of all design measures Ξ in the Löwner sense is called Löwner optimal. However, it is very difficult to find the Löwner optimal design ξ, in general. A popular way is to specify an optimality criterion, which is defined as a real-valued function of the information matrix $M(\xi; \beta)$ of the model. The most commonly used function is logarithm of its determinant $\log |M(\xi; \beta)|$ and the corresponding optimality is known as D-optimality. A design ξ is called a locally D-optimal in the Poisson mixed model (9.1) if for a given nominal value of β, it maximizes $\log |M(\xi; \beta)|$, i.e.,

$$\xi^D = \arg\max_{\xi} \log |\mathbf{M}(\beta)|. \tag{9.7}$$

It is known that a D-optimal design ξ^D minimizes the content of the confidence region of β and so minimizes the volume of the ellipsoid [2]. Note that the information matrix $M(\xi; \beta)$ for a general model usually depends on the parameters β, and then the design ξ^D is called locally D-optimal. In Sect. 9.4, the locally D-optimal designs for the Poisson mixed models in (9.5) and (9.6) are calculated at the prior means of β, respectively.

Niaparast and Schwabe [14] provides an equivalence theorem for checking the optimality for a given candidate design for the Poisson mixed models. The D-efficiency of an arbitrary design ξ compared to the D-optimal design ξ^D is defined as [2]

$$D_{\text{eff}} = \left(\frac{|M(\xi; \beta)|}{|M(\xi^D; \beta)|} \right)^{\frac{1}{p}}, \tag{9.8}$$

where, p is the number of parameters for the fixed effects of the model.

A Bayesian D-optimal design, ξ^{BD}, helps to overcome the problem of design dependence on the unknown parameters, is defined as follows:

$$\xi^{BD} = \arg\max_{\xi} \int_{\beta} \log |M(\xi; \beta)| \eta(\beta) d\beta, \tag{9.9}$$

where $\eta(\beta)$ is a chosen prior distribution of β. The integration here will be calculated numerically by quasi-Monte Carlo (QMC) methods. It is known that the QMC

methods for multi-dimensional numerical integration are much more efficient than traditional Monte Carlo methods [5].

9.3.2 RPD-and RPMMD-Optimalities

Foo and Duffull [9] proposed a hypercube design criterion termed HCD-optimality, which is a specific case of product optimality, where component models are formed by the same structure model but with sets of parameter values taken at the 2.5th and 97.5th percentiles values of the prior distribution of β. A maximin design criterion was also considered in [9] by setting the domain of parameters as Θ_{HC} composed of all the combinations of the 2.5th and 97.5th percentile values, which is called HCMMD-optimality. The HCD-optimal design is defined by

$$\xi^{HCD} = \arg\max_{\xi} \sum_{\beta \in \Theta_{HC}} \log |M(\xi; \beta)|, \tag{9.10}$$

and the HCMMD-optimal design is defined by

$$\xi^{HCMMD} = \arg\max_{\xi} \min_{\beta \in \Theta_{HC}} \log |M(\xi; \beta)|. \tag{9.11}$$

The method in [9] is attractive for its short operating time and acceptable effective at some nominal parameter values. The maximin optimal designs [4, 7, 17] are particularly attractive since an appropriate range for the unknown parameters is only required to specify. The major problem is that the maximin optimality criterion is not differentiable and the equivalence theorem is elusive.

Note that the set of percentiles may not represent as much information of a Multivariate distribution as possible. We now consider the use of RPs of the prior distribution by NTMs. Fang and Wang [5] introduced two kinds of RPs based on the F-discrepancy criterion and MSE criterion, respectively. Under the F-discrepancy criterion, there exists a set of optimal RPs for a given continuous univariate distribution by directly using the inverse transformation method. For the multivariate distributions with independent components, their RPs may also be obtained by using the inverse transformation method. For the multivariate distributions with dependence structures, Fang and Wang [5] proposed the NTSR algorithm to generate their RPs, which can be implemented to obtain the RPs of the spherically symmetric distribution, multivariate l_1-norm distribution, Liouville distribution, and so on. Zhou and Wang [18] considered the RPs of Student's t_n distribution for minimizing the MSE criterion. Very recently, Zhou and Fang [19] proposed a new criterion, termed FM-criterion, to choose n RPs of a given distribution, which minimize the L_2-norm of the difference between the empirical distribution and the given distribution under the constraint that the first $n-1$ sample moments equal the population moments. The empirical study in [19] shows that the RPs under the FM-criterion are better than

other types of RPs. It is known that finding RPs under the MSE criterion is more difficult, but more appropriate in the case of small sample size.

In what follows, by $\Theta_{n\text{-}RP}$ we denote a set of n RPs generated by the inverse transformation method under F-discrepancy criterion from a prior distribution of β with independent components. We define two robust design criteria to against the uncertainty of the fixed effects in the mixed model (9.1) by using the RPs in $\Theta_{n\text{-}RP}$, and compare them with the existing criteria in (9.10) and (9.11).

A design is called RPD-optimal if it maximizes

$$\Phi_{RPD}(\xi) = \sum_{\beta \in \Theta_{n\text{-}RP}} \log |M(\xi; \beta)|, \tag{9.12}$$

and a design is called RPMMD-optimal if it maximizes

$$\Phi_{RPMMD}(\xi) = \min_{\beta \in \Theta_{n\text{-}RP}} \log |M(\xi; \beta)|. \tag{9.13}$$

9.4 Numerical Studies

In this section we present Numerical studies for the RPD-and RPMMD-optimal designs for the first-order model in (9.5) with three different covariance structures of the random effects, and the second-order model in (9.6) with a diagonal covariance matrix of the random effects, respectively. The design region is taken as $[c, 1]$ with $c = 0.01, 0.2, 0.4$, respectively, as used in [15].

We assume that β has a continuous multivariate prior distribution $H(\beta)$ with independent components. i.e., $H(\beta) = H(\beta_1, \cdots, \beta_p) = \prod_{i=1}^{p} H_i(\beta_i)$, where $H_i(\beta_i)$ $(i = 1, \cdots, p)$ are the marginnal distribution functions of β. We use the NTMs as demonstrated in [5] to find the set of RPs of the prior distribution. Letting $\{c_k = (c_{k1}, \cdots, c_{kp}), k = 1, \cdots, n\}$ is a set of n points which are uniformly scattered in the unit cube $C^s = [0, 1]^s$, e.g., a good lattice points (glp) set, then the set $\Theta_{n\text{-}RP}$ is obtained by using the inverse transformation method, i.e., $\Theta_{n\text{-}RP} = \{\beta_k = (H_1^{-1}(c_{k1}), \cdots, H_p^{-1}(c_{kp})), k = 1, \cdots, n\}$.

To find the optimal m-exact designs that maximize the criteria defined in last section, we use the simulated annealing (SA) algorithm. In our computation for $m = 8, 12, 24$, the initial temperature in the SA algorithm is taken as $T_0 = 10^6$, and the temperature reduction factor is 0.9. It is known that the SA algorithm allows the search patterns to move away from a path of strict descent, migrates through a sequence of local extremum in search of the global solution, and recognizes when the global extremum has been located [3, 6, 9].

It must be noted that the Bayesian D-optimality criterion (9.9) requires a complicated integration over the prior distribution. The computation of the Bayesian optimal designs involves two steps: (i) computation of criterion for a given design, and (ii) finding an optimal design by maximization of the criterion value. To compute

the criterion (9.9) for a given design, we use the NTMs which is more efficient than Monte Carlo methods to obtain a good approximation of integration.

9.4.1 Designs for the First-Order Poisson Mixed Model

For the first-order Poisson mixed model given in (9.5), we consider the following three kinds of covariance matrices G of random effects $b = (b_0, b_1)^T$:

$$G_1 = \begin{pmatrix} 0.5 & 0 \\ 0 & 0 \end{pmatrix}, \quad G_2 = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}, \quad G_3 = \begin{pmatrix} 0.5 & 0.25 \\ 0.25 & 0.5 \end{pmatrix}.$$

9.4.1.1 The Case of Normal Prior Distributions

Assume that the prior distribution of $\beta = (\beta_0, \beta_1)^T$ is a normal distribution with mean $\bar{\beta} = (\bar{\beta}_0, \bar{\beta}_1)^T = (1, -3)^T$ and an identity covariance matrix I_2.

In order to compare with the design criterion using percentile points in Θ_{HC} which contains 4 values of β, we will use the set $\Theta_{3\text{-}RP}$ of the RPs of the prior distribution $\beta \sim N_2(\bar{\beta}, I_2)$. According to Theorem 1.2 in Fang and Wang [5], the set $\Theta_{3\text{-}RP}$ can be obtained by taking an inverse transformation of the following glp set in C^2,

$$\left\{\left(\frac{1}{6}, \frac{3}{6}\right), \left(\frac{3}{6}, \frac{1}{6}\right), \left(\frac{5}{6}, \frac{5}{6}\right)\right\}.$$

The two sets $\Theta_{3\text{-}RP}$ and Θ_{HC} chosen from the prior distribution $N_2(\bar{\beta}, I_2)$ of β are shown in Table 9.1.

The optimal m-exact designs ($m = 8,16,24$) under the five optimality criteria in (9.9)–(9.13) for the first-order model (9.5) with random effects covariance matrix G_i ($i = 1, 2, 3$) are calculated numerically, where the sets $\Theta_{n\text{-}RP}$ and Θ_{HC} used in these criteria are given in Table 9.1. To save space, we only show the optimal 8-exact designs for the covariance matrix G_2 in Table 9.2.

It is observed from this table that for a given value of c, the designs have two support points except for the HCD-optimal designs on the cases $c = 0.01, 0.2$. The left endpoint of each design region is the common support of these designs, but the weights on it can be different.

In the following, taking for example, we make an efficiency comparison among the optimal 8-exact designs in the case of $c = 0.01$. We compute the D-efficiencies defined by (9.8) of the optimal 8-exact designs on the region $[0.01, 1]$ obtained under the criteria (9.9)–(9.13), respectively, with respect to each of the 100 locally D-optimal designs where the 100 values of β are randomly sampled from its prior distribution $N_2(\bar{\beta}, I_2)$. The results for each model with random effects covariance matrix G_j ($j = 1, 2, 3$) are shown in Figs. 9.1–9.3. In each plot, column 1 stands for the box plot of D-efficiency of the RPD-optimal design, column 2 for the box plot of

Table 9.1 The sets $\Theta_{3\text{-}RP}$ and Θ_{HC} for the prior distribution $\beta \sim N_2(\bar{\beta}, I_2)$

$\Theta_{3\text{-}RP}$	β_0	β_1	Θ_{HC}	β_0	β_1
β_{RP}^1	0.0326	−3.0000	β_{HC}^1	$1-1.96$	$-3-1.96$
β_{RP}^2	1.0000	−3.9674	β_{HC}^2	$1-1.96$	$-3+1.96$
β_{RP}^3	1.9674	−2.0326	β_{HC}^3	$1+1.96$	$-3-1.96$
			β_{HC}^4	$1+1.96$	$-3+1.96$

Table 9.2 The optimal 8-exact designs on $[c, 1]$ for the first-order model (9.5) with random effects covariance matrix $G_2 = 0.5\, I_2$ based on the sets $\Theta_{3\text{-}RP}$ and Θ_{HC} in Table 14.1

Criterion	$c = 0.01$	$c = 0.2$	$c = 0.4$
Local D	$\begin{pmatrix} 0.01 & 0.7250 \\ 0.3125 & 0.6875 \end{pmatrix}$	$\begin{pmatrix} 0.2 & 0.8802 \\ 0.375 & 0.625 \end{pmatrix}$	$\begin{pmatrix} 0.4 & 1 \\ 0.375 & 0.625 \end{pmatrix}$
HCD	$\begin{pmatrix} 0.01 & 0.5567 & 1 \\ 0.375 & 0.5 & 0.125 \end{pmatrix}$	$\begin{pmatrix} 0.2 & 0.7611 & 1 \\ 0.375 & 0.5 & 0.125 \end{pmatrix}$	$\begin{pmatrix} 0.4 & 0.9886 \\ 0.5 & 0.5 \end{pmatrix}$
HCMMD	$\begin{pmatrix} 0.01 & 0.4688 \\ 0.375 & 0.5 \end{pmatrix}$	$\begin{pmatrix} 0.2 & 0.656 \\ 0.375 & 0.625 \end{pmatrix}$	$\begin{pmatrix} 0.4 & 0.8524 \\ 0.5 & 0.5 \end{pmatrix}$
RPD	$\begin{pmatrix} 0.01 & 0.7075 \\ 0.375 & 0.625 \end{pmatrix}$	$\begin{pmatrix} 0.2 & 0.9081 \\ 0.375 & 0.625 \end{pmatrix}$	$\begin{pmatrix} 0.4 & 1 \\ 0.375 & 0.625 \end{pmatrix}$
RPMMD	$\begin{pmatrix} 0.01 & 0.8199 \\ 0.375 & 0.625 \end{pmatrix}$	$\begin{pmatrix} 0.2 & 1 \\ 0.375 & 0.625 \end{pmatrix}$	$\begin{pmatrix} 0.4 & 1 \\ 0.375 & 0.625 \end{pmatrix}$
Bayesian	$\begin{pmatrix} 0.01 & 0.7198 \\ 0.375 & 0.625 \end{pmatrix}$	$\begin{pmatrix} 0.2 & 0.9185 \\ 0.375 & 0.625 \end{pmatrix}$	$\begin{pmatrix} 0.4 & 1 \\ 0.5 & 0.5 \end{pmatrix}$

D-efficiency of the HCD-optimal design, column 3 for the box plot of D-efficiency of the Bayesian D-optimal design, column 4 for the box plot of D-efficiency of the RPMMD-optimal design, and column 5 stands for the box plot of D-efficiency of HCMMD-optimal design.

Figure 9.1 shows the results for the first-order Poisson model with random intercept. The median of the D-efficiency of the RPMMD-optimal design is the highest, even better than Bayesian optimal design, and the performance of the HCMMD-optimal design is the worst. Although the D-efficiency of the RPD-optimal design is a little lower than that of the HCD-optimal design, its median is above 0.8, which is acceptable in practice. Figures 9.2–9.3 show the results for the first-order Poisson model with both random intercept and random slope. These results show that the difference of the five designs shrinks, and their performances are comparable. It is noticed that the RPD-, RPMMD-and Bayesian D-optimal designs perform better than the HCD-and HCMMD-optimal designs. In conclusion, the optimality criteria based on the RPs is more efficient than that based on the hypercube method to overcome the problem of dependence of designs on the unknown parameters of the model.

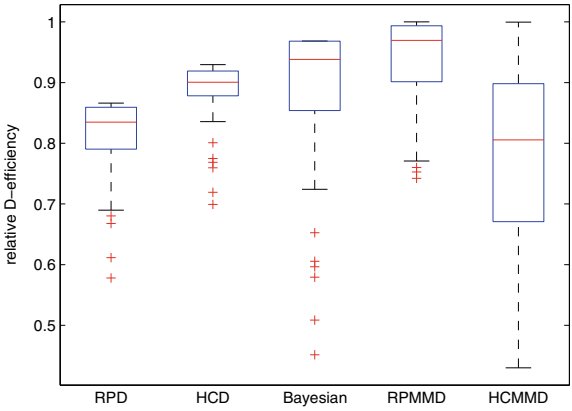


Fig. 9.1 Box plots of the D-efficiencies of the five optimal 8-exact designs with respect to the 100 locally D-optimal 8-exact designs on [0.01, 1] for the first-order model (9.5) with random effects covariance matrix G_1

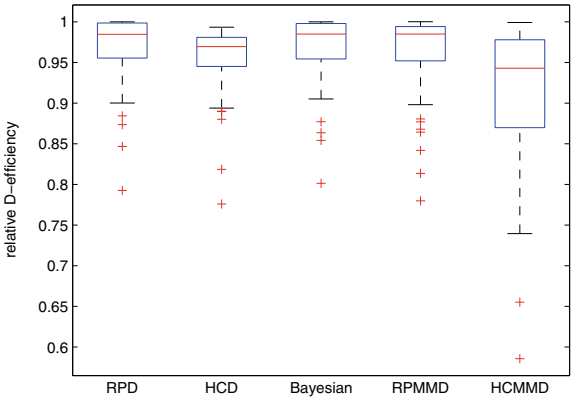


Fig. 9.2 Box plots of the D-efficiencies of the 8-exact optimal 8-exact designs with respect to the 100 locally D-optimal 8-exact designs on [0.01, 1] for the first-order model (9.5) with random effects covariance matrix G_2

Furthermore, we examine the affect of the number of RPs on the RPD- and RPMMD-optimal exact designs on [0.01, 1] for the first-order model (9.5) with random effects covariance matrix G_j ($j = 1, 2, 3$). The RPD-and RPMMD-optimal 8-exact designs are carried out under three sets $\Theta_{n\text{-}RP}$ ($n = 3, 5, 8$), and the D-efficiencies of these designs are calculated with respect to the locally D-optimal 8-exact designs at each of 100 values of β which are randomly sampled from the prior distribution $N_2(\bar{\beta}, I_2)$. For space reason, in Fig. 9.4 we only report a part of these D-efficiencies of the RPD-and RPMMD-optimal designs for the first-order model (9.5) with random effects covariance matrix G_1. These results show that the number of RPs has a slight impact on the RPD-and RPMMD-optimal designs.

9.4.1.2 The Case of Noncentral *t* Prior Distributions

We consider the case of non-normal prior distributions of the fixed effects. For illustration purpose, we assume that β_0 and β_1 in the first-order model (9.5) are indepen-

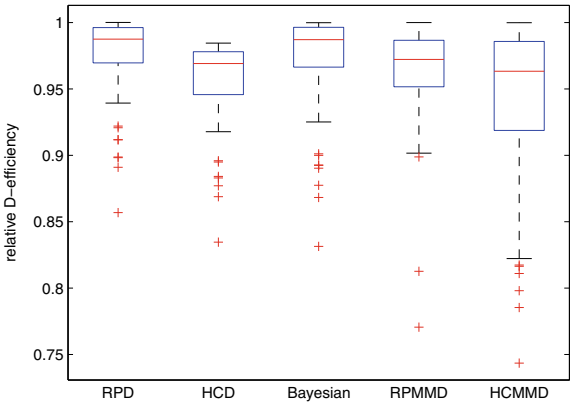


Fig. 9.3 Box plots of the D-efficiencies of the five optimal 8-exact designs with respect to the 100 locally D-optimal 8-exact designs on [0.01, 1] for the first-order model (9.5) with random effects covariance matrix G_3

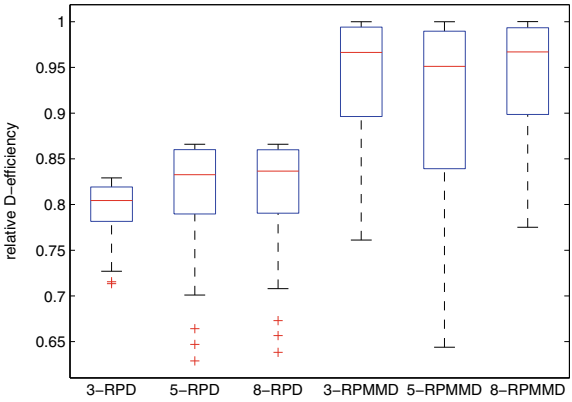


Fig. 9.4 Box plots of the D-efficiencies of the RPD-and RPMMD-optimal 8-exact designs under three sets $\Theta_{n\text{-}RP}$ ($n = 3, 5, 8$) with respect to the 100 locally D-optimal 8-exact designs on [0.01, 1] for the first-order model (9.5) with random effects covariance matrix G_1

dent and follow noncentral t distributions having means 1 and -3, respectively. Let $\beta_0 \sim t(q_0, \delta_0)$ and $\beta_1 \sim t(q_1, \delta_1)$. By assuming the degrees of freedom $q_0 = 4$ and $q_1 = 3$, the noncentrality parameters are then obtained by solving the equations

$$E(\beta_i) = \frac{\delta_i \Gamma(\frac{q_i - 1}{2})}{\Gamma(\frac{q_i}{2})} \sqrt{\frac{q_i}{2}}, \quad i = 0, 1,$$

which are $\delta_0 = 0.7979$ and $\delta_1 = -2.1708$, respectively. The set of RPs can also be obtained by the NTMs. Our computation is carried out in Matlab, and the sets $\Theta_{3\text{-}RP}$ and Θ_{HC} of β with this prior distribution are given in Table 9.3. The results in Table 9.4 are the optimal 8-exact designs under the five optimality criteria in (9.10)–(9.13) for the first-order model (9.5) with random effects covariance matrix $G_2 = 0.5\, I_2$ and the noncetral t prior distribution of β, where the sets $\Theta_{n\text{-}RP}$ and Θ_{HC} used in these criteria are as in Table 9.3. Compared with the results in Table 9.2, we observed that both the RPD-and RPMMD-optimal designs on the region $[c, 1]$ are very similar (except

Table 9.3 The sets $\Theta_{3\text{-}RP}$ and Θ_{HC} of $\beta = (\beta_0, \beta_1)^T$ whose components are independent and follow prior distributions $t(4, 0.7979)$ and $t(3, -2.1708)$ respectively

Θ_{RP}	β_0	β_1	Θ_{HC}	β_0	β_1
β_{RP}^1	−0.1823	−2.3957	β_{HC}^1	−1.4604	−9.4003
β_{RP}^2	0.8505	−4.4759	β_{HC}^2	−1.4604	−0.2209
β_{RP}^3	2.1199	−1.1984	β_{HC}^3	4.3557	−9.4003
			β_{HC}^4	4.3557	−0.2209

Table 9.4 The optimal 8-exact designs on $[c, 1]$ for the first-order model (9.5) with random effects covariance matrix $G_2 = 0.5\, I_2$ and the noncentral t prior distribution of β, based on the sets $\Theta_{3\text{-}RP}$ and Θ_{HC} in Table 9.3

criterion	$c = 0.01$	$c = 0.2$	$c = 0.4$
HCD	$\begin{pmatrix} 0.01 & 0.2456 & 1 \\ 0.25 & 0.5 & 0.25 \end{pmatrix}$	$\begin{pmatrix} 0.2 & 0.4402 & 1 \\ 0.375 & 0.5 & 0.125 \end{pmatrix}$	$\begin{pmatrix} 0.4 & 0.6527 & 1 \\ 0.375 & 0.5 & 0.125 \end{pmatrix}$
HCMMD	$\begin{pmatrix} 0.01 & 0.24 \\ 0.375 & 0.625 \end{pmatrix}$	$\begin{pmatrix} 0.2 & 0.4209 \\ 0.5 & 0.5 \end{pmatrix}$	$\begin{pmatrix} 0.4 & 0.6208 \\ 0.5 & 0.5 \end{pmatrix}$
RPD	$\begin{pmatrix} 0.01 & 0.6979 \\ 0.375 & 0.625 \end{pmatrix}$	$\begin{pmatrix} 0.2 & 0.8862 \\ 0.375 & 0.625 \end{pmatrix}$	$\begin{pmatrix} 0.4 & 1 \\ 0.375 & 0.625 \end{pmatrix}$
RPMMD	$\begin{pmatrix} 0.01 & 0.7204 \\ 0.375 & 0.625 \end{pmatrix}$	$\begin{pmatrix} 0.2 & 0.7420 \\ 0.375 & 0.625 \end{pmatrix}$	$\begin{pmatrix} 0.4 & 0.9498 \\ 0.375 & 0.625 \end{pmatrix}$
Bayesian	$\begin{pmatrix} 0.01 & 0.7110 & 1 \\ 0.375 & 0.25 & 0.125 \end{pmatrix}$	$\begin{pmatrix} 0.2 & 0.8658 & 1 \\ 0.5 & 0.25 & 0.25 \end{pmatrix}$	$\begin{pmatrix} 0.4 & 0.9648 & 1 \\ 0.5 & 0.125 & 0.375 \end{pmatrix}$

RPMMD at $c = 0.2$), while others are much different, based on the two kinds of the prior distribution of β.

9.4.2 Designs for the Second-Order Poisson Mixed Model

A similar discussion to the previous subsection is done to the second-order Poisson mixed model in (9.6). For illustration, we assume that the covariance matrix of the random effects $b = (b_0, b_1, b_2)^T$ is $G = 0.5\, I_3$, and the prior distribution of the fixed

effects $\beta = (\beta_0, \beta_1, \beta_2)^T$ is normal distribution with mean $\bar{\beta} = (1, -3, -0.9)^T$ and covariance matrix I_3. In this case, the percentile set Θ_{HC} contains 8 points, and for comparison we choose the set $\Theta_{7\text{-}RP}$ having seven RPs of the prior distribution $\beta \sim N_3(\bar{\beta}, I_3)$, which is obtained by the inverse transformation method from the following glp set,

$$\left\{\left(\frac{1}{14}, \frac{5}{14}, \frac{9}{14}\right), \left(\frac{3}{14}, \frac{11}{14}, \frac{5}{14}\right), \left(\frac{5}{14}, \frac{3}{14}, \frac{1}{14}\right), \left(\frac{7}{14}, \frac{9}{14}, \frac{11}{14}\right), \left(\frac{9}{14}, \frac{1}{14}, \frac{7}{14}\right), \left(\frac{11}{14}, \frac{7}{14}, \frac{3}{14}\right), \left(\frac{13}{14}, \frac{13}{14}, \frac{13}{14}\right)\right\}.$$

The two sets $\Theta_{7\text{-}RP}$ and Θ_{HC} are shown in Table 9.5.

Table 9.6 shows the six kinds of optimal 8-exact designs on the region $[c, 1]$ with $c = 0.01, 0.2$ for the second-order model (9.6) with the random effects covariance matrix $Cov(b) = 0.5\, I_3$. These designs are obtained numerically under the six optimality criteria given in (9.7), (9.9)–(9.13), where the sets $\Theta_{7\text{-}RP}$ and Θ_{HC} in Table 9.5 are used in (9.10)–(9.13) correspondingly.

As in the previous subsection, we are going to make a comparison among these designs. We generate randomly 100 values of β from the prior distribution $\beta \sim N_3(\bar{\beta}, I_3)$, and find out the locally D-optimal 8-exact designs on the region $[0.01, 1]$ at each of these values of β. Then we calculate the D-efficiencies of the RPD-, HCD-, Bayesian D-, RPMMD-and HCMMD-optimal 8-exact designs relative to each of these locally D-optimal designs. The box plots of these D-efficiencies are shown in Fig. 9.5.

As shown in Fig. 9.5, the medians of D-efficiencies of the RPD-, HCD-, Bayesian D-, RPMMD-optimal designs are all greater than 0.95, while the median of D-efficiencies of the HCMMD-optimal design is 0.8. The performance of the RPD-optimal design is slightly better than the HCD-and Bayesian D-optimal designs. And the performance of the RPMMD-optimal design is much better than the HCMMD-optimal design.

Table 9.5 The sets $\Theta_{7\text{-}RP}$ and Θ_{HC} for the prior distribution $\beta \sim N_3(\bar{\beta}, I_3)$

$\Theta_{7\text{-}RP}$	β_0	β_1	β_2	Θ_{HC}	β_0	β_1	β_2
β_{RP}^1	−0.4652	−3.3661	−0.5339	β_{HC}^1	$1-1.96$	$-3-1.96$	$-0.9-1.96$
β_{RP}^2	0.2084	−2.2084	−1.2661	β_{HC}^2	$1-1.96$	$-3+1.96$	$-0.9-1.96$
β_{RP}^3	0.6339	−3.7916	−2.3652	β_{HC}^3	$1-1.96$	$-3-1.96$	$-0.9+1.96$
β_{RP}^4	1.0000	−2.6339	−0.1084	β_{HC}^4	$1-1.96$	$-3+1.96$	$-0.9+1.96$
β_{RP}^5	1.3661	−4.4652	−0.9	β_{HC}^5	$1+1.96$	$-3-1.96$	$-0.9-1.96$
β_{RP}^6	1.7916	−3	−1.6916	β_{HC}^6	$1+1.96$	$-3+1.96$	$-0.9-1.96$
β_{RP}^7	2.4652	−1.5348	0.5652	β_{HC}^7	$1+1.96$	$-3-1.96$	$-0.9+1.96$
				β_{HC}^8	$1+1.96$	$-3+1.96$	$-0.9+1.96$

Table 9.6 The optimal 8-exact designs on $[c, 1]$ for the second-order model (9.6) with the random effects covariance matrix $Cov(b) = 0.5\, I_3$

criterion	$c = 0.01$	$c = 0.2$
Local D	$\begin{pmatrix} 0.01 & 0.3365 & 0.9665 \\ 0.25 & 0.375 & 0.375 \end{pmatrix}$	$\begin{pmatrix} 0.2 & 0.4923 & 1 \\ 0.25 & 0.375 & 0.375 \end{pmatrix}$
HCD	$\begin{pmatrix} 0.01 & 0.2807 & 0.7627 & 1 \\ 0.25 & 0.375 & 0.25 & 0.125 \end{pmatrix}$	$\begin{pmatrix} 0.2 & 0.4582 & 0.8337 & 1 \\ 0.25 & 0.375 & 0.125 & 0.25 \end{pmatrix}$
RPD	$\begin{pmatrix} 0.01 & 0.3186 & 0.8872 & 1 \\ 0.25 & 0.375 & 0.25 & 0.125 \end{pmatrix}$	$\begin{pmatrix} 0.2 & 0.4772 & 0.9316 & 1 \\ 0.25 & 0.375 & 0.125 & 0.25 \end{pmatrix}$
HCMMD	$\begin{pmatrix} 0.01 & 0.2122 & 0.6354 \\ 0.25 & 0.375 & 0.375 \end{pmatrix}$	$\begin{pmatrix} 0.2 & 0.3868 & 0.8163 \\ 0.25 & 0.375 & 0.375 \end{pmatrix}$
RPMMD	$\begin{pmatrix} 0.01 & 0.3412 & 1 \\ 0.25 & 0.375 & 0.375 \end{pmatrix}$	$\begin{pmatrix} 0.2 & 0.4644 & 1 \\ 0.25 & 0.375 & 0.375 \end{pmatrix}$
Bayesian D	$\begin{pmatrix} 0.01 & 0.3219 & 0.8124 & 0.9686 \\ 0.25 & 0.375 & 0.25 & 0.125 \end{pmatrix}$	$\begin{pmatrix} 0.2 & 0.4766 & 0.9905 \\ 0.5 & 0.375 & 0.125 \end{pmatrix}$

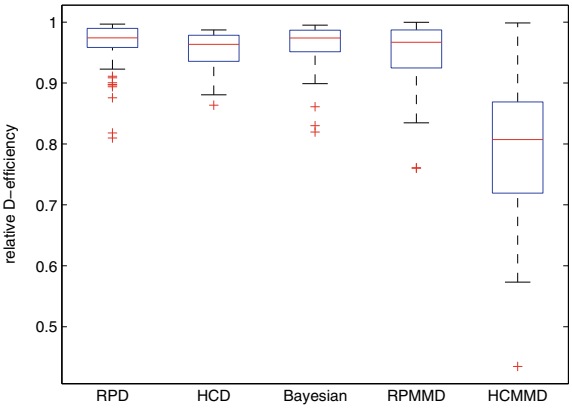


Fig. 9.5 Box plots of the D-efficiencies of the five optimal 8-exact designs relative to the 100 locally D-optimal 8-exact designs on [0.01, 1] for the second-order model (9.6) with random effects covariance matrix $Cov(b) = 0.5\, I_3$

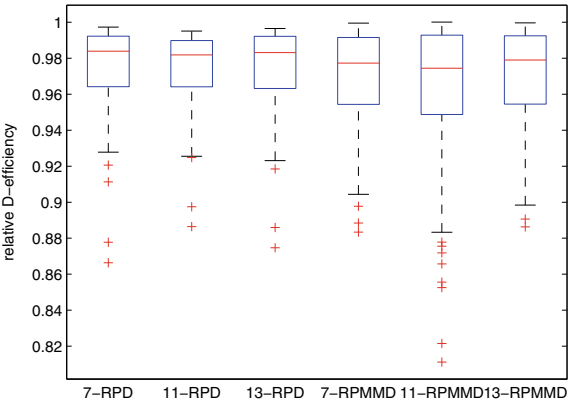


Fig. 9.6 Box plots of the D-efficiencies of the RPD-and RPMMD-optimal 8-exact designs under three sets $\Theta_{7\text{-}RP}$,$\Theta_{11\text{-}RP}$, $\Theta_{13\text{-}RP}$ with respect to the 100 locally D-optimal 8-exact designs on [0.01, 1] for the second-order model (9.6) with random effects covariance matrix $Cov(b) = 0.5\, I_3$

We here examine the affect of the number of RPs used in the RPD- and RPMMD-optimality criteria on the optimal designs for the model (9.6). The RPD-and RPMMD-optimal 8-exact designs on [0.01, 1] under three sets $\Theta_{n\text{-}RP}$ ($n = 7, 11, 13$) are calculated numerically, and the D-efficiencies of these designs are computed with respect to the locally D-optimal 8-exact designs at each of the 100 values of β which are randomly sampled from the prior distribution $N_3(\bar{\beta}, I_3)$. These results in Fig. 9.6 show that the number of RPs has a slight impact on the RPD-and RPMMD-optimal designs.

9.5 Concluding Remarks

This paper concerns with optimal and robust design problems for Poisson mixed models. Two optimality criteria, termed RPD-optimality and RPMMD-optimality, for the Poisson mixed model are introduced by using the RPs of the prior distribution of fixed effects. The purpose of these two criteria is to overcome the dependence problem of D-optimality on the values of unknown parameters. By assuming the prior distribution of fixed effects is a multivaraite normal distribution with independent components, we obtain the RPs by using the transformation method. The numerical results for the first-and second-order models show that the optimal designs based on the RPs are more robust than those based on the hypercube method. Moreover, the number of RPs has a slight impact on both RPD-and RPMMD-optimal designs. Therefore, a small number of RPs used in the RPD-and RPMMD-optimality criteria may yield a good robustness against parameter uncertainty. Hence, our results will give more options to the experimenters.

In aspects of computation, the running times of constructing the RPD-and RPMMD-optimal designs are much less than that of the HCD-and HCMMD-optimal designs, respectively. The computation time of constructing the Bayesian D-optimal

design is much longer than others due to the long time required in computation of the Bayesian criterion for a given design.

Moreover, in our computation the prior distribution of the fixed effects is assumed to have independent components, and then the RPs are obtained by using the inverse transformation method. If the prior distributions of the fixed effects have correlated components, the RPs can be generated by other methods proposed in, e.g., [5, 18, 19] and the references therein.

Acknowledgments The work is supported by the NSFC grants 11971318, 11871143.

References

1. Atkinson, A.C., Cox, D.C.: Planning experiments for discriminating between models. J. Roy. Statist. Soc. Ser. B **36**, 321–348 (1974)
2. Atkinson, A.C., Donev, A.N., Tobias, R.D.: Optimal Experimental Designs, With SAS. Oxford university Press, New York (2007)
3. Bohachevsky, I.O., Johnson, M.E., Stein, M.L.: Generalized simulated annealing for function optimization. Technometrics **28**, 209–217 (1986)
4. Dette, H., Haines, L.M., Imhof, L.A.: Maximin and Bayesian optimal designs for regression models. Statist. Sinica **17**, 463–480 (2007)
5. Fang, K.T., Wang, Y.: Number-Theoretic Methods in Statistics. Chapman and Hall, London (1994)
6. Haines, L.M.: The application of the annealing algorithm to the construction of exact optimal designs for linear regression models. Technometrics **29**, 439–447 (1987)
7. Haines, L.M.: A geometric approach to optimal design for one-parameter non-linear models. J. Roy. Statist. Soc. Ser. B **57**, 575–598 (1995)
8. Khuri, A.I., Mukherjee, B., Sinha, B.K., Ghosh, M.: Design issue for generalized linear models: a review. Statist. Sci. **21**, 376–399 (2006)
9. Foo, L.K., Duffull, S.: Methods of robust design of nonlinear models with an application to pharmacokinetics. J. Biopharm. Statist. **20**, 886–902 (2010)
10. McCullagh, P.M., Nelder, J.A.S.: Generalized Linear Models, 2nd edn. Chapman and Hall, London (1989)
11. McGree, J.M., Duffull, S.B., Eccleston, J.A.: Optimal design for studying bioimpedance. Physiol. Meas. **28**, 1465–1483 (2007)
12. Niaparast, M.: On optimal design for a Poisson regression model with random intercept. Statist. Probab. Lett. **79**, 741–747 (2009)
13. Niaparast, M.: On optimal design for mixed effects Poisson regression models. Ph.D. thesis, Otto-von-Guericke University, Magdeburg (2010)
14. Niaparast, M., Schwabe, R.: Optimal design for quasi-likelihood estimation in Poisson regression with random coefficients. J. Statist. Plann. Inference **143**, 296–306 (2013)
15. Wang, Y.P.: Optimal experimental designs for the Poisson regression model in toxicity studies. Ph.D. thesis, Virginia Polytechnic Institute and State University (2002)
16. Wedderburn, R.W.M.: Quasi-likelihood functions, generalized linear models, and the Gauss-Newton method. Biometrika **61**, 439–447 (1974)
17. Wong, W.K.: A unified approach to the construction of minimax designs. Biometrika **79**, 611–619 (1992)

18. Zhou, M., Wang, W.J.: Representative points of Student's t_n distribution and their application in statistical simulation (in Chinese). Acta. Math. Appl. Sinica. Chin. Ser. **39**, 620–640 (2016)
19. Zhou, Y.D., Fang, K.T.: FM-criterion for representative points (in Chinese). Sci. Chin. Math. **49**, 1009–1020 (2019)

Chapter 10
Study of Central Composite Design and Orthogonal Array Composite Design

Si Qiu, Minyu Xie, Hong Qin, and Jianhui Ning

Abstract Response surface methodology (RSM) is an effective tool for exploring the relationships between the response and the input factors. Central composite design (CCD) and orthogonal array composite design (OACD) are useful second-order designs in response surface methodology. In this work, we consider the efficiencies of the two classes of composite designs for general case. Assuming the second-order polynomial model, the D-efficiency of CCDs and OACDs are studied for general value of α in star points. Moreover, the determination of α is also discussed from the perspective of space-filling criterion.

Keywords Central composite design · Orthogonal array composite design · Centered L_2-discrepancy · D-efficiency

10.1 Introduction

Considering a process or system involves a response y that depends on factors $\mathbf{x} = (x_1, \cdots, x_k)$, and their relationship can be modeled by

$$y = f(x_1, \cdots, x_k) + \varepsilon, \tag{10.1}$$

where the function f is unknown and ε is the error term that represents the sources of variability not captured by f. In order to estimate the f, many methods have been

S. Qiu · M. Xie · H. Qin · J. Ning (✉)
School of Mathematics and Statistics, Central China Norm University, Wuhan, China
e-mail: jhning@mail.ccnu.edu.cn

S. Qiu
e-mail: ccnuqiu@qq.com

M. Xie
e-mail: myxie@mail.ccnu.edu.cn

H. Qin
e-mail: qinhong@mail.ccnu.edu.cn

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_10

proposed, response surface methodology (RSM, [7]) is an appealing technique to achieve the goal that involves experimentation, modeling, data analysis and optimization. The main idea of RSM is to use a sequence of designed experiments to obtain an optimal response. Typically, there are three stages in RSM: The first stage is to detect the significant factors by modeling a first-order polynomial model with a factorial experiment design or a fractional factorial design. The second stage is to search the optimum region; Once the first two stages has been done successfully, then a more complicated model is employed to approximate the f [7]. suggested using a second-order polynomial model to do this due to its easy estimation and application. A comprehensive account of response surface methodology can be referred to [5, 14], and [18].

Given the quantitative factors denoted by $x_1, \cdots, x_k$, a second-order polynomial model is

$$y = \beta_0 + \sum_{i=1}^{k} \beta_i x_i + \sum_{i=1}^{k} \beta_{ii} x_i^2 + \sum_{i=1<j\leq k} \beta_{ij} x_i x_j + \varepsilon, \tag{10.2}$$

where β_0, β_i, β_{ii}, β_{ij} are the intercept, linear, quadratic and bilinear terms respectively, and ε is the error term. Many second-order designs, which allow all parameters in (10.2) to be estimated, have been proposed in the literature. The common way that researchers address second-order model in the literature is to use the central composite designs (CCDs) introduced by [7]. According to [25], for k factors, a composite design consists of three parts: (i) n_1 cube points (or corner points) with all $x_i = 1$ or -1; (ii) n_2 star points (or additional points) with all $x_i = \alpha$ or $-\alpha$; (iii) n_0 center points with all $x_i = 0$. The composite design has a total of $n_1 + n_2 + n_0$ points and has three or five different levels. If $\alpha = 1$, the composite design has three different levels, otherwise, the design has five different levels. Central composite designs use $n_2 = 2k$ axial points of the form $(0, \cdots, x_i, \cdots, 0)$ with $x_i = \alpha$ or $-\alpha$ for $i = 1, \cdots, k$ as the star points.

There are also other variations such as small composite designs (SCDs) proposed by [8] and augmented-pair designs (APDs) proposed by [17]. Motivated by an antiviral drug experiment [24], introduced a new class of composite designs, called orthogonal array composite designs (OACDs), which use a three-level orthogonal array as the star points. An orthogonal array of n runs, k columns, s levels, and strength t, denoted by $OA(n, s^k, t)$, is an $n \times k$ matrix in which all s^t level combinations appear equally often in every $n \times t$ submatrix. Detailed discussion of orthogonal array can see [1, 11].

Reference [25] developed some general theoretical results for CCDs and OACDs, and derived bounds of their efficiencies for estimating all and part of the parameters in a second-order model for $\alpha = 1$. In this paper, we generalized their theory to more general value of α, and discuss how to choose a better α from the perspective of space-filling view.

The rest of paper is organized as follows. In Sect. 10.2, we present some preliminaries on the CCD, OACD and optimum criteria. In Sect. 10.3 , we show the D-efficiency values and bounds of CCDs and OACDs respectively. In Sect. 10.4,

the determination of the α is also discussed. Section 10.5 is devoted to conclusion remarks, and the proofs of the main conclusions are listed in the appendix.

10.2 Preliminaries

CCD and OACD

A CCD with k input factors consist of following three parts:

- n_1 cube points with $x_i = 1$ or -1 for $i = 1, \cdots, k$;
- n_2 axial points of the form $(0, \cdots, x_i, \cdots, 0)$ with $x_i = \alpha$ or $-\alpha$ for $i = 1, \cdots, k$;
- n_0 center points with all $x_i = 0$ for $i = 1, \cdots, k$.

An OACD with k factors has three parts as follows:

- n_1 cube points with $x_i = 1$ or -1 for $i = 1, \cdots, k$;
- a 3-level orthogonal array with n_2 runs;
- n_0 center points with all $x_i = 0$ for $i = 1, \cdots, k$.

D-efficiencies

Let d be the k-factor composite design, $X = (\mathbf{1}, Q, L, B)$ be the model matrix of the second-order model (10.2), where $\mathbf{1}$ is a column of ones, Q, L and B are quadratic, linear and bilinear terms, respectively. Let d_i be the part i of the design for $i = 1, 2$. The total number of runs of d is $N = n_1 + n_2 + n_0$. The D-efficiency of d is

$$D(d) = N^{-1}|X^T X|^{1/p}, \tag{10.3}$$

where $p = (k + 1)(k + 2)/2$ is the number of parameters in the second-order model (10.2).

Sometimes the partial efficiency (D_s-efficiency) describes the precision for estimating a subset s of the model parameters. D_s-efficiency can be defined as

$$D_s(d) = N^{-1}|X_s^T X_s - X_s^T X_{(s)}(X_{(s)}^T X_{(s)})^{-1} X_{(s)}^T X_s|^{1/t},$$

where s is a subset of factors, X_s and $X_{(s)}$ are the sub-matrices of X corresponding to the parameters in s or not in s, respectively, and t is the number of parameters in s. Since

$$|X^T X| = |X_{(s)}^T X_{(s)}||X_s^T X_s - X_s^T X_{(s)}(X_{(s)}^T X_{(s)})^{-1} X_{(s)}^T X_s|,$$

then the D_s-efficiency can also be calculated by

$$D_s(d) = N^{-1}\left(\frac{|X^T X|}{|X_{(s)}^T X_{(s)}|}\right)^{1/t}. \tag{10.4}$$

For convenience, we denote D_L, D_B and D_Q as the D_s-efficiency, when the subset s is the linear, bilinear and quadratic terms, respectively.

10.3 *D*-efficiencies of CCDs and OACDs

In this section, we deduce some results of CCDs and OACDs based on D-efficiencies with general α value. We consider a CCD with k factors and n_0 center points. When an $OA(n_1, 2^k, 4)$ (or a full factorial for $k < 4$) is used for the two-level portion of CCD, the linear, quadratic, and bilinear terms are orthogonal to each other, that is, $Q^T L = 0, Q^T B = 0, L^T B = 0$. The information matrix of the central composite design is a block diagonal matrix

$$X'X = \begin{pmatrix} N & (n_1+2\alpha^2)\mathbf{1}_k' & \mathbf{0} & \mathbf{0} \\ (n_1+2\alpha^2)\mathbf{1}_k & n_1 J_k + 2\alpha^4 I_k & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & (n_1+2\alpha^2)I_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & n_1 I_q \end{pmatrix},$$

where $\mathbf{1}_k$ is a column of k ones, I_k is $k \times k$ identity matrix, J_k is the $k \times k$ matrix of ones. So it is easy to obtain that

$$|X'X| = n_1^q (2\alpha^4 n_1 + 4\alpha^6)^k \left[\left(1 + \frac{kn_1}{2\alpha^4}\right) n_0 + \left(1 - \frac{k}{\alpha^2}\right)^2 n_1\right],$$

$$|X'_{(L)} X_{(L)}| = n_1^q (2\alpha^4)^k \left[\left(1 + \frac{kn_1}{2\alpha^4}\right) n_0 + \left(1 - \frac{k}{\alpha^2}\right)^2 n_1\right],$$

$$|X'_{(B)} X_{(B)}| = (2\alpha^4 n_1 + 4\alpha^6)^k \left[\left(1 + \frac{kn_1}{2\alpha^4}\right) n_0 + \left(1 - \frac{k}{\alpha^2}\right)^2 n_1\right],$$

$$|X'_{(Q)} X_{(Q)}| = N(n_1 + 2\alpha^2)^k n_1^q.$$

From these equations, we obtain the following Theorem 10.1.

Theorem 10.1 *For a CCD with k factors and n_0 center points, if the 2-level portion is an $OA(n_1, 2^k, 4)$, its D-, D_L-, D_B- and D_Q- efficiency are, respectively,*

$$D(CCD) = \frac{1}{N}\left\{n_1^q(2\alpha^4 n_1 + 4\alpha^6)^k\left[\left(1+\frac{kn_1}{2\alpha^4}\right)n_0 + \left(1-\frac{k}{\alpha^2}\right)^2 n_1\right]\right\}^{1/p}, \tag{10.5}$$

$$D_L(CCD) = \frac{1}{N}(n_1 + 2\alpha^2), \tag{10.6}$$

$$D_B(CCD) = \frac{n_1}{N},$$

$$D_Q(CCD) = \frac{2\alpha^4}{N^{\frac{k+1}{k}}}\left[\left(1+\frac{kn_1}{2\alpha^4}\right)n_0 + \left(1-\frac{k}{\alpha^2}\right)^2 n_1\right]^{1/k}, \tag{10.7}$$

where $N = n_1 + 2k + n_0$ *is the total number of runs,* $p = (k+1)(k+2)/2$ *and* $q = k(k-1)/2$ *is the number of all parameters and bilinear terms parameters respectively in the second-order model (10.2).*

It can be seen from Theorem 10.1 that the α value affects the efficiencies of CCDs except the D_B-efficiency. Obviously, D-efficiency, D_L-efficiency and D_Q-efficiency increase as α increases. From the perspective of D-efficiency, a larger α is more favorable. In practice, D-efficiency is not the only criterion we care about. Some other criteria are also important. Hence, the α value should be determined carefully.

Next we consider the efficiencies of OACDs. Because of the information matrix and efficiencies for OACDs depend on the specific 3-level orthogonal array, we cannot get general theoretical results for the D-efficiency and D_s-efficiencies. The lower bounds of the efficiencies can be summarized in following Theorem 10.2 and Theorem 10.3. For the D_L-efficiency, we can obtain the upper bound.

Theorem 10.2 *Let an* $OA(n_1, 2^k, 4)$ *be the first part and an* $OA(n_2, 3^k)$ *be the second part of the OACD. Then the determinant of its information matrix and D-efficiency have the following lower bounds, respectively,*

$$|X^T X| \geq n_1^q\left[\frac{(4n_2\alpha^2 + 6n_1)\alpha^2 n_2}{27}\right]^k \cdot \eta,$$

$$D(OACD) \geq LB(OACD) = N^{-1}n_1^{q/p}\left[\frac{(4n_2\alpha^2 + 6n_1)\alpha^2 n_2}{27}\right]^{k/p} \cdot \eta^{1/p}, \tag{10.8}$$

where $\eta = (1+2k\alpha^2)n_0 + n_2 + n_1\left(1 + 2k\alpha^2 + \frac{9kn_0}{2n_2\alpha^2} + \frac{9k}{2\alpha^2} - 6k\right)$, $N = n_1 + n_2 + n_0$, $q = k(k-1)/2$, $p = (k+1)(k+2)/2$.

Theorem 10.3 *Suppose that an OACD satisfies the conditions in Theorem 10.2. Its* D_L*-,* D_B*-,* D_Q*-efficiencies have the following lower bounds, respectively,*

$$D_L(OACD) \geq \frac{1}{N}\left(\frac{9n_1}{9n_1 + 4n_2\alpha^4}\right)^{q/k}\left(\frac{2}{3}n_2\alpha^2 + n_1\right), \tag{10.9}$$

$$D_B(OACD) \geq \frac{n_1}{N}, \tag{10.10}$$

$$D_Q(OACD) \geq \frac{2n_2\alpha^2}{9N^{\frac{k+1}{k}}} \left(\frac{9n_1}{9n_1 + 4n_2\alpha^4}\right)^{q/k} \cdot \eta^{1/k}, \tag{10.11}$$

where the η, N and q are the same as in the Theorem 10.2. Furthermore, its D_L-efficiency has an upper bound

$$D_L(OACD) \leq \frac{1}{N}\left(\frac{2}{3}n_2\alpha^2 + n_1\right), \tag{10.12}$$

the equality holds when the linear terms of second part of the design are orthogonal to the bilinear terms of second part of the design.

Theorems 10.2 and 10.3 show that the lower bounds of D-efficiency, D_L-efficiency, D_Q-efficiency and the upper bound of D_L-efficiency increase as α increases. By comparing the results of the three theorem, we know that an OACD has larger D-efficiency than a CCD for the same α, especially when k is large. In order to compare the two classes of composite designs intuitively, next we give an example to compare the efficiency of the composite designs.

Example 10.4 We compare OACDs with CCDs consisting of the same 2-level portion for $k = 4, \cdots, 12$. We choose a full factorial design 2^k for $k = 4$ or a regular 2^{k-m} design with resolution at least V for $k = 5, \cdots, 11$. For $k = 12$, we use an $OA(128, 2^{15}, 4)$ from [23] as the 2-level portion. For the 3-level OA, we choose the first k columns of oa.9.4.3.2.txt, oa.18.7.3.2.txt, and oa.27.13.3.2.txt from Sloane's website http://neilsloane.com/oadir/.

Table 10.1 shows the D-efficiencies of OACDs and CCDs as well as the lower bound for $\alpha = 1$ and $\alpha = 1.5$, respectively, with $n_0 = 5$ center points. For every $k \geq 4$, an OACD has larger D-efficiency than a CCD for every α. And the lower bound of OACD is also larger than CCD's D-efficiency when $\alpha = 1$. When $\alpha = 1.5$, the lower bound of OACD is no longer larger than CCD's D-efficiency for $k = 4, \cdots, 7$, but the D-efficiency of OACDs still larger than CCD's. From Fig. 10.1 we can get this result visually.

As we can see that orthogonal array plays an importance role in OACD. Actually Different orthogonal array used in the OACDs, even the isomorphic orthogonal array, may lead to different efficiency. It means if we permute the levels of the orthogonal array, we may find a better OACD with higher D-efficiency. We use the following example to illustrate it.

Example 10.5 We consider two OACDs d_1 and d_2 for $k = 4$ and $\alpha = 1$, the only different between them is the orthogonal array. The OA's used in d_1 and d_2 are

Table 10.1 *D*-efficiency of OACDs and CCDs

			$\alpha = 1$			$\alpha = 1.5$		
k	d_1	d_2	$D(OACD)$	$LB(OACD)$	$D(CCD)$	$D(OACD)$	$LB(OACD)$	$D(CCD)$
4	2^4	$OA(9, 3^4)$	0.42108	0.40179	0.39835	0.72977	0.53363	0.56000
5	2_V^{5-1}	$OA(18, 3^5)$	0.44523	0.38248	0.37968	0.93602	0.51456	0.52616
6	2_{VI}^{5-1}	$OA(18, 3^6)$	0.48160	0.43847	0.41672	0.86086	0.55279	0.55928
7	2_{VII}^{7-1}	$OA(18, 3^7)$	0.50102	0.47804	0.44163	0.77859	0.57529	0.57837
8	2_V^{8-2}	$OA(27, 3^8)$	0.52416	0.48455	0.44919	0.86378	0.58210	0.57862
9	2_V^{9-2}	$OA(27, 3^9)$	0.54165	0.51972	0.46666	0.79590	0.60460	0.59011
10	2_V^{10-3}	$OA(27, 3^{10})$	0.55634	0.53500	0.47925	0.79343	0.61569	0.59763
11	2_V^{11-4}	$OA(27, 3^{11})$	0.56969	0.54881	0.48990	0.79161	0.62566	0.60319
12	$OA(128, 2^{12}, 4)$	$OA(27, 3^{12})$	0.58118	0.56132	0.49885	0.78745	0.63466	0.60717

$n_0 = 5$ is used for all the designs

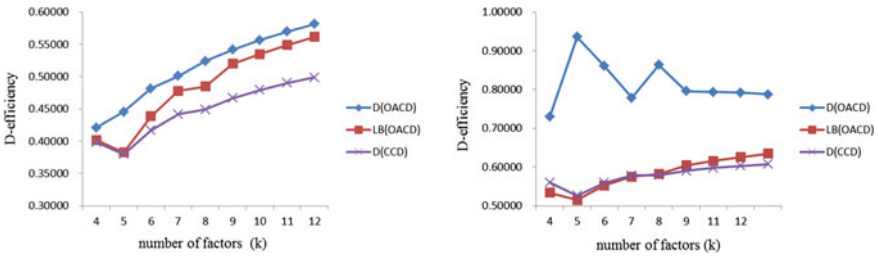


Fig. 10.1 Comparison of D-efficiency between OACDs and CCDs for $\alpha = 1$ (left) and $\alpha = 1.5$ (right) with $n_0 = 5$

$$A_1 = \begin{pmatrix} -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 \\ 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 & -1 \\ -1 & 0 & 1 & 0 & 1 & -1 & 1 & -1 & 0 \\ -1 & 1 & 0 & 0 & -1 & 1 & 1 & 0 & -1 \end{pmatrix}^T$$

and

$$A_2 = \begin{pmatrix} -1 & -1 & -1 & 0 & 0 & 0 & 1 & 1 & 1 \\ -1 & 0 & 1 & -1 & 0 & 1 & -1 & 0 & 1 \\ -1 & 0 & 1 & 0 & 1 & -1 & 1 & -1 & 0 \\ -1 & 1 & 0 & 0 & -1 & 1 & 1 & 0 & -1 \end{pmatrix}^T,$$

respectively. Actually, A_1 and A_2 are isomorphic, since A_2 can be obtained by permuting the levels in the second column of A_1. When $n_0 = 5$, the D-efficiency of the OACD d_1 is 0.40179. This value happens to be the lower bound of OACD for $k = 4$ as shown in Table 10.1. And D-efficiency of the OACD d_2 is 0.42108. So permuting levels may get an OACD with a higher efficiency.

10.4 The Determination of the α Value

For CCD [22], gives some suggestions on the determination of the α value. In general, α should be chosen between 1 and $\sqrt{k}$, and rarely outside this range. For $\alpha = 1$, the star points are placed at the center of the faces of the cube, $\alpha = \sqrt{k}$ makes the star points and cube points lie on the same sphere. The efficiency of the parameter estimates is increased by pushing the star points toward the extreme, however, for large k, this choice should be taken with caution because the star points are too far from the center point and no information is available for the response surface in the intermediate range of the factors, especially along the axes. From Example 10.4, we can find the effect of the α value on efficiency, the larger α is, the greater efficiency is. In general, if it is desired to collect information closer to the faces of the cube, a smaller α value should be chosen. If the estimation efficiency is concerned, the star points should be pushed toward the extremes of the region, namely choosing a larger α.

Reference [6] provides a criterion, called rotatability, for CCD to determine the α value. Denote the predicted response at $X = (x_1, \cdots, x_k)$ by $\widehat{y}(X)$. A design is called rotatable if $Var(\widehat{y}(X))$ depends only on $\|X\| = \left(x_1^2 + \cdots + x_k^2\right)^{1/2}$. If the central composite design is rotatable, then $\alpha = 2^{(k-m)/4}$. Another criterion is orthogonality. In central composite design, let b_0, b_i, b_{ii}, b_{ij} denote the least square estimators of $\beta_0, \beta_i, \beta_{ii}, \beta_{ij}$ respectively, all the covariances between estimated regression coefficient except $cov(b_{ii}, b_{ij})$ are zero. But if the inverse of the information matrix $(X'X)^{-1}$ is a diagonal matrix, then $cov(b_{ii}, b_{ij})$ also becomes zero. This property is called orthogonality. The condition for making a central composite design orthogonal is by setting $\alpha = \left(\frac{\sqrt{Nn_1}-n_1}{2}\right)^{1/2}$, see [4] for more details.

In this paper, we want to determine the α value from another view. What we need to pay attention to is that, above discussion is under the second-order model assumption. When the second-order polynomial model does not fit the true model very well, we need to consider some more robust criterion, instead of the D-efficiency. As well known, the uniform design is a widely used robust space-filling design method. That is the reason that we propose to determine the α value from the perspective of measure of uniformity. The uniform design is usually measured by discrepancy, such as the centered L_2-discrepancy, more detail introduction can refer to [9]. Here, we only discussed the method under centered L_2-discrepancy. Other discrepancies can be similarly done. The centered L_2-discrepancy can be defined as:

$$CD_2(P) = \left\{ \left(\frac{13}{12}\right)^k - \frac{2}{N}\sum_{i=1}^{N}\prod_{j=1}^{k}\left(1 + \frac{1}{2}|x_{ij} - 0.5| - \frac{1}{2}|x_{ij} - 0.5|^2\right) + \frac{1}{N^2}\sum_{i=1}^{N}\sum_{s=1}^{N}\prod_{j=1}^{k}\left(1 + \frac{1}{2}|x_{ij} - 0.5| + \frac{1}{2}|x_{sj} - 0.5| - \frac{1}{2}|x_{ij} - x_{sj}|\right)\right\}^{1/2}$$

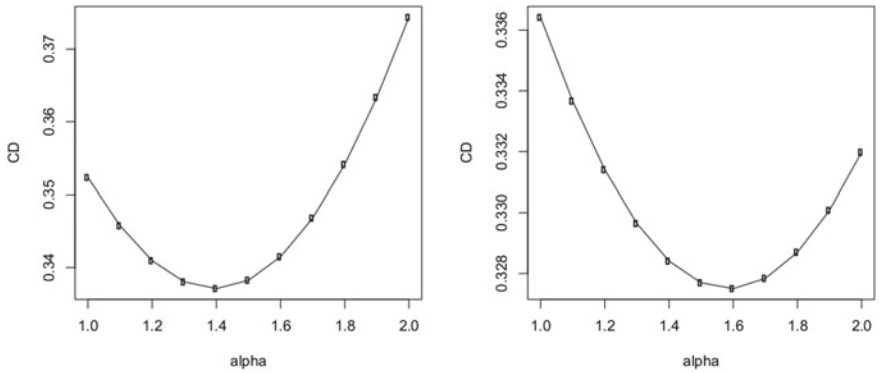


Fig. 10.2 L_2-Centered discrepancy of OACDs (left) and CCDs (right) for $k = 4$

where P is the design and N is the runs of the design. Before calculating the L_2-centered discrepancy, we need to make a transformation to let points distributed on $[0, 1]^k$. Next we give a simple example to show our idea.

Example 10.6 Suppose that the response y depends on four input factors x_1, x_2, x_3, x_4. $x_1 \in [11.20, 13.60]$, $x_2 \in [1.98, 3.04]$, $x_3 \in [0.75, 1.75]$, $x_4 \in [1.00, 3.00]$. For the sake of later calculation easier and eliminate the influence of variable dimension, we convert x_i to coded variables firstly. We transfer the lower bound of the actual level into -2 and the upper bound into 2, then α can be arbitrary value from 1 to 2. We calculate L_2-centered discrepancy of OACDs and CCDs for different α.

Figure 10.2 shows the tendency of L_2-centered discrepancy changing with different α. From the two figures, we can find that L_2-centered discrepancy of two classes composite designs decreases first and then increases. The α value really has influence on uniformity of the design. We can find that the OACD is the most uniformly design for $\alpha = 1.4$ and the CCD is the most uniformly design for $\alpha = 1.6$. So we can determine the α value by calculating the discrepancy and choosing a α that make the discrepancy minimum.

10.5 Conclusion Remarks

We study the estimation efficiencies of CCDs and OACDs under a second-order polynomial model for general α value. We find that OACD are more effective in estimating the parameters than CCD especially the number of factors is large. The OACD provide a good trade-off between estimation efficiency and run size economy, so it can be used as an alternative to the popular CCD. We also suggest some criteria to choose the α value. Different criteria will make different results. In practice, the determination of the α value also depends on the objectives of each experiment

and the geometric nature of and the practical constraints on the design region. We proposed an idea to determine α value from the the perspective of space-filling, make the design points distributed on the target area more uniformly, but we haven't given the theoretical results. We hope to address this issue in future work.

Acknowledgements This research was partially supported by a grant from the Natural Science Foundation of China (No.11571133 and 11871237).

Appendix

Lemma 10.7 *Let $a \neq 0$, $b \neq 0$,*

$$\begin{pmatrix} c_0 & c\boldsymbol{1}_k' \\ c\boldsymbol{1}_k & a\boldsymbol{J}_k + b\boldsymbol{I}_k \end{pmatrix} = b^{k-1}(bc_0 + k(ac_0 - c^2)).$$

Lemma 10.8 *Let E and F be two $n \times n$ nonnegative definite matrices with partions*

$$E = \begin{pmatrix} E_1 & \boldsymbol{0} \\ \boldsymbol{0} & E_2 \end{pmatrix} \geq 0, \ F = \begin{pmatrix} F_1 & F_3 \\ F_3' & F_2 \end{pmatrix} \geq 0,$$

where E_1 and F_1 are $m \times m$ matrices. Then

$$|E + F| \geq |E_2| \cdot |E_1 + F_1|.$$

Proof of Theorem 10.2 Denote $X_0 = (\mathbf{1}_{n_0}, \mathbf{0}, \mathbf{0}, \mathbf{0})$ and $X_i = (\mathbf{1}_{n_i}, Q_i, L_i, B_i)$, where Q_i, L_i, B_i respectively are the quadratic, linear and bilinear terms of d_i in the second-order model, $i = 1, 2$.

$$X_1'X_1 = \begin{pmatrix} n_1 & n_1\mathbf{1}_k' & \mathbf{0} & \mathbf{0} \\ n_1\mathbf{1}_k & n_1J_k & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & n_1I_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & n_1I_q \end{pmatrix} = \begin{pmatrix} n_1J_{k+1} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & n_1I_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & n_1I_q \end{pmatrix},$$

$$X_2'X_2 = \begin{pmatrix} n_2 & \frac{2}{3}n_2\alpha^2\mathbf{1}_k' & \mathbf{0} & \mathbf{0} \\ \frac{2}{3}n_2\alpha^2\mathbf{1}_k & \frac{4}{9}n_2\alpha^4J_k + \frac{2}{9}n_2\alpha^4I_k & \mathbf{0} & Q_2'B_2 \\ \mathbf{0} & \mathbf{0} & \frac{2}{3}n_2\alpha^2I_k & L_2'B_2 \\ \mathbf{0} & B_2'Q_2 & B_2'L_2 & B_2'B_2 \end{pmatrix},$$

let $Y = X_2'X_2 + X_0'X_0$, then

$$Y = \begin{pmatrix} n_2 + n_0 & \frac{2}{3}n_2\alpha^2\mathbf{1}_k' & \mathbf{0} & \mathbf{0} \\ \frac{2}{3}n_2\alpha^2\mathbf{1}_k & \frac{4}{9}n_2\alpha^4 J_k + \frac{2}{9}n_2\alpha^4 I_k & \mathbf{0} & Q_2'B_2 \\ \mathbf{0} & \mathbf{0} & \frac{2}{3}n_2\alpha^2 I_k & L_2'B_2 \\ \mathbf{0} & B_2'Q_2 & B_2'L_2 & B_2'B_2 \end{pmatrix},$$

denote

$$B_{11} = \begin{pmatrix} n_2 + n_0 & \frac{2}{3}n_2\alpha^2\mathbf{1}_k' \\ \frac{2}{3}n_2\alpha^2\mathbf{1}_k & \frac{4}{9}n_2\alpha^4 J_k + \frac{2}{9}n_2\alpha^4 I_k \end{pmatrix}, B_{13} = \begin{pmatrix} \mathbf{0} \\ Q_2'B_2 \end{pmatrix},$$

then

$$X'X = X_1'X_1 + Y = \begin{pmatrix} B_{11} + n_1 J_{k+1} & \mathbf{0} & B_{13} \\ \mathbf{0} & (\frac{2}{3}n_2\alpha^2 + n_1)I_k & L_2'B_2 \\ B_{13}' & B_2'L_2 & n_1 I_q + B_2'B_2 \end{pmatrix}, \quad (10.13)$$

from Lemma 10.8, we get

$$|X'X| = |X_1'X_1 + Y| \geq |n_1 I_q| \cdot \begin{vmatrix} B_{11} + n_1 J_{k+1} & \mathbf{0} \\ \mathbf{0} & (\frac{2}{3}n_2\alpha^2 + n_1)I_k \end{vmatrix} = n_1^q \left(\frac{2}{3}n_2\alpha^2 + n_1\right)^k |B_{11} + n_1 J_{k+1}|,$$

from Lemma 10.7, we have

$$|B_{11} + n_1 J_{k+1}| = \left(\frac{2}{9}n_2\alpha^2\right)^k \left[(1 + 2k\alpha^2)n_0 + n_2 + n_1\left(1 + 2k\alpha^2 + \frac{9kn_0}{2n_2\alpha^2} + \frac{9k}{2\alpha^2} - 6k\right)\right],$$

therefore

$$|X'X| \geq n_1^q \left[\frac{(4n_2\alpha^2 + 6n_1)\alpha^2 n_2}{27}\right]^k \left[(1 + 2k\alpha^2)n_0 + n_2 + n_1\left(1 + 2k\alpha^2 + \frac{9kn_0}{2n_2\alpha^2} + \frac{9k}{2\alpha^2} - 6k\right)\right], \quad (10.14)$$

then we can obtain Theorem 10.2.

Proof of Theorem 10.3 When $s = L$, from Eq. (10.13) and Fischer inequality, we have

$$|X_{(L)}'X_{(L)}| \leq |B_{11} + n_1 J_{k+1}| \cdot |n_1 I_q + B_2'B_2|,$$

because all of the diagonal elements of $B_2'B_2$ are $\frac{4}{9}n_2\alpha^4$, we have

$$|n_1 I_q + B_2'B_2| \leq \left(n_1 + \frac{4}{9}n_2\alpha^4\right)^q, \quad (10.15)$$

so

$$|X_{(L)}'X_{(L)}| \leq |B_{11} + n_1 J_{k+1}| \cdot \left(n_1 + \frac{4}{9}n_2\alpha^4\right)^q,$$

then using Eq. (10.4) and (10.14), we obtain the lower bound of D_L-efficiency. Moreover, from Fischer inequality, we have

$$|X'X| \leq |X'_{(L)}X_{(L)}| \cdot |X'_L X_L|,$$

so

$$\frac{|X'X|}{|X'_{(L)}X_{(L)}|} \leq |X'_L X_L| = \left|\left(\frac{2}{3}n_2\alpha^2 + n_1\right) I_k\right| = \left(\frac{2}{3}n_2\alpha^2 + n_1\right)^k,$$

then get the upper bound of D_L-efficiency, if the linear terms of d_2 are orthogonal to the bilinear terms of d_2, then

$$\frac{|X'X|}{|X'_{(L)}X_{(L)}|} = |X'_L X_L|,$$

and the upper bound of D_L-efficiency is achieved.

When $s = B$, from Eq. (10.13),

$$|X'_{(B)}X_{(B)}| = \begin{vmatrix} B_{11} + n_1 J_{k+1} & \mathbf{0} \\ \mathbf{0} & (\frac{2}{3}n_2\alpha^2 + n_1)I_k \end{vmatrix} = |B_{11} + n_1 J_{k+1}| \cdot \left(\frac{2}{3}n_2\alpha^2 + n_1\right)^k,$$

then follows from Eqs. (10.4) and (10.14), we get the lower bound of D_B-efficiency.

When $s = Q$, from Eq. (10.13) and Fischer inequality,

$$\begin{aligned} |X'_{(Q)}X_{(Q)}| &= \begin{vmatrix} N & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & (\frac{2}{3}n_2\alpha^2 + n_1)I_k & L'_2 B_2 \\ \mathbf{0} & B'_2 L_2 & n_1 I_q + B'_2 B_2 \end{vmatrix} \leq N\left(\frac{2}{3}n_2\alpha^2 + n_1\right)^k |n_1 I_q + B'_2 B_2| \\ &\leq N\left(\frac{2}{3}n_2\alpha^2 + n_1\right)^k \left(n_1 + \frac{4}{9}n_2\alpha^4\right)^q, \end{aligned}$$

then follows from Eq. (10.4), Theorem 10.2 and Eq. (10.15), we get the lower bound of D_Q-efficiency.

References

1. Ai, M., Kong, X., Li, K.: A general theory for orthogonal array based latin hypercube sampling. Stat. Sin. **26**(2), 761–777 (2016)
2. Ai, M., Li, P.-F., Zhang, R.-C.: Optimal criteria and equivalence for nonregular fractional factorial designs. Metrika **62**(1), 73–83 (2005)
3. Asadi, N., Zilouei, H.: Optimization of organosolv pretreatment of rice straw for enhanced biohydrogen production using enterobacter aerogenes. Bioresour. Technol. **227**, 335–344 (2017)
4. Oyejola, B.A., Nwanya, J.C.: Selecting the right central composite design. Int. J. Stat. Appl. **5**(1), 21–30 (2015)

5. Box, G.E.P., Draper, N.R.: Response Surfaces, Mixtures, and Ridge Analyses. John Wiley & Sons Inc, Hoboken, NJ, USA (2007)
6. Box, G.E.P., Hunter, J.S.: Multi-factor experimental designs for exploring response surfaces. Ann. Math. Stat **28**(1), 195–241 (1957)
7. Box, G.E.P., Wilson, K.B.: On the experimental attainment of optimum conditions. J. R. Stat. Soc. Ser. B, **13**(1), 1–45 (1951)
8. Draper, N.R., Lin, D.K.J.: Small response-surface designs. Technometrics **32**(2), 187 (1990)
9. Fang, K.-T., Lin, D.K., Winker, P., Zhang, Y.: Uniform design: theory and application. Technometrics **42**(3), 237–248 (2000)
10. Farrell, R.H., Kiefer, J., Walbran, A.: Optimum multivariate designs. In: Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, vol. 1, pp. 113–138. University of California Press, Berkeley, Calif (1967)
11. Hedayat, A.S., Sloane, N.J.A., Stufken, J.: Orthogonal Arrays: Theory and Applications. Springer Series in Statistics, Springer, New York (2013)
12. Jaynes, J., Zhao, Y., Xu, H., Ho, C.-M.: Use of orthogonal array composite designs to study lipid accumulation in a cell-free system. Qual. Reliab. Eng. Int. **32**(5), 1965–1974 (2016)
13. Karlin, S., Studden, W.J.: Optimal experimental designs. Ann. Math. Stat. **37**(4), 783–815 (1966)
14. Khuri, A.I., Cornell, J.A.: Response Surfaces: Designs and Analyses, volume 152 of Statistics : Textbooks and Monographs. Dekker, New York, 2nd, rev. and expanded. edition (1996)
15. Kiefer, J.: Optimum designs in regression problems, ii. Ann. Math. Stat. **32**(1), 298–325 (1961)
16. Lucas, J.M.: Optimum composite designs. Technometrics **16**(4), 561–567 (1974)
17. Morris, M.D.: A class of three-level experimental designs for response surface modeling. Technometrics **42**(2), 111–121 (2000)
18. Myers, R.H., Montgomery, D.C., Anderson-Cook, C.M.: Response Surface Methodology: Process and Product Optimization Using Designed Experiments. Wiley series in probability and statistics, 4th edn. Wiley, Hoboken, New Jersey (2016)
19. Park, S., Fowler, J.W., Mackulak, G.T., Keats, J.B., Carlyle, W.M.: D-optimal sequential experiments for generating a simulation-based cycle time-throughput curve. Oper. Res. **50**(6), 981–990 (2002)
20. Pesotchinsky, L.L.: D-optimum and quasi-d-optimum second-order designs on a cube. Biometrika **62**(2), 335–340 (1975)
21. Wald, A.: On the efficient design of statistical investigations. Ann. Math. Stat. **14**(2), 134–140 (1943)
22. Wu, C.-F.J., Hamada, M.: Experiments: Planning, Analysis, and Optimization. Wiley series in probability and statistics, 2nd edn. Wiley, Hoboken, N.J., (2009)
23. Xu, H.: Some nonregular designs from the nordstrom-robinson code and their statistical properties. Biometrika **92**(2), 385–397 (2005)
24. Xu, H., Jaynes, J., Ding, X.: Combining two-level and three-level orthogonal arrays for factor screening and response surface exploration. Stat. Sinica **24**, 269–289 (2014)
25. Zhou, Y., Xu, H.: Composite designs based on orthogonal arrays and definitive screening designs. J. Am. Stat. Assoc. **112**, 1675–1683 (2017)

Chapter 11
Uniform Design on Manifold

Yu Tang

Abstract Uniform design aims to scatter points as evenly as possible on certain domain. Although in real applications, the experimental domain is often quite arbitrary, the discrepancies frequently used to measure the uniformity of experimental designs are often defined on the unit cube. In this paper, we will introduce a unified framework to measure the uniformity of an experimental design on manifold. We will give some examples to illustrate the construction of uniform designs on some specific manifolds and provide a stochastic algorithm to construct uniform designs on the unit semi-spherical surface and on the unit spherical surface. Numerical results show that the algorithm performs well.

11.1 Background

Uniform design has been applied to various fields since it was proposed in Fang [4], Wang and Fang [22]. As its name implies, the basic idea of a uniform design is to seek design points scattered uniformly on certain domain. So in general, the combinatorial structure of a uniform design (or a low-discrepancy design) is quite arbitrary, which is different from that of an orthogonal array. To evaluate uniformity of a design, one must need a criterion, named discrepancy in uniform design theory. In fact, the concept of discrepancy came from number theory (quasi-Monte Carlo) method. As indicated in Fang and Wang [7], Fang et al. [6], many discrepancies including star discrepancy, L_p-discrepancy, and modified discrepancies proposed in Hickernell [10, 11] have their clear geometrical meanings. Uniform designs based on various discrepancies have been investigated extensively in existing literatures. Many properties and construction methods related to uniform design can be found in Fang et al. [5, 6]. To make it easier, most discrepancies are defined on the unit cube $C^s = [0, 1)^s$, but in some real problems, the experimental domain may be quite complicated. For example, Chuang and Hung [1] proposed the centered composite

Y. Tang (✉)
School of Mathematical Science, Soochow University, Jiangsu, China
e-mail: ytang@suda.edu.cn

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_11

discrepancy for a general domain, and Lin et al. [14] used it to construct uniform designs on the flexible region

$$R_m = \{(x_1, \ldots, x_s) \in [-1, 1]^s \big| |x_1|^m + |x_2|^m + \cdots + |x_s|^m \leq 1\},$$

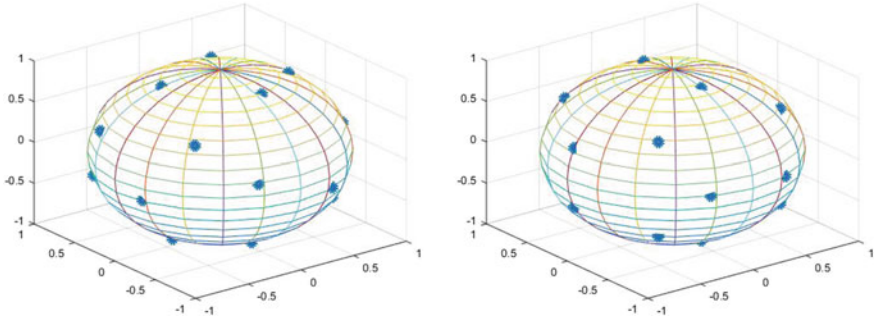
which was considered in Draper and Guttman [2]. The centered composite discrepancy can be regarded as a generalization of the centered L_2-discrepancy. Although it has no close form for an arbitrary domain, it can still be numerically calculated using efficient algorithms, just showed in Lin et al. [14]. Liu and Liu [15] considered uniform designs for mixture experiments with complex constraints, which could also be regarded as uniform designs on irregular domain. However, in some scenarios, the experimental domain will be even special. For example, in aerospace and military fields, people often want to scatter points on certain manifolds [3, 9]. To solve such problems, we need to further generalize the definition of discrepancy. Fang and Wang [7] suggests using inverse transformation to construct uniform design for some symmetrical domain including the unit spherical surface. For simplicity, throughout the paper, we only consider the three-dimensional case. Denote the unit spherical surface

$$\mathscr{U}^3 = \{(z_1, z_2, z_3) : z_1^2 + z_2^2 + z_3^2 = 1\}.$$

Let $\mathscr{P} = \{x^{(k)} = (x_{k1}, x_{k2}), k = 1, \ldots, n\}$ be a set of n points uniformly distributed on $\mathscr{C}^2 = [0, 1)^2$. Consider a transformation from $\mathscr{C}^2$ to $\mathscr{U}^3$ defined as

$$\begin{cases} z_{k1} = 1 - 2x_{k1}, \\ z_{k2} = 2\sqrt{x_{k1}(1 - x_{k1})}\cos(2\pi x_{k2}), \\ z_{k3} = 2\sqrt{x_{k1}(1 - x_{k1})}\sin(2\pi x_{k2}), \end{cases}$$

where $k = 1, 2, \ldots, n$. Although it can be proved that the resultant point set $\{z^{(k)} = (z_{k1}, \ldots, z_{ks}), k = 1, \ldots, n\}$ is uniformly scattered on the unit spherical surface $\mathscr{U}^s$, the actual result of the inverse transformation method does not seem effective, especially when the number of points is small. As Fang and Wang [7] pointed out, the indirect method using inverse transformation may not measure the uniformity of designs accurately. Figure 11.1 will illustrate it. The left part (a) in Fig. 11.1 indicates 20 points on the unit spherical surface obtained using the inverse transformation, while the right part (b) shows the 20 vertices of a regular dodecahedron, whose circumscribed sphere is the unit spherical surface. Intuitively, the latter seems more uniform compared with the former. It is well-known that there are only five different types of regular polyhedrons. Thus when the number of points is other than four, six, eight, twelve and twenty, we need consider other construction methods. Moreover, we will also show that even when a regular polyhedron exists, it will not always be the best one given some specific criterion. The paper is organized as follows. We will first define a general discrepancy based on geodesic distance for uniform design on manifold in Sect. 11.2. In Sects. 11.3 and 11.4, we consider uniform designs on the unit semi-spherical surface and the unit spherical surface,



(a). Transferred from $\mathscr{C}^2$ (b). Regular dodecahedron

Fig. 11.1 Two methods to construct uniform designs on $\mathscr{U}^3$

respectively. We also propose an algorithm to construct uniform designs for these two cases. Numerical examples will show that the algorithm is quite effective. Finally, we will give some conclusion and discussion in the last section.

11.2 General Discrepancy on Manifold

The concept of discrepancy arises in the Quasi-Monte Carlo method, which is used to solve multivariate integration problem. In many cases, we want to obtain the integration of certain function $f(x)$ over a specific domain $\mathscr{D}$. However, since the function $f(x)$ may be much complicated and we cannot get the exact value of the integration

$$I(f) \equiv \int_{\mathscr{D}} f(x)\, dx,$$

we will sometimes use the approximation to evaluate $I(f)$. A simple and easy way of doing so is to choose a set of n-point, P, which is uniformly scattered on the domain $\mathscr{D}$, and calculate all the values of $f(x)$ on these points, sum up them all and divide by n, i.e.,

$$Q(f; P) \equiv \frac{1}{n} \sum_{z \in P} f(z).$$

The approximation part $Q(f; P)$ is often called quadrature rule. Obviously, different set of points P may result in different quadrature rule. So we need to define a criterion to evaluate the uniformity of the point set P. As discussed in the previous section, some modified L_2-discrepancies, including the centered L_2-discrepancy and the wrap-around L_2-discrepancy, are often used in practice. However, most of them are defined on the unit cube and cannot be directly used when the experimental domain is a manifold. In Li [13], the author proposed the λ-discrepancy for uniform

design on a general domain. In this paper, we will generalize the λ-discrepancy, in order to let it be suitable for uniform design on manifold.

Definition 11.1 Let $\mathscr{D}$ be a domain and $\partial\mathscr{D}$ be its boundary. For any point $z \in \mathscr{D}$, define

$$\mathscr{B}_z = \bigvee_{x \in \partial\mathscr{D}} \boldsymbol{d}_B(z, x), \tag{11.1}$$

where $\boldsymbol{d}_B(\cdot,\cdot)$ is a well-defined distance function and the notation "$\bigvee$" represents an overall function (such as summation, integral, maximum or minimum), then $\mathscr{B}_z$ is called the **boundary deviation of** z.

Definition 11.2 Let $\mathscr{D}$ be a domain, and $\mathscr{Z} = \{z^{(1)}, z^{(2)}, \ldots, z^{(n)}\}$ be an n-point set, where each $z^{(i)} \in \mathscr{D}$. For any point $z \in \mathscr{D}$, define

$$\mathscr{P}_z = \bigvee_{z^{(i)} \in \mathscr{Z}} \boldsymbol{d}_P(z, z^{(i)}), \tag{11.2}$$

where $\boldsymbol{d}_P(\cdot,\cdot)$ also represents a well-defined distance function and "$\bigvee$" represents an overall function (such as summation, integral, maximum or minimum), then $\mathscr{P}_z$ is called the **point deviation of** z.

Definition 11.3 Let $\mathscr{D}$ be a domain, and $\mathscr{Z} = \{z^{(1)}, z^{(2)}, \ldots, z^{(n)}\}$ be an n-point set, where each $z^{(i)} \in \mathscr{D}$. Define

$$\mathscr{M}_\lambda(\mathscr{Z}) = C + \bigvee_{z^{(i)} \in \mathscr{Z}} \mathscr{B}_{z^{(i)}} + \lambda \bigvee_{z^{(i)} \in \mathscr{Z}} \mathscr{P}_{z^{(i)}}, \tag{11.3}$$

where C is a constant, λ is a positive parameter and "$\bigvee$" represents an overall function (such as summation, integral, maximum or minimum), then $\mathscr{M}_\lambda(\mathscr{Z})$ is called the **λ*e*-discrepancy of set** $\mathscr{Z}$.

Definition 11.4 Let $\mathscr{D}$ be a domain, and $\mathscr{Z} = \{z^{(1)}, z^{(2)}, \ldots, z^{(n)}\}$ be an n-point set, where each $z^{(i)} \in \mathscr{D}$. If $\mathscr{M}_\lambda(\mathscr{Z})$ can achieve the best value (minimum or maximum with respect to the choices of the overall function "$\bigvee$") over $\mathscr{D}$, then $\mathscr{Z}$ will be called a unform design on $\mathscr{D}$ under the λ-discrepancy.

Remark 11.1 the two distance function $\boldsymbol{d}_B(\cdot,\cdot)$ and $\boldsymbol{d}_P(\cdot,\cdot)$ in the above definitions can be different. Obviously, uniform designs under the λ-discrepancy may vary from different choices of the two distance functions $\boldsymbol{d}_B(\cdot,\cdot)$ and $\boldsymbol{d}_P(\cdot,\cdot)$. However, for uniform design on manifold, the natural selection of $\boldsymbol{d}_B(\cdot,\cdot)$ and $\boldsymbol{d}_P(\cdot,\cdot)$ is the geodesic distance function. Throughout this paper, we will take both of these two functions as the same geodesic distance function defined on the unit spherical surface.

Remark 11.2 different overall function "$\bigvee$" can be chosen for different purpose. Hickernell [10, 11] defined many modified discrepancies, i.e, overall functions, based

on different types of kernels, i.e, distance functions, on the unit cube. However, these discrepancies may not be suitable for uniform designs on manifolds. Throughout this paper, we will follow the idea of Johnson et al. [12], and use the maximin criterion to define the overall function. That is to say, we will first define the overall boundary deviation as the minimum distance over the domain boundary, and define the overall point deviation as the minimum distance among all distinct (otherwise the overall point deviation shall always be zero) pairwise points in the point set $\mathscr{Z}$. Then we try to maximize these two overall functions.

The λ-discrepancy in Definition 11.2 tries to balance the both the boundary effect and the point effect simultaneously, and thus has clear geometrical meanings. For clarity, we will divide into two cases to investigate uniform designs on spherical surface and on semi-spherical surface, respectively.

11.3 Uniform Design on Semi-spherical Surface

In this section, we will use the λ-discrepancy defined in the previous section as the measure of uniformity to consider uniform designs on the unit semi-spherical surface. To make it clear, throughout this section, the unit semi-spherical surface considered will always be assumed to be the above one, i.e, the point set of the domain is

$$\mathscr{U}_+^3 = \{(z_1, z_2, z_3) : z_3 \geq 0 \text{ and } z_1^2 + z_2^2 + z_3^2 = 1\}.$$

So the boundary of the domain is a circle:

$$\partial\mathscr{U}_+^3 = \{(z_1, z_2, z_3) : z_3 = 0 \text{ and } z_1^2 + z_2^2 + z_3^2 = 1\}.$$

As stated in the previous section, here we use maximin criterion [12] to measure the uniformity of the design. That is to say, the λ-discrepancy in (11.3) becomes the following form:

$$\mathscr{M}_\lambda(\mathscr{Z}) = C + \min_{z^{(i)}\in\mathscr{Z}} \min_{z^{(j)}\in\partial\mathscr{U}_+^3} \boldsymbol{d}_B(z^{(i)}, z^{(j)}) + \lambda \min_{z^{(i)}\in\mathscr{Z}} \min_{z^{(j)}\neq z^{(i)}} \boldsymbol{d}_P(z^{(i)}, z^{(j)}). \quad (11.4)$$

Here the distance functions $\boldsymbol{d}_B(\cdot,\cdot)$ and $\boldsymbol{d}_P(\cdot,\cdot)$ in (11.4) are both chosen to be the geodesic distance on the unit spherical surface. It is well-known that the geodesics on spherical surface are great circles.

Let $z^{(1)} = (x_1, x_2, x_3)$ and $z^{(2)} = (y_1, y_2, y_3)$ be two points on the unit spherical surface. Denote $\boldsymbol{d}_e(z^{(1)}, z^{(2)}) = \sqrt{((x_1 - y_1)^2 + (x_2 - y_2)^2 + (x_3 - y_3)^2)}$ as the Euclidean distance between $z^{(1)}$ and $z^{(2)}$. Then the geodesic distance between $z^{(1)}$ and $z^{(2)}$ is actually the length of arc $\overset{\frown}{z^{(1)}z^{(2)}}$ on the unit spherical surface, i.e,

$$\boldsymbol{d}_P(z^{(1)}, z^{(2)}) = \arccos\left[1 - 0.5\,\boldsymbol{d}_e^2(z^{(1)}, z^{(2)})\right].$$

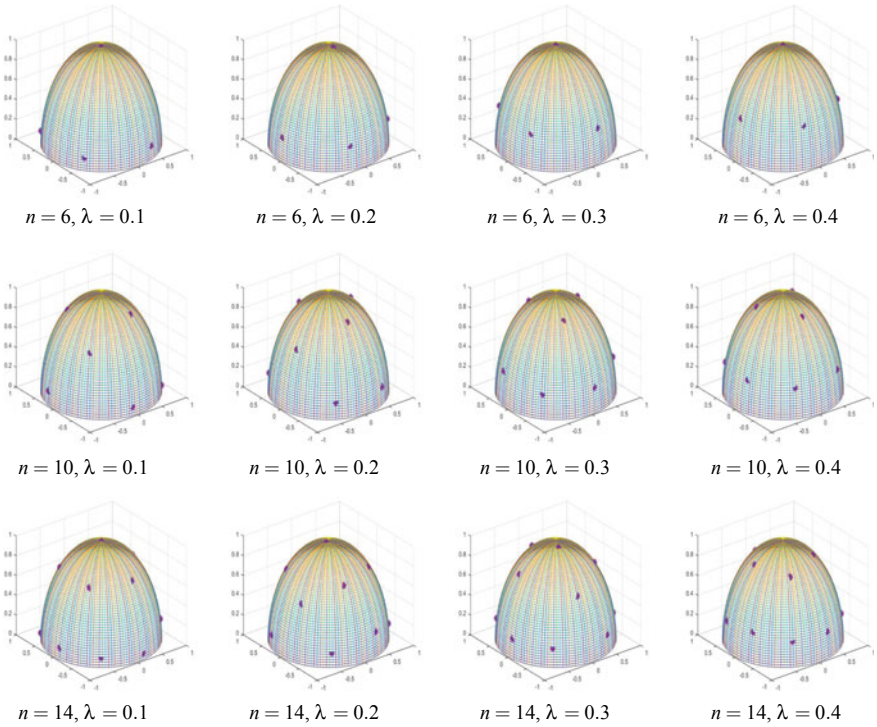


Fig. 11.2 Searching results for different number of points and λ's

Moreover, for any fixed point $\boldsymbol{z}^{(i)} = (x_1, x_2, x_3)$, its nearest point within the boundary $\partial\mathscr{U}_+^3$ will be $\boldsymbol{z}_0^{(i)} = (x_1', x_2', 0)$, where $x_1' = \frac{x_1}{\sqrt{x_1^2+x_2^2}}$ and $x_2' = \frac{x_2}{\sqrt{x_1^2+x_2^2}}$, thus

$$\min_{\boldsymbol{z}^{(j)} \in \partial\mathscr{U}_+^3} \boldsymbol{d}_B(\boldsymbol{z}^{(i)}, \boldsymbol{z}^{(j)}) = \boldsymbol{d}_P(\boldsymbol{z}^{(i)}, \boldsymbol{z}_0^{(i)}).$$

Now the objective function, i.e, the λ-discrepancy $\mathscr{M}_\lambda(\mathscr{Z})$ in (11.4), can be fully determined if an n-point set is given. So we can use a standard optimization algorithm to search for a design with less λ-discrepancy. The basic framework of the pseudo code is presented in Algorithm 5.

For the sake of simplicity, here we implement the algorithm on restricted lattice points. The candidates are equal distance grid points in the polar coordinate system. The searching results of Algorithm 5 are shown in Fig. 11.2. It can easily be seen that when the parameter λ becomes larger, the points tends to be scattered away from the boundary. This seems reasonable as we add more penalty to the boundary deviation when the points are near the boundary. Such a flexible solution can provide an alternative way to control the experimental points according to specific requirements in different real applications.

Algorithm 5 Pseudo code for prototype local search heuristic.

1: Initialize λ and τ (number of iterations)
2: Generate a starting design $\mathscr{Z}^c$ and let $\mathscr{Z}^{\max} := \mathscr{Z}^c$
3: **while** number of iterations $< \tau$ **do**
4: Generate $\mathscr{Z}^{\text{new}} \in \mathscr{N}(\mathscr{Z}^c)$ (neighbor to current solution)
5: Compute $\nabla = \mathscr{M}_\lambda(\mathscr{Z}^{\text{new}}) - \mathscr{M}_\lambda(\mathscr{Z}^c)$ and generate u (uniform random variable)
6: **if** $(\nabla > 0)$ or acceptance criterion(∇, u) met **then** $\mathscr{Z}^c = \mathscr{Z}^{\text{new}}$
7: **if** $\mathscr{Z}^c > \mathscr{Z}^{\max}$ **then** $\mathscr{Z}^{\max} := \mathscr{Z}^c$
8: **end while**

11.4 Uniform Design on Spherical Surface

Now we will consider uniform designs on the unit spherical surface. Similar with the case in the previous section, we also take the maximin criterion to define the λ-discrepancy. Since the unit spherical surface has no boundary, the λ-discrepancy of (11.3) will be equivalent with the following quantity:

$$\mathscr{M}_\lambda(\mathscr{Z}) = \min_{z^{(i)} \in \mathscr{Z}} \min_{z^{(j)} \neq z^{(i)}} \boldsymbol{d}_P(z^{(i)}, z^{(j)}), \tag{11.5}$$

where the distance function $\boldsymbol{d}_p(\cdot, \cdot)$ also represents the geodesic distance on the unit spherical surface.

Spherical trigonometry theory [21] tells us that there are many existing properties related to angles, sides and areas of spherical triangles and other configurations. These properties can not only help calculate the λ-discrepancies during our searching for uniform designs on the unit spherical surface numerically, but also provide upper bounds. In fact, Tammes [20] firstly considered the problem of arranging n points on a unit sphere which maximizes the minimum distance between any two distinct points. It is not an easy task to determine the best arrangement of the Tammes problem for some sporadic numbers of points, let alone to provide a systematical solution. For example, Musin and Tarasov [16, 17] provided final solutions to the Tammes problem when the numbers of points are thirteen and fourteen, respectively. The current paper does not study the Tammes problem theoretically and only aims to give an algorithmic solution to it. To evaluate our searching results, here we present some upper bounds of the λ-discrepancies.

Theorem 11.1 *Let $\mathscr{Z}$ be n points on the unit spherical surface $\mathscr{U}^3$, and $\mathscr{M}_\lambda(\mathscr{Z})$ be its λ-discrepancy as defined in (11.5). Then we have*

$$\mathscr{M}_\lambda(\mathscr{Z}) \leq 4 \arcsin(\sqrt{1/n}). \tag{11.6}$$

Proof Denote $r = \mathscr{M}_\lambda(\mathscr{Z})/2$. Since $0 \leq \mathscr{M}_\lambda(\mathscr{Z}) \leq 2\pi$, we have $0 \leq r \leq \pi$. For each point $z \in \mathscr{Z}$ on the unit spherical surface, define a set $\Omega_z = \{x \in \mathscr{U}^3 : \boldsymbol{d}_p(x, z) \leq r\}$. Easy to see, all points of Ω_z form a spherical crown. Thus the area of Ω_z is $A_z = 4\pi \sin^2(r/2)$. Sum up the area of all these spherical crowns, we

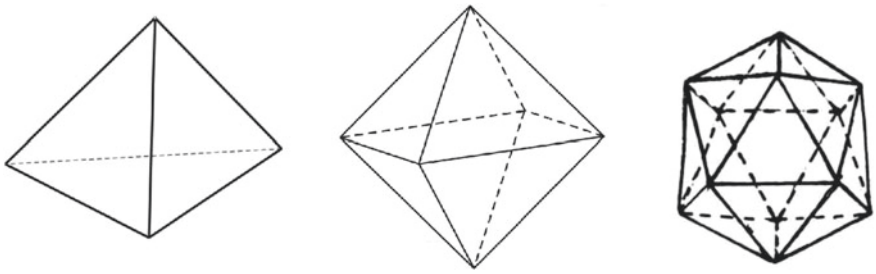


Fig. 11.3 Triangulations of regular polyhedrons with four, six and twelve points

have $nA_z \leq 4\pi$, where 4π represents the total area of the unit spherical surface. So $n \leq 1/\sin^2(r/2)$, i.e, $\mathscr{M}_\lambda(\mathscr{Z}) = 2r \leq 4\arcsin(\sqrt{1/n})$. □

The proof of Theorem 11.1 is quite intuitive, however, the bound of (11.6) can be further improved. As a matter of fact, many authors have provided upper bounds for the Tammes problem using graph theory, convex optimization and other techniques. Specifically, the following result was stated in Fejes-Tath [8].

Theorem 11.2 *Let $\mathscr{Z}$ be n points on the unit spherical surface, and $\mathscr{M}_\lambda(\mathscr{Z})$ be its λ-discrepancy as defined in (11.5). Then we have*

$$\mathscr{M}_\lambda(\mathscr{Z}) \leq \arccos[(cot^2\omega - 1)/2], \tag{11.7}$$

where $\omega = \frac{n}{n-2} \cdot \frac{\pi}{6}$.

The proof of Theorem 11.2 is not straightforward. Here we only give some explanation. The right hand side of (11.7) is the side length of an equilateral spherical triangle of area $\frac{4\pi}{(2n-4)}$, where 4π means the total area of the unit spherical surface and $2n-4$ represents the number of triangular faces induced by the n points. Theorem 11.2 says that when all the $2n-4$ triangular faces are equilateral spherical triangles with the same side length, the λ-discrepancy will achieve the upper bound (11.7). As it has been pointed out in many existing papers, when the number of points are four, six and twelve, the λ-discrepancies of the regular polyhedrons reach the upper bound (11.7). The triangulations of these regular polyhedrons are illustrated in Fig. 11.3, respectively.

Implement similar pseudo code as that of Algorithm 5, we can obtain a series of uniform designs on the unit spherical surface. Table 11.1 shows the numerical results for designs with different number of points. Notice that the values in the second and the third columns of Table 11.1 represent the λ-discrepancies of the resultant designs obtained by the inverse transformation from $\mathscr{C}^2$ and by implementing Algorithm 5, respectively. Easy to see, the algorithmic approach should be recommended. In fact, when the number of points are eight and twenty, the λ-discrepancies of the resultant

Table 11.1 Searching results of uniform designs on the unit spherical surface

#points	Transform	**Algorithm**	Bound (11.6)	Bound (11.7)	Polyhedron
4	1.633435	**1.908769**	2.094395	1.910633	1.910633
6	1.094689	**1.566865**	1.682137	1.570796	1.570796
8	0.822482	**1.299689**	1.445468	1.359080	1.230959
10	0.740723	**1.146458**	1.287002	1.214081	–
12	0.548902	**1.092115**	1.171371	1.107149	1.107149
14	0.525865	**0.950100**	1.082199	1.024176	–
16	0.461221	**0.888048**	1.010721	0.957398	–
18	0.358579	**0.834576**	0.951765	0.902163	–
20	0.434371	**0.795415**	0.902054	0.855491	0.7297277

designs are even better than those of polyhedrons. The fourth and the fifth columns of Table 11.1 list the upper bounds in (11.6) and (11.7), respectively. Compared with the former, the latter shall be much better.

11.5 Conclusion and Discussion

In this paper, we introduce a general definition of discrepancy based on geodesic distance to measure the uniformity of designs on manifold. We provide an algorithmic approach of a unified framework to search for low-discrepancy designs on the unit semi-spherical surface as well as on the unit spherical surface. Numerical results show the effectiveness of our proposed algorithm.

Some issues reported in this paper can be further investigated. For example, here we use the maximin criterion to define the overall function for the λ-discrepancy (11.3). However, using geodesic distance on specific manifold, criteria including minimax, mean squared-error [18] and entropy [19] can also be defined as the objective functions. Moreover, during the implementation of the algorithm, we restrict the candidates within equal distance grid points in the polar coordinate system. Such an approach may reduce the calculation burden, it can also bring negative effect on the λ-discrepancies of the designs.

Acknowledgements The author would like to thank the referees for their valuable and helpful comments. This research is supported by Natural Science Foundation of China (11671290) and Jiangsu Provincial Key Subject on Statistics (GD10700118).

References

1. Chuang, S.C., Hung, Y.C.: Uniform design over general input domains with applications to target region estimation in computer experiments. Comput. Stat. Data Anal. **54**(1), 219–232 (2010)
2. Draper, N.R., Guttman, I.: Response surface designs in flexible regions. J. Am. Stat. Assoc. **81**, 1089–1094 (1986)
3. Fan, H., Gao, X., Wang, Y.: Research on group inter-ception strategy of multiple kill vehicle (in Chinese). Syst. Eng. Electron. **32**(8), 1700–1702 (2010)
4. Fang, K.T.: The uniform design: application of number-theoretic methods in experimental design. Acta Math. Appl. Sin. **3**, 363–372 (1980)
5. Fang, K.T., Li, R., Sudjianto, A.: Design and Moodeling for Computer Experiments. Chapman and Hall/CRC, New York (2006)
6. Fang, K.T., Liu, M.Q., Qin, H., Zhou, Y.D.: Theory and Application of Uniform Experimental Designs. Science Press and Springer (2018)
7. Fang, K.T., Wang, Y.: Number-Theoretic Methods in Statistics. Chapman and Hall, London (1994)
8. Fejes-Tath, L.: On the densest packing of spherical caps. Am. Math. Monthly **56**, 330–331 (1949)
9. Fredrickson, S., Duran, S., Mitchell, J.: Mini AERCam Inspection Robot for Human Space Missions. In: Space 2004 Conference and Exhibit, California, USA, 28–30 Sept (2004)
10. Hickernell, F.J.: A generalized discrepancy and quadrature error bound. Math. Comput. **67**, 299–322 (1998)
11. Hickernell, F.J.: Lattice rules: how well do they measure up? In: Hellekalek, P., Larcher, G. (eds.) Random and Quasi-Random Point Sets. Lecture Notes in Statistics, vol. 138, pp. 109–166. Springer, New York (1998)
12. Johnson, M.E., Noore, L.M., Ylvisaker, D.: Minimax and maximin distance designs. J. Stat. Plann. Inference **26**, 131–148 (1990)
13. Li, J.: Uniform Designs on Symmetrical Domains and Their Constructions (in Chinese). Master Thesis, Soochow University (2011)
14. Lin, D.K.J., Sharpe, C., Winker, P.: Optimized U-type designs on flexible regions. Comput. Stat. Data Anal. **54**(6), 1505–1515 (2010)
15. Liu, Y., Liu, M.: Construction of uniform designs for mixture experiments with complex constraints. Commun. Stat. Theory Methods **45**, 2172–2180 (2016)
16. Musin, O.R., Tarasov, A.S.: The strong thirteen spheres problem. Discret. Comput. Geom. **48**, 128–141 (2012)
17. Musin, O.R., Tarasov, A.S.: The Tammes problem for $N = 14$. Exp. Math. **24**(4), 460–468 (2015)
18. Sacks, J., Schiller, S.B.: Spatial designs. In: Gupta, S.S., Berger, J.O. (eds.) Statistical Decision Theory and Related Topics IV, pp. 385–399. Springer Verlag, New York (1988)
19. Shewry, M.C., Wynn, H.P.: Maximum entropy sampling. J. Appl. Stat. **14**, 165–170 (1987)
20. Tammes, R.M.L.: On the origin of number and arrangement of places of exit on the surface of pallen grains. Rec. Trav. Bot. Neerl. **27**, 1–84 (1930)
21. Todhunter, I.: Spherical Trigonometry (5th ed.). MacMillan (1886)
22. Wang, Y., Fang, K.T.: A note on uniform distribution and experimental design. Chin. Sci. Bull. **26**, 485–489 (1981)

Part III
Multivariate Analysis

Chapter 12
An Application of the Theory of Spherical Distributions in Multiple Mean Comparison

Jiajuan Liang, Man-Lai Tang, Jing Yang, and Xuejing Zhao

Abstract Multiple normal mean comparison without the equal-variance assumption is frequently encountered in medical and biological problems. Classical analysis of variance (ANOVA) requires the assumption of equal variances across groups. When variations across groups are found to be different, classical ANOVA method is essentially inapplicable for multiple mean comparison. Although various approximation methods have been proposed to solve the classical Behrens-Fisher problem, there exists computational complexity in approximating the null distributions of the proposed tests. In this paper we employ the theory of spherical distributions to construct a class of exact F-tests and a simple generalized F-test for multiple mean comparison. The methods in this paper actually provide a simple exact solution and a simple approximate solution to the classical Behrens-Fisher problem in the case of balanced sample designs. A simple Monte Carlo study shows that the recommended tests have fairly good power performance. An analysis on a real medical dataset illustrates the application of the new methods in medicine.

J. Liang (✉)
University of New Haven, West Haven, CT, USA
e-mail: jliang@newhaven.edu

M.-L. Tang
Hang Seng Management College, Hong Kong, China
e-mail: mltang@hsu.edu.hk

J. Yang
Tianjin Medical University, Tianjin, China
e-mail: yangjingmath@163.com

X. Zhao
Lanzhou University, Lanzhou, China
e-mail: zhaoxj@lzu.edu.cn

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_12

12.1 Introduction

The theory of spherical distributions and spherical matrix distributions was comprehensively studied by Fang et al. [7], and Fang and Zhang [6]. Many applications based on the theory of spherical distributions have been developed since 1990. Simply speaking, the family of spherical distributions consists of those continuous multivariate distributions that possess the property of orthogonal rotation-invariance. The stochastic representation method is usually employed to characterize the family of spherical distributions. Let

$$\mathscr{S}_p(\phi) = \{\mathbf{x} : \ \boldsymbol{\Gamma} x \stackrel{\mathrm{d}}{=} x \text{ for any constant } p \times p \text{ orthogonal matrix } \boldsymbol{\Gamma}\}, \quad (12.1)$$

where $\phi(\cdot)$ stands for the characteristic function of a distribution. $\mathscr{S}_p(\phi)$ is called the family of spherically symmetric distributions or simply called spherical distributions. It is obvious that $\mathscr{S}_p(\phi)$ includes that the standard normal distribution $N_p(\mathbf{0}, \boldsymbol{I})$ and some commonly known multivariate distributions such as the multivariate Student t-distribution with zero mean and identity covariance matrix. It is known that $x \in \mathscr{S}_p(\phi)$ if and only if

$$\mathbf{x} \stackrel{\mathrm{d}}{=} R\mathscr{U}^{(p)}, \quad (12.2)$$

where $\mathscr{U}^{(p)}$ stands for the uniform distribution on the surface of the unit sphere in R^p (the p-dimensional real space), that is, $\|\mathscr{U}^{(p)}\| = 1$ ($\|\cdot\|$ stands for the usual Euclidean norm), and $R > 0$ is a random variable that is independent of $\mathscr{U}^{(p)}$. Equation (12.2) is called the stochastic representation for a spherical distribution. For any $\mathbf{x} \in \mathscr{S}_p(\phi)$ with $P(\mathbf{x} = \mathbf{0}) = 0$, it is always true that

$$\mathbf{x} \stackrel{\mathrm{d}}{=} \|\mathbf{x}\| \cdot \frac{\mathbf{x}}{\|\mathbf{x}\|}, \quad (12.3)$$

where $\|\mathbf{x}\|$ and $\mathbf{x}/\|\mathbf{x}\|$ are independent, and $\mathbf{x}/\|\mathbf{x}\| \stackrel{\mathrm{d}}{=} \mathscr{U}^{(p)}$.

An $n \times p$ random matrix X is said to have a left-spherical matrix distribution, denote by $X \sim LS_{n\times p}(\phi)$, if for any constant orthogonal matrix $\boldsymbol{\Gamma}$ $(n \times n)$

$$\boldsymbol{\Gamma} X \stackrel{\mathrm{d}}{=} X. \quad (12.4)$$

It is known that $X \sim LS_{n\times p}(\phi)$ if and only if X has the stochastic representation [6]:

$$X \stackrel{\mathrm{d}}{=} UV, \quad (12.5)$$

where U $(n \times p)$ is independent of V $(p \times p)$ and $U \sim \mathscr{U}^{(n\times p)}$, which is uniformly distributed on the Stielfel manifold

$$\mathscr{Q}(n, p) = \{H_{n\times p} : H'H = I_p\}. \quad (12.6)$$

If $X = (x_1, \ldots, x_n)'$ $(n \times p)$ consists of i.i.d. observations from $N_p(0, \boldsymbol{\Sigma})$, then $X \sim LS_{n\times p}(\phi)$ and X has a stochastic representation (12.5). For any random matrix $D_{p\times q}$ $(q \leq p)$, which is a function of X in the quadratic form $D = f(X'X)$, it can be proved that $XD \sim LS_{n\times p}(\phi)$ [6]. So XD also has a stochastic representation similar to (12.5), say, $XD \overset{\mathrm{d}}{=} U\boldsymbol{A}$ and $U \sim \mathscr{U}^{(n\times q)}$ that is independent of $\boldsymbol{A}_{q\times q}$. As a result of this stochastic representation, any affine-invariant statistic $T(XD) \overset{\mathrm{d}}{=} T(U)$, whose distribution is uniquely determined no matter how to choose the quadratic function $D = f(X'X)$.

Some successful applications of the theory of spherical distributions and spherical matrix distributions have been developed by some international scholars. For example, Läuter [15], Läuter et al. [16, 17], and Glimm and Läuter [11] employed the major theory of spherical matrix distributions in Fang and Zhang [6] to developed a class of exact multivariate tests for normal statistical inference. These tests can be still effectively applicable under high dimension with a small sample size, which may be smaller than the dimension of sample data. The tests developed by Läuter and his associates provide exact solutions to multivariate normal mean comparisons under high dimension with a small sample size. These tests extend the traditional Hotelling's T^2-test to the multiple mean comparisons as in multivariate analysis of variance (so-called MANOVA) and general linear tests for regression coefficients in multivariate regression models. Their tests are still applicable with fair power performance even in the case that the sample size is smaller than the dimension of sample data, see Kropf et al. [13]. By using the theory of spherical distributions and spherical matrix distributions in Fang et al. [7], and Fang and Zhang [6], Fang and Liang and their collaborators developed a class of nonparametric tests for goodness-of-fit purpose, see, for example, Fang et al. [8, 9], Liang and Fang [20], Liang et al. [23–26]. Liang and Ng [21], Liang and Tang [22], Ai et al. [1], and Liang [18, 19].

The classical problem of multiple mean comparison came from comparing the difference between experimental effects called treatment effects. It belongs to the topic of analysis of variance (ANOVA). Among others, the classical two-sample t-test may be the easiest one for comparing two normal means with the equal-variance assumption. The problem of two-sample mean comparison with unequal means was long noticed as early as Welch [28]. The problem has been continuing to be challenging in the case of multiple mean comparison with possible unequal variances. Approximate solutions to the problem have been proposed, for example, the Turkey test [27], the Kramer test [12], the Wald test, the likelihood ratio test and the score test [3], and the Kruskal-Wallis one-way ANOVA by ranks [14]. Some exact solutions to the problem of multiple mean comparison with unequal variances were also proposed. But the null distributions of the proposed tests do not have simple analytical expressions for easy computation of the p-value. This makes them inconvenient for various applications, see, for example, the procedures reviewed and compared by Dudewicz et al. [5].

In this paper we propose a simple solution to the problem of multiple mean comparison without assuming equal variances by using the theory of spherical distributions in Fang et al. [7] and the spherical matrix distributions in Fang and Zhang [6].

The case of a balanced sample design is assumed. The new approach consists of a class of exact F-tests and a generalized F-test. Section 12.2 presents the theoretical details on the construction of the tests. Section 12.3 gives a Monte Carlo study on the performance of the tests and illustrates the application of the tests by using real medical data. Some concluding remarks are summarized in the last section.

12.2 Construction of the Exact F-tests and the Generalized F-test

Assume that there is a balanced sample design (with an equal sample size across the normal populations) to obtain i.i.d. samples $\{x_i = (x_{i1}, \ldots, x_{in})' : \ n \times 1, \ i = 1, \ldots, k\}$ from the normal distribution $N(\mu_i, \sigma_i^2)$ for each population $i = 1, \ldots, k$ $(k \geq 2)$. Here it is also assumed that samples from different populations $N(\mu_i, \sigma_i^2)$ and $N(\mu_j, \sigma_j^2)$ $(i \neq j)$ are independent. We want to test the hypothesis of multiple mean comparison:

$$\begin{aligned} H_0 &: \ \mu_1 = \ldots = \mu_k, \\ H_1 &: \ \text{at least two means differ.} \end{aligned} \tag{12.7}$$

Randomly selecting a population as population k, we construct the observation matrix

$$X = \begin{pmatrix} x_{11} - x_{k1} & x_{21} - x_{k1} & \ldots & x_{k-1,1} - x_{k1} \\ x_{12} - x_{k2} & x_{22} - x_{k2} & \ldots & x_{k-1,2} - x_{k2} \\ \vdots & \vdots & \vdots & \vdots \\ x_{1n} - x_{kn} & x_{2n} - x_{kn} & \ldots & x_{k-1,n} - x_{kn} \end{pmatrix} : \quad n \times (k-1). \tag{12.8}$$

Theorem 12.1 *Let the observation matrix X be given by (12.8). Define the following eigenvalue-eigenvector problem:*

$$\left(\frac{1}{n} X'X\right) \boldsymbol{d}_i = \lambda \boldsymbol{d}_i, \tag{12.9}$$

where $\boldsymbol{d}_i = (d_{i1}, \ldots, d_{i,k-1})'$ for $i = 1, \ldots, r$ with $r = \min(n, k-1) - 1$ being the number of positive eigenvalues $\lambda_1 > \ldots > \lambda_r > 0$ in (12.9). Define

$$\boldsymbol{z}_i = (z_{i1}, \ldots, z_{in})' = X\boldsymbol{d}_i, \ \bar{z}_i = \frac{1}{n}\sum_{j=1}^{n} z_{ij}, \ F_i = n(\bar{z}_i)^2 \Big/ \left[\frac{1}{n-1}\sum_{j=1}^{n}(z_{ij} - \bar{z}_i)^2\right] \tag{12.10}$$

for $i = 1, \ldots, r$. Under the null hypothesis (12.7), F_i has an exact F-distribution $F(1, n-1)$ for $i = 1, \ldots, r = \min(n, \ k-1) - 1$.

Proof Denote by $X = (x_1, \ldots, x_n)'$: $n \times (k-1)$. Under the null hypothesis (12.7), it is easy to verify that the vectors $x_1, \ldots, x_n$ are i.i.d. and have the normal distribution $N_{k-1}(\mathbf{0}, \boldsymbol{\Sigma})$ with

$$\boldsymbol{\Sigma} = \text{diag}(\sigma_1^2 + \sigma_k^2, \ldots, \sigma_{k-1}^2 + \sigma_k^2).$$

Then X has a matrix-normal distribution:

$$X \sim N_{n\times(k-1)}(0, \boldsymbol{I}_n \otimes \boldsymbol{\Sigma}),$$

where "$\otimes$" stands for the Kronecker product of matrices. It is also easy to verify that X has a left-spherical matrix distribution [6] satisfying $\boldsymbol{\Gamma} X \overset{\text{d}}{=} X$. Note that the vector $\mathbf{d}_i$ in (12.9) is a function of $X'X$, denote by $\mathbf{d}_i = f_i(X'X)$. By using the stochastic representation (12.4), we obtain

$$(\boldsymbol{\Gamma} X)d_i = (\boldsymbol{\Gamma} X) f_i[(\boldsymbol{\Gamma} X)'(\boldsymbol{\Gamma} X)] \overset{\text{d}}{=} X f_i(X'X) = X d_i. \tag{12.11}$$

This results in

$$\boldsymbol{\Gamma} z_i \overset{\text{d}}{=} z_i \tag{12.12}$$

for any given $n \times n$ constant orthogonal matrix $\boldsymbol{\Gamma}$. Therefore, each z_i in (12.10) has a spherical distribution. The F-type statistic F_i in (12.10) is location-scale invariant. According to Fang et al. [7],

$$F_i(z_i) \overset{\text{d}}{=} F_i(z_0) \sim F(1, n-1), \quad i = 1, \ldots, r = \min(n,\ k-1) - 1, \tag{12.13}$$

under the null hypothesis (12.7), where $z_0 \sim N_n(0, \boldsymbol{I}_n)$ stands for the n-dimensional standard normal. □

Each of the F_i-statistic given by (12.10) can be employed to test the hypothesis (12.7). For any given $i = 1, \ldots, r = \min(n,\ k-1) - 1$, reject the null hypothesis in (12.7) at a given level $0 < \alpha < 1$ for a large value of $F_i > F(1-\alpha; 1, n-1)$, which stands for the $100(1-\alpha)$-percentile of the traditional F-distribution $F(1, n-1)$. By using Theorem 3 in Liang and Tang [22], we can obtain the following corollary.

Corollary 12.1 *Let z_i and the F_i-statistics be given by (12.10) for $i = 1, \ldots, r = \min(n,\ k-1) - 1$. Define the GF-statistic by*

$$GF(Z) = \max_{1\le i\le r}\{F_i(z_i)\}, \quad Z = (z_1, \ldots, z_r)' :\ r \times n. \tag{12.14}$$

Under the null hypothesis (12.7), GF has an approximate (n is large) "generalized F-distribution" with the cumulative distribution function (c.d.f.) given by

$$F_g(x) = P(GF(Z) < x) \approx [F(x; 1, n-1)]^r, \quad x > 0, \tag{12.15}$$

where $F(x; 1, n-1)$ stands for the c.d.f of the F-distribution $F(1, n-1)$.

From this corollary, we can propose a test for hypothesis (12.7): a large value of $GF(Z)$ indicates the null hypothesis is not true. The p-value of the GF-test can be approximated by using (12.15). Now we have a class of exact F-test and a generalized F-test for multiple mean comparison (12.7) without assuming equal variances:

$$F_1(z_1), \ldots, F_r(z_r), \ GF(Z), \tag{12.16}$$

where $F_i(z_i)$ for $i = 1, \ldots, r = \min(n, \ k-1) - 1$ are given by (12.10) and $GF(Z)$ by (12.14). The following section gives a Monte Carlo study on the empirical performance of these tests.

12.3 A Monte Carlo Study and a Real Example

12.3.1 Empirical Power Performance

Experiment. We choose the following six designs. Designs 1–4 are for the cases of $n \geq k$ and designs 5–6 for $n \leq k$.

$$\begin{aligned}
&\text{Design 1}: (n,k) = (10,5), \quad (\sigma_1^2, \ldots, \sigma_5^2) = (1, 5^2, 10^2, 20^2, 50^2),\\
&\qquad \text{mean difference} = |\mu_{i+1} - \mu_i| = 2.5c \text{ with } c = 0, 1, 2, \ldots, 12\\
&\qquad \text{for } i = 1, 2, 3, 4;\\
&\text{Design 2}: (n,k) = (20,5), \quad (\sigma_1^2, \ldots, \sigma_5^2) = (1, 5^2, 10^2, 20^2, 50^2),\\
&\qquad \text{mean difference} = |\mu_{i+1} - \mu_i| = c = 0, 1, 2, \ldots, 12\\
&\qquad \text{for } i = 1, 2, 3, 4;\\
&\text{Design 3}: (n,k) = (10,10), \quad (\sigma_1^2, \ldots, \sigma_{10}^2) = (1, 10^2, \ldots, 90^2),\\
&\qquad \text{mean difference} = |\mu_{i+1} - \mu_i| = c = 0, 1, 2, \ldots, 15\\
&\qquad \text{for } i = 1, 2, \ldots, 9;\\
&\text{Design 4}: (n,k) = (20,10), \quad (\sigma_1^2, \ldots, \sigma_{10}^2) = (1, 10^2, \ldots, 90^2),\\
&\qquad \text{mean difference} = |\mu_{i+1} - \mu_i| = c = 0, 1, 2, \ldots, 15\\
&\qquad \text{for } i = 1, 2, \ldots, 9;\\
&\text{Design 5}: (n,k) = (10,20), \quad (\sigma_1^2, \ldots, \sigma_{20}^2) = (1, 10^2, 20^2, \ldots, 190^2),\\
&\qquad \text{mean difference} = |\mu_{i+1} - \mu_i| = c = 1.5 \times (0, 1, 2, \ldots, 15)\\
&\qquad \text{for } i = 1, 2, \ldots, 15;\\
&\text{Design 6}: (n,k) = (20,20), \quad (\sigma_1^2, \ldots, \sigma_{20}^2) = (1, 10^2, 20^2, \ldots, 190^2),\\
&\qquad \text{mean difference} = |\mu_{i+1} - \mu_i| = c = (0, 1, 2, \ldots, 15)\\
&\qquad \text{for } i = 1, 2, \ldots, 15.
\end{aligned} \tag{12.17}$$

The following Four statistics as given by (12.10) and (12.14) are chosen:

$$F_1, \ F_{r_1}, \ F_{r_2}, \ GF, \tag{12.18}$$

where $r_1 = [r/3]$ and $r_2 = [r/2]$ with $r = \min(n, k-1) - 1$ in the sample design (12.17), here the notation $[x]$ stands for the nearest integer not exceeding $x > 0$. That is, F_1 is the statistic constructed from the eigenvector associated with the largest eigenvalue in (12.9); F_{r_1} is the statistic constructed from the eigenvector associated with the r_1th largest eigenvalue in (12.9); F_{r_2} is the statistic constructed from the eigenvector associated with the r_2th largest eigenvalue in (12.9). The power performance at level 0.05 of these four tests is displayed in Figs. 12.1 and 12.2. The immediate observation is:

(1) the four tests can control their nominal levels very well;
(2) the generalized F-test GF performs the best in most cases;
(3) the F_1-test performs the best among all exact F-tests;
(4) F_{r_1} and F_{r_2} have increasing power at beginning but have decreasing power when the mean difference becomes bigger. This indicates that only the F_1-test has comparable power performance with that of the GF-test.

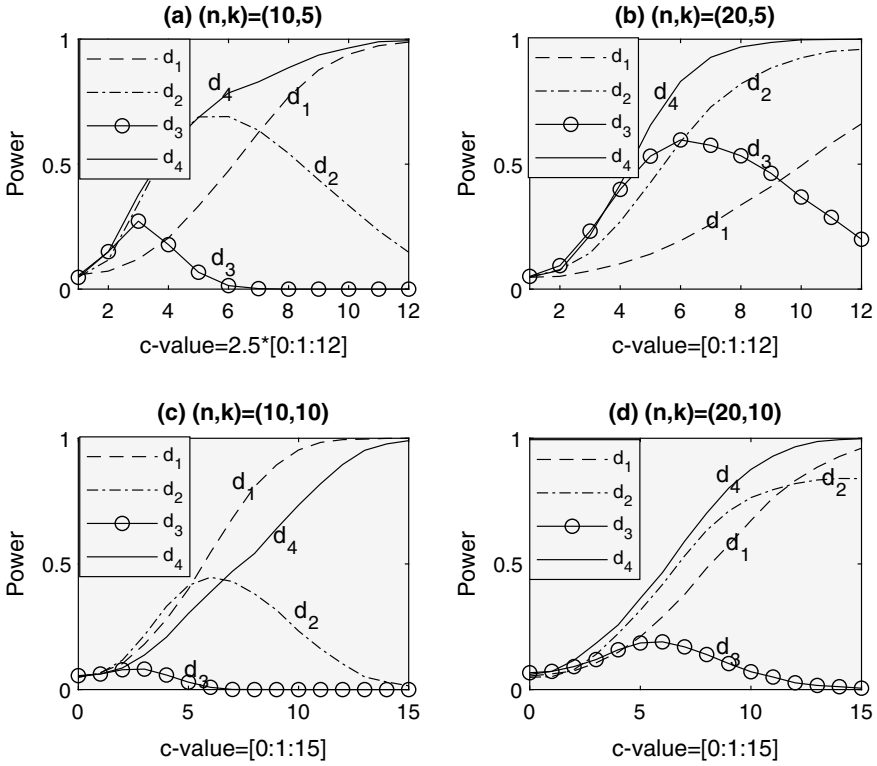


Fig. 12.1 Illustration of power performance of the exact F_1-tests and the generalized F-test: the direction d_1 is associated with largest eigenvalue; d_2 is associated with the $r_1 = [r/3]$th largest eigenvalue; d_3 is associated with the $r_2 = [r/2]$th largest eigenvalue (cases of $n \geq k$)

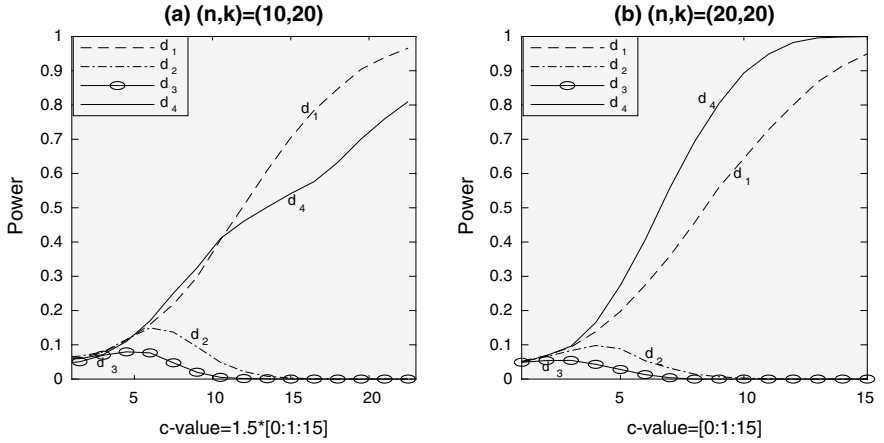


Fig. 12.2 Illustration of power performance of the exact F_1-tests and the generalized F-test: the direction d_1 is associated with largest eigenvalue; d_2 is associated with the $r_1 = [r/3]$th largest eigenvalue; d_3 is associated with the $r_2 = [r/2]$th largest eigenvalue (cases of $n \leq k$)

Based on the Monte Carlo simulation results, we recommend the F_1-test and the GF-test for hypothesis (12.7) for general multiple mean comparison without equal-variance assumption.

12.3.2 An Illustrative Application

A research project was carried out by Tianjin Medical University, China [10]. Rats were collected for experiment by different treatments to see the treatment effects. There are four different treatments. Each treatment consists of 46 levels with sample size $n = 6$. In the experiment on 6 rats, the ratio of organ wet weight to body weight (organ coefficient) was observed. The purpose is to evaluate organ development during the treatment. Details on the experiment and medical analysis can be found in Gao et al. [10]. In one-way ANOVA, we can consider each factor level as a group or population. In the experiment on 6 male rats with 46 levels, we consider if the ratio of organ wet weight to body weight has changed during the treatment. Let

$$\mu_i = \text{the average ratio of organ wet weight to body weight at level } i \qquad (12.19)$$

for $i = 1, \ldots, 46$. Then we need to test the hypothesis

$$\begin{aligned} H_0 &: \ \mu_1 = \ldots = \mu_{46}, \\ H_1 &: \ \text{at least two means differ.} \end{aligned} \qquad (12.20)$$

Note that the balanced sample size $n = 6$ with $k = 46$ groups or populations. We could apply both the traditional one-way ANOVA F-test $F(k-1, N-k) = F(45, 230)$, the $F_1 \sim F(1, n-1)$-test, and the GF-test to hypothesis (12.20). The one-way ANOVA F-test gives a p-value ≈ 0, the $F_1 \sim F(1, n-1)$-test gives a p-value $= 4.31 \times 10^{-5}$, the GF-test gives a p-value $= 2.16 \times 10^{-4}$, and the Turkey-Kramer pairwise approximate t-tests give results of all significantly different means in the pairwise comparisons. The Bartlett test [2] and the Brown-Forsythe test [4] are employed to test the homogeneity of variances of the 46 levels (groups) of the treatment factor. It turns out that the Bartlett test has a p-value $= 7.1702 \times 10^{-175}$ and the Brown-Forsythe test has a p-value $= 1.3229 \times 10^{-104}$, indicating very strong variance homoscedasticity. This implies that the traditional F-test from one-way ANOVA is essentially inapplicable.

12.4 Concluding Remarks

The exact F-tests and the generalized F-test in this paper are applicable for multiple mean comparisons without assuming homogeneity of variances across the populations. They provide an exact solution to the problem of multiple mean comparison with simple F-tests under a balanced sample design. Existing methods in the literature are facing computational complexity in computing the critical values or the p-values of the test statistics. The Monte Carlo study in Sect. 12.3 shows not all of the exact F-tests have desirable power performance. But F_1-test constructed from the eigenvector associated with the largest eigenvalue and the generalized F-test GF have fairly good power performance. They are recommended for general comparison of multiple means. Theorem 12.1 in Sect. 12.2 implies that the exact F-tests and the generalized F-test heavily depend on the normal assumption on the raw data. Although the robustness of the exact F-tests against a possible departure from the normal assumption is generally unknown, the proof of Theorem 12.1 shows that it only requires the observation matrix X defined by (12.8) to have a left-spherical matrix distribution. Therefore, the exact F-tests and the generalized F-test are robust in the distribution family of left-spherical matrix distributions, which includes the normal assumption as a special case. The methods in this paper actually provide a simple exact solution and a simple approximate solution to the classical Behrens-Fisher problem in the case of balanced sample designs. The exact F-tests and the generalized F-test with a balanced sample design in this paper can be generalized to the case of unbalanced sample designs. Our research is in progress and much stronger results on exact solutions to the general Behrens-Fisher problem will be obtained soon.

Acknowledgements The authors would like to thank Prof. Zengrong Sun and her research associates in Tianjin Medical University, China, for providing the real medical data in gene comparisons under different experimental conditions.

References

1. Ai, M., Liang, J., Tang, M.L.: Generalized T_3-plot for testing high-dimensional normality. Front. Math. China **11**, 1363–1378 (2016)
2. Bartlett, M.S.: Properties of sufficiency and statistical tests. Proc. Roy. Statist. Soc. (Ser. A) **160**, 268–282 (1937)
3. Best, D.J., Rayner, J.C.W.: Welch's approximate solution for the Behrens-Fisher problem. Technometrics **29**, 205–210 (1987)
4. Brown, M.B., Forsythe, A.B.: Robust tests for the equality of variances. J. Amer. Stat. Assoc. **69**, 364–367 (1974)
5. Dudewicz, E.J., Ma, Y., Mai, E., Su, H.: Exact solutions to the Behrens Fisher problem: asymptotically optimal and finite sample efficient choice among. J. Stat. Plann. Inference **137**, 1584–1605 (2007)
6. Fang, K.T., Zhang, Y.T.: Generalized Multivariate Analysis. Science Press and Springer, Beijing and Berlin (1990)
7. Fang, K.T., Kotz, S., Ng, K.W.: Symmetric Multivariate and Related Distributions. Chapman and Hall Ltd., London and New York (1990)
8. Fang, K.T., Li, R., Liang, J.: A multivariate version of Ghosh's T_3-plot to detect non-multinormality. Comput. Stat. Data Anal. **28**, 371–386 (1998)
9. Fang, K.T., Liang, J., Hickernell, F.J., Li, R.: A stabilized uniform Q-Q plot to detect non-multinormality. In: Hsiung, A.C., Ying, Z., Zhang, C.H. (eds.) Random Walk, Sequential Analysis and Related Topics, pp. 254–268. World Scientific, New Jersey (2007)
10. Gao, N., Hu, R., Huang, Y., Dao, L., Zhang, C., Liu, Y., Wu, L., Wang, X., Yin, W., Gore, A.C., Zengrong Sun, Z.: Specific effects of prenatal DEHP exposure on neuroendocrine gene expression in the developing hypothalamus of male rats. Arch. Toxicol. **92**, 501–512 (2018)
11. Glimm, E., Läuter, J.: On the admissibility of stable spherical multivariate tests. J. Multivar. Anal. **86**, 254–265 (2003)
12. Kramer, C.Y.: Extension of multiple range tests to group means with unequal numbers of replications. Biometrics **12**, 307–310 (1956)
13. Kropf, S., Läuter, J., Kosea, D., von Rosen, D.: Comparison of exact parametric tests for high-dimensional data. Comput. Stat. Data Anal. **53**, 776–787 (2009)
14. Kruskal, W.H., Wallis, W.A.: Use of ranks in one-criterion variance analysis. J. Amer. Stat. Assoc. **47**, 583–621 (1952)
15. Läuter, J.: Exact t and F tests for analyzing studies with multiple endpoints. Biometrics **52**, 964–970 (1996)
16. Läuter, J., Glimm, E., Kropf, S.: New multivariate tests for data with an inherent structure. Biomet. J. **38**, 5–23 (1996)
17. Läuter, J., Glimm, E., Kropf, S.: Multivariate tests based on left-spherically distributed linear scores. Ann. Stat. **26**, 1972–1988 (1998)
18. Liang, J.: Exact F-tests for a class of elliptically contoured distributions. J. Adv. Stat. **1**, 212–217 (2016)
19. Liang, J.: A generalized F-test for the mean of a class of elliptically contoured distributions. J. Adv. Stat. **2**, 10–15 (2017)
20. Liang, J., Fang, K.T.: Some applications of Läuter's technique in tests for spherical symmetry. Biometrical J. **42**(8), 923–936 (2000)
21. Liang, J., Ng, K.W.: A multivariate normal plot to detect non-normality. J. Comput. Graph. Stat. **18**, 52–72 (2009)
22. Liang, J., Tang, M.L.: Generalized F-tests for the multivariate normal mean. Comput. Stat. Data Anal. **57**, 1177–1190 (2009)
23. Liang, J., Fang, K.T., Hickernell, F.J.: Some necessary uniform tests for sphericcal symmetry. Ann. Inst. Stat. Math. **60**, 679–696 (2008)
24. Liang, J., Tang, M.L., Chan, P.S.: A generalized Shapiro-Wilk W statistic for testing high-dimensional normality. Comput. Stat. Data Anal. **53**, 3883–3891 (2009)

25. Liang, J., Li, R., Fang, H., Fang, K.T.: Testing multinormality based on low-dimensional projection. J. Stat. Plann. Inference **86**, 129–141 (2000)
26. Liang, J., Tang, M.L., Zhao, X.: Testing high-dimensional normality based on classical skewness and kurtosis with a possible small sample size. Commun. Stat. Theory Methods **48**(23), 5719–5732 (2019)
27. Turkey, J.W.: Comparing individual means in the analysis of variance. Biometrics **5**, 99–114 (1949)
28. Welch, B.L.: The generalization of 'Students' problem when several different population variances are involved. Biometrika **34**, 28–35 (1947)

Chapter 13
Estimating the Location Vector for Spherically Symmetric Distributions

Jian-Lun Xu

Abstract When a $p \times 1$ random vector $\mathbf{X}$ has a spherically symmetric distribution with the location vector $\boldsymbol{\theta}$, Brandwein and Strawderman [7] proved that estimators of the form $\mathbf{X} + a\mathbf{g}(\mathbf{X})$ dominate the $\mathbf{X}$ under quadratic loss if the following conditions hold: (i) $\|\mathbf{g}\|^2/2 \leq -h \leq -\nabla \circ \mathbf{g}$, where $-h$ is superharmonic, (ii) $E[-R^2 h(\mathbf{V})]$ is nondecreasing in R, where $\mathbf{V}$ has a uniform distribution in the sphere centered at $\boldsymbol{\theta}$ with a radius $R = \|\mathbf{X} - \boldsymbol{\theta}\|$, and (iii) $0 < a \leq 1/[pE(R^{-2})]$. In this paper we not only use a weaker condition than their (ii) to show the dominance of $\mathbf{X} + a\mathbf{g}(\mathbf{X})$ over the $\mathbf{X}$, but also obtain a new bound $E(R)/[pE(R^{-1})]$ for a, which is always better than bounds obtained by Brandwein and Strawderman [7] and Xu and Izmirlian [24]. The generalization to concave loss function is also considered. In addition, estimators of the location vector are investigated when the observation contains a residual vector and the scale is unknown.

13.1 Introduction

It is well-known that the normal distribution and its related statistical inference such as estimation of its mean are crucial in application. Ever since Stein [19] discovered the inadmissibility of the best invariant estimator of the p-dimensional ($p \geq 3$) normal mean under quadratic loss, there has been considerable interest in improving upon the best invariant estimator of a location vector by relaxing the normality assumption, studying more general estimators, or considering different loss functions.

Under the quadratic loss, James and Stein [15] presented a class of dominating estimators, $\left(1 - a/\|\mathbf{X}\|^2\right)\mathbf{X}$ for $0 < a < 2(p-2)$ if $\mathbf{X}$ has a normal distribution with the identity covariance matrix I_p. This result remains true if the distribution of $\mathbf{X}$ is spherically symmetric about its location vector and $p \geq 4$ as shown by Brandwein [2],

J.-L. Xu (✉)
Biometry Research Group, National Cancer Institute,
9609 Medical Center Drive, Room 5E128, Bethesda, MD 20892-9789, USA
e-mail: xujia@mail.nih.gov

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_13

Brandwein and Strawderman [3, 4, 7], Fan and Fang [12–14], Maruyama [16], and Brown and Zhao [9], Tosh and Dasgupta [21] and others; see the review articles by Brandwein and Strawderman [5, 6]. When the dimension is at least three, Brown [8] also proved that the best invariant estimator of a location vector is inadmissible for a wide class of distributions and loss functions. When the components of $\mathbf{X}$ are independent, identically and symmetrically (iis) distributed about their respective means, Shinozaki [18] studied the dominance conditions of the James-Stein type estimator

$$\boldsymbol{\delta}_{a,b}(\mathbf{X}) = \left(1 - \frac{a}{b + \|\mathbf{X}\|^2}\right)\mathbf{X}, \tag{13.1}$$

over $\mathbf{X}$ and obtained the bounds of a and b in (13.1) that depend on the second and fourth moments of the component distributions. Xu [23] investigated the bounds of a and b in (13.1) when $\mathbf{X}$ has a sign-invariant distribution.

For more general estimators and different loss functions, Miceli and Strawderman [17] restricted the distribution of $\mathbf{X}$ to the subclass of iis distributions called independent component variance mixtures of normals and replaced a in (13.1) by $ar(X_1^2, \ldots, X_p^2)$, where $r(X_1^2, \ldots, X_p^2)$ is a function of $X_1^2, \ldots, X_p^2$. Their loss function is nonquadratic. When $\mathbf{X}$ has a spherically symmetric distribution about its location vector $\boldsymbol{\theta}$ and loss function is a quadratic loss, a concave function of quadratic loss, or the general quadratic loss, Brandwein and Strawderman [7] elegantly used the *divergence* theorem to prove the dominance of the estimators

$$\boldsymbol{\delta}_{a,\mathbf{g}}(\mathbf{X}) = \mathbf{X} + a\mathbf{g}(\mathbf{X}). \tag{13.2}$$

over $\mathbf{X}$ under conditions (i) $\|\mathbf{g}\|^2/2 \le -h \le -\nabla \circ \mathbf{g}$, where $-h$ is superharmonic, (ii) $E[-R^2 h(\mathbf{V})]$ is nondecreasing in R, where $\mathbf{V}$ has a uniform distribution in the sphere centered at $\boldsymbol{\theta}$ with a radius $R = \|\mathbf{X} - \boldsymbol{\theta}\|$, and (iii) $0 < a \le 1/[pE(R^{-2})]$. Clearly, the estimators $\boldsymbol{\delta}_{a,\mathbf{g}}(\mathbf{X})$ given by (13.2), together with conditions (i) and (iii) extend the classical James-Stein estimator to a broader class of estimators, while their condition (ii) is a technical condition. Xu and Izmirlian [24] dropped their technical condition (ii) and obtained a bound $0 < a < [\mu_1/(p^2\mu_{-1})][1 - (p-1)\mu_1/(p\mu_{-1}\mu_2)]^{-1}$ for a, where $\mu_i = E(R^i)$ for $i = -1, 1, 2$. As stated by Xu and Izmirlian [24], their bound of a is sometimes worse than the bound obtained by Brandwein and Strawderman [7]. A question of theoretical interest is raised: Is this possible that bounds of a obtained by Brandwein and Strawderman [7] and Xu and Izmirlian [24] can be improved under a weaker condition than Brandwein and Strawderman's [7] technical condition (ii)? In this paper we provide an affirmative answer to this question. Specifically, we use the fact that the average of $-h$ over the sphere is nonincreasing in the radius to show dominance of $\boldsymbol{\delta}_{a,\mathbf{g}}(\mathbf{X})$ over $\mathbf{X}$ and obtain a new bound $0 < a \le \mu_1/(p\mu_{-1})$ for a, which is always better than $1/(p\mu_{-2})$ and $[\mu_1/(p^2\mu_{-1})][1 - (p-1)\mu_1/(p\mu_{-1}\mu_2)]^{-1}$.

The paper is organized as follows: In Sect. 13.2 we present the main result that states the dominance conditions of the estimators $\boldsymbol{\delta}_{a,\mathbf{g}}(\mathbf{X})$ with respect to the quadratic loss. To illustrate the construction of the function h and the performance of the new

bound, three examples are also studied in Sect. 13.2. In Sect. 13.3 we extend the main result in Sect. 13.2 to other loss functions that are nondecreasing concave functions of quadratic loss. The estimators of the location vector when the scale is unknown and the observation $(\mathbf{X}^T, \mathbf{Y}^T)^T$ contains a residual vector $\mathbf{Y}$ are also considered in Sect. 13.3. Section 13.4 is devoted to some concluding remarks, while the last section consists of proofs of results in Sects. 13.2 and 13.3.

13.2 Main Results

Let $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)^T$ be an estimator of $\boldsymbol{\theta}$ and let $R(\boldsymbol{\delta}, \boldsymbol{\theta}) = E[L(\boldsymbol{\delta}, \boldsymbol{\theta})]$ be the risk of $\boldsymbol{\delta}$, where the loss function $L(\boldsymbol{\delta}, \boldsymbol{\theta})$ is defined by

$$L(\boldsymbol{\delta}, \boldsymbol{\theta}) = \|\boldsymbol{\delta} - \boldsymbol{\theta}\|^2 = \sum_{i=1}^{p} (\delta_i - \theta_i)^2. \tag{13.3}$$

That is, the loss function $L(\boldsymbol{\delta}, \boldsymbol{\theta})$ we consider in this section is quadratic. Furthermore, we employ the following notation introduced by Xu and Izmirlian [24]:

$$\begin{aligned} m(t) &= -E_{\mathbf{U}}[h(t\mathbf{U} + \boldsymbol{\theta})], \\ M_*(t) &= M(t) - M(0) = \int_0^t m(z)dz \end{aligned} \tag{13.4}$$

for $t \geq 0$, where $-h$ is a nonnegative and superharmonic function and the random vector $\mathbf{U}$ has a uniform distribution on the surface of the unit sphere. Note that $m(t)$ is a nonincreasing function of t and $M_*(t)$ is a nonnegative and nondecreasing concave function of t (see Du Plessis [[11], p. 54]).

Theorem 13.1 *Suppose that* $\mathbf{X} \sim \mathrm{SS}_p(\boldsymbol{\theta}, I_p)$ *(spherically symmetric about mean vector* $\boldsymbol{\theta}$*) and* $\boldsymbol{\delta}_{a,\mathbf{g}}(\mathbf{X})$ *is defined by (13.2). Then under quadratic loss (13.3),* $\boldsymbol{\delta}_{a,\mathbf{g}}(\mathbf{X})$ *has a smaller risk than* $\boldsymbol{\delta}_{0,\mathbf{g}}(\mathbf{X}) = \mathbf{X}$ *if*

(i) $\|\mathbf{g}\|^2/2 \leq -h \leq -\nabla \circ \mathbf{g}$, *where* $-h$ *is superharmonic,*

(ii) $r \int_0^1 m(rz) p z^{p-1} dz \geq c \int_0^1 M_*(rz) p z^{p-2} dz$ *when* $r > \sqrt{ap}$, *where* m *and* M_* *are defined by (13.4) and* $1 \leq c \leq p - 1$ *is a constant, and*

(iii) $0 < a \leq \mu_1/(p\mu_{-1})$, *where* $\mu_{-i} = E(R^{-i})$ *for* $i = -1, 1$ *and* $R = \|\mathbf{X} - \boldsymbol{\theta}\|$.

Remark 13.1 The condition (ii) of Theorem 13.1 is slightly weaker than the condition (ii) of Brandwein and Strawderman [7]. To see this, we use integration by parts to obtain that

$$r \int_0^1 m(rz) p z^{p-1} dz = p \left(M_*(r) - \int_0^1 M_*(rz)(p-1) z^{p-2} dz \right).$$

Thus, the condition (ii) above is equivalent to $N(r) \geq 0$ when $r > \sqrt{ap}$, where $N(r)$ is defined by

$$N(r) = M_*(r) - (p - 1 - c) \int_0^1 M_*(rz) z^{p-2} dz.$$

Taking the derivative of $N(r)$ gives that

$$N'(r) = m(r) - (p - 1 - c) \int_0^1 m(rz) z^{p-1} dz. \tag{13.5}$$

Since the condition (ii) of Brandwein and Strawderman [7] is equivalent to

$$\int_0^1 m(rz) z^{p-1} dz \leq \frac{1}{p-2} m(r), \quad r > 0. \tag{13.6}$$

Applying (13.6) to (13.5) will yield that

$$N'(r) \geq m(r) - \frac{p - 1 - c}{p - 2} m(r) = \frac{c - 1}{p - 2} m(r) \geq 0$$

because $c \geq 1$. This shows that $N(r)$ is a nondecreasing function of r. Using the fact that $\lim_{r \to 0^+} N(r) = 0$, we can conclude that $N(r) \geq 0$ when $r > 0$.

It is also worth mentioning that we only require the condition (ii) to be true when $r > \sqrt{ap}$. When $r \leq \sqrt{ap}$, there is no any assumption.

Remark 13.2 Let F denote the distribution function (df) of $R = \|\mathbf{X} - \boldsymbol{\theta}\|$. Then applying Lemma 13.1 in Sect. 13.5 with $f_1(r) = r$, $g_1(r) = 1/r^2$, $f_2(r) = g_2(r) = 1$ and $d\alpha = dF$ yields that

$$\mu_{-1} = E\left(\frac{1}{R}\right) = E\left(R \frac{1}{R^2}\right) \leq E(R) E\left(\frac{1}{R^2}\right) = \mu_1 \mu_{-2}.$$

Using this fact, we can conclude that the new bound for a is better than that of Brandwein and Strawderman [7] because

$$\frac{1}{p\mu_{-2}} \leq \frac{\mu_1}{p\mu_{-1}}.$$

Remark 13.3 The new bound for a is also better than that of Xu and Izmirlian [24]. This can be seen from a direct comparison with the fact that $\mu_1 \leq \mu_{-1}\mu_2$, which follows from an application of Lemma 13.1 in Sect. 13.5 with $f_1(r) = 1/r$, $g_1(r) = r^2$, $f_2(r) = g_2(r) = 1$ and $d\alpha = dF$.

Remark 13.4 It needs to be mentioned that the requirement of dimensionality such as $p \geq 4$ usually arises in the condition (i) of Theorem 13.1. Meanwhile, although the

function h used in Theorem 13.1 has many choices, we usually take $h(\mathbf{X}) = \nabla \circ \mathbf{g}(\mathbf{X})$ when $\nabla \circ \mathbf{g}(\mathbf{X})$ is a subharmonic function.

Example 13.1 Consider the James-Stein [15] estimator which is given by

$$\delta_{a,0}(\mathbf{X}) = \left(1 - \frac{a}{\|\mathbf{X}\|^2}\right)\mathbf{X}$$

and discussed by many authors including Brandwein and Strawderman [7] and Fan and Fang [13]. Clearly, taking $\mathbf{g}(\mathbf{X}) = -\mathbf{X}/\|\mathbf{X}\|^2$ in (13.2) will see that $\delta_{a,0}(\mathbf{X})$ is a special case of estimators (13.2). Let

$$h(\mathbf{X}) = \nabla \circ \mathbf{g}(\mathbf{X}) = -\frac{p-2}{\|\mathbf{X}\|^2}.$$

Then $-h$ is superharmonic if $p \geq 4$ because

$$-\sum_{i=1}^{p} \frac{\partial^2 h}{\partial x_i^2} = -\frac{(p-2)(p-4)}{\|\mathbf{X}\|^2} \leq 0.$$

The condition (i) in Theorem 13.1 is clearly satisfied. Meanwhile, condition (ii) in Theorem 13.1 is also true because Brandwein and Strawderman's [7] technical condition (ii) is true, see Lemma 2.2 of Fan and Fang [13].

To illustrate the performance of the new bound of a, we consider two examples below. We use a_{new} to denote the new bound $\mu_1/(p\mu_{-1})$ of a. We denote by $a_{\text{bs}} = 1/(p\mu_{-2})$, the bound of a in Brandwein and Strawderman's [7] Theorem 2.1, and $a_{\text{xi}} = [\mu_1/(p^2\mu_{-1})][1 - (p-1)\mu_1/(p\mu_{-1}\mu_2)]^{-1}$, the bound of a in Xu and Izmirlian's [24] Theorem 1.

Example 13.2 Let $\mathbf{X}$ have a normal distribution with mean $\boldsymbol{\theta}$ and covariance matrix I_p. Then $R^2 = \|\mathbf{X} - \boldsymbol{\theta}\|^2$ has a χ_p^2-distribution, which implies that $\mu_{-2} = 1/(p-2)$, $\mu_{-1} = \Gamma((p-1)/2)/[\sqrt{2}\Gamma(p/2)]$, and $\mu_1 = \sqrt{2}\Gamma((p+1)/2)/\Gamma(p/2)$. Table 13.1 below provides the values of three bounds of a for different p.

One can see from Table 13.1 that the new bound of a is the best, especially, it is much better than other two bounds when the dimensionality is small.

Table 13.1 Bounds of a

p	4	5	6	7	8	9	10	15	20	30	40	50	75	100
a_{new}	0.750	0.800	0.833	0.857	0.875	0.889	0.900	0.933	0.950	0.967	0.975	0.980	0.987	0.990
a_{bs}	0.500	0.600	0.667	0.714	0.750	0.778	0.800	0.867	0.900	0.933	0.950	0.960	0.973	0.980
a_{xi}	0.429	0.444	0.455	0.462	0.467	0.471	0.474	0.483	0.487	0.492	0.494	0.495	0.497	0.497

Table 13.2 Bounds of a

p	4	5	6	7	8	9	10	15	20	30	40	50	75	100
a_{new}	0.150	0.133	0.119	0.107	0.097	0.089	0.082	0.058	0.045	0.031	0.024	0.019	0.013	0.010
a_{bs}	0.125	0.120	0.111	0.102	0.094	0.086	0.080	0.058	0.045	0.031	0.024	0.019	0.013	0.010
a_{xi}	0.115	0.105	0.096	0.088	0.081	0.075	0.070	0.052	0.041	0.029	0.023	0.019	0.013	0.010

Example 13.3 Let $\mathbf{X}$ have a uniform distribution in the unit sphere centered at $\boldsymbol{\theta}$. Then $R = \|\mathbf{X} - \boldsymbol{\theta}\|$ has a probability density function (pdf) pr^{p-1}, $0 \le r \le 1$ and $\mu_i = p/(p+i)$, $i = -2, -1, 1, 2$. Thus, $a_{\text{bs}} = (p-2)/p^2$, $a_{\text{xi}} = (p-1)/(p^2 + 3p - 2)$, and $a_{\text{new}} = (p-1)/[p(p+1)]$. Table 13.2 below provides the values of three bounds of a for different p.

One can see from Table 13.2 that the new bound of a is the best. Meanwhile, all three bounds will approach to zero when the dimensionality increases.

13.3 Extensions to Other Loss Functions and the Unknown Scale Case

Similar to Xu and Izmirlian [24], we consider two extensions in this section. The first one is to show that Theorem 13.1 in Sect. 13.2 can be generalized to a larger class of loss functions, while the second one is to estimate the location vector with an unknown scale parameter.

The loss function used in the first extension is

$$L(\boldsymbol{\delta}, \boldsymbol{\theta}) = W\left(\|\boldsymbol{\delta} - \boldsymbol{\theta}\|^2\right), \tag{13.7}$$

where W is a nonnegative and nondecreasing concave function. The loss function (13.7) has been studied for the spherically symmetric distributions by many investigators including Bock [1], Brandwein and Strawderman [4, 6, 7], Fan and Fang [12–14], Xu [23], and Xu and Izmirlian [24].

Theorem 13.2 *Let F be the df of $R = \|\mathbf{X} - \boldsymbol{\theta}\|$ satisfying*

$$0 < \int_0^\infty W'\left(r^2\right) dF(r) < \infty,$$

where W' is the derivative of W. Suppose that $\mathbf{X}$ is spherically symmetric about $\boldsymbol{\theta}$ and $\boldsymbol{\delta}_{a,\mathbf{g}}(\mathbf{X})$ is defined by (13.2). Then under loss function (13.7), $\boldsymbol{\delta}_{a,\mathbf{g}}(\mathbf{X})$ has a smaller risk than $\mathbf{X}$ if the conditions (i) and (ii) of Theorem 13.1 hold and

(iii) $0 < a < \nu_1/(p\nu_{-1})$, *where* $\nu_i = E_G(R^i)$ *for* $i = -1, 1$ *and* G *is a weighted df of* F *with the weight function* $W'\left(r^2\right)$ *defined by*

$$G(t) = \left(\int_0^\infty W'\left(r^2\right) dF(r)\right)^{-1} \int_0^t W'\left(r^2\right) dF(r), \quad t \geq 0.$$

Now we investigate the problem of estimating the location vector $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$ when the observation $(\mathbf{X}^T, \mathbf{Y}^T)^T$ contains an $m \times 1$ residual vector $\mathbf{Y}$ such that $\mathbf{X}_*^T = (1/\sigma)(\mathbf{X}^T, \mathbf{Y}^T)$ follows a spherically symmetric distribution $\mathrm{SS}_{p+m}(\boldsymbol{\theta}_*, \sigma^2 I_{p+m})$, where $\boldsymbol{\theta}_*^T = (\boldsymbol{\theta}^T, \mathbf{0}_m^T)$, $\mathbf{0}_m$ is an $m \times 1$ vector in which all elements are zero, and σ is an unknown scale. The improved estimators we consider is given by

$$\boldsymbol{\delta}^*_{a,\,\mathbf{g}}(\mathbf{X}_*) = \mathbf{X} + a\mathbf{Y}^T\mathbf{Y}\mathbf{g}(\mathbf{X}). \tag{13.8}$$

Theorem 13.3 *Suppose that* $\mathbf{X}$ *is a* $p \times 1$ *random vector and* $\mathbf{Y}$ *is an* $m \times 1$ *random vector such that* $\mathbf{X}_* = (1/\sigma)\,(\mathbf{X}^T, \mathbf{Y}^T)^T \sim \mathrm{SS}_{p+m}(\boldsymbol{\theta}^*, \sigma^2\, I_{p+m})$. *Let* $\boldsymbol{\delta}^*_{a,\,\mathbf{g}}(\mathbf{X}_*)$ *be defined by (13.8). Then under the scaled quadratic loss function*

$$L(\boldsymbol{\delta}, \boldsymbol{\theta}) = \|\boldsymbol{\delta} - \boldsymbol{\theta}\|^2/\sigma^2,$$

$\boldsymbol{\delta}^*_{a,\,\mathbf{g}}(\mathbf{X}_*)$ *dominates* $\mathbf{X}$ *if conditions (i) and (ii) of Theorem 13.1 hold and*

(iii) $0 < a < (p-1)/[p(m+2)]$.

The bound of a in Theorem 13.3 doesn't depend on the distribution of $\mathbf{X}_*$. Cellier, Fourdrinier and Robert [10] first observed this type of robustness phenomenon for the James-Stein estimator.

13.4 Discussion

If $-h$ is superharmonic, Brandwein and Strawderman [7] used the fact that its average over the ball ("volume") is greater than its average over the sphere ("surface area") to show the dominance of the estimators of the form $\boldsymbol{\delta}_{a,\,\mathbf{g}}(\mathbf{X})$ over the estimator $\mathbf{X}$. In this paper we use the fact that the average of $-h$ over the sphere is a nonincreasing function of the radius of the sphere. The new approach allows us not only to weaken their technical condition (ii), but also to obtain a better bound for a. The new bound of a is also better than those of Brandwein and Strawderman [7] and Xu and Izmirlian [24]. In addition, we consider two extensions. The first is to extend the quadratic loss (13.3) to the loss function (13.7), while the second is to study the estimators of the location vector when the observation $(\mathbf{X}^T, \mathbf{Y}^T)^T$ contains a residual vector $\mathbf{Y}$ and the scale is unknown. While the bounds of a given by the theorems in Sects. 13.2 and 13.3 are better than those of Brandwein and Straderman [7] and Xu and Izmirlian [24], they are not necessarily optimal and should be considered a guide post. Clearly, one may

be able to obtain better bounds than those given here if the distribution of R is known. Stein [20], for example, used integration by parts to obtain $0 < a \leq 1(= \mu_2/p)$ under normality. Thus, it would be interesting to see if our new bound $0 < a \leq \mu_1/(p\mu_{-1})$ can be further improved to $0 < a \leq \mu_2/p$ for $\mathbf{X} \sim \mathrm{SS}_p(\boldsymbol{\theta}, I_p)$. As a final point, it would be interesting, but perhaps very difficult, to study the dominance conditions of the estimator $\boldsymbol{\delta}_{a,\mathbf{g}}(\mathbf{X}) = \mathbf{X} + a\mathbf{g}(\mathbf{X})$ over $\mathbf{X}$ for other distributions. As mentioned by Xu and Izmirlian [24], results of estimator (13.1) obtained by Shinozaki [18] for the class of distributions with independently and identically distributed components and by Xu [23] for the sign-invariant distribution are very limited.

13.5 Proofs

In this section we use $f_{c,s}(z)$ to denote the pdf of the Beta distribution Beta(c, s) given by

$$f_{c,s}(z) = \frac{\Gamma(c+s)}{\Gamma(c)\Gamma(s)} z^{c-1}(1-z)^{s-1}, \quad 0 < z < 1,$$

where $c > 0$ and $s > 0$ are parameters. To shorten the proofs of results in Sects. 13.2 and 13.3, we need the following lemmas in which the first one is taken from Wijsman's [22] Theorem 2.

Lemma 13.1 *Let α be a measure on the real line $\mathbb{R}$ and let f_j, g_j $(j = 1, 2)$ be Borel-measurable functions: $\mathbb{R} \to \mathbb{R}$ such that $f_2 \geq 0$, $g_2 \geq 0$, and $\int |f_i g_j| d\alpha < \infty$ $(i, j = 1, 2)$. If f_1/f_2 and g_1/g_2 are monotonic in the same direction, then*

$$\int f_1 g_1 d\alpha \int f_2 g_2 d\alpha \geq \int f_1 g_2 d\alpha \int f_2 g_1 d\alpha, \tag{13.9}$$

whereas if f_1/f_2 and g_1/g_2 are monotonic in the opposite directions, then inequality in (13.9) is reversed. The equality in (13.9) holds if and only if $f_2 = 0$ or $g_2 = 0$ or $f_1/f_2 =$ constant or $g_1/g_2 =$ constant almost everywhere with respect to the measure ρ defined by $d\rho = (|f_1| + |f_2|)(|g_1| + |g_2|)d\alpha$.

Lemma 13.2 *Let the function M_* be defined by (13.4). Then*

$$\int_0^1 M_*(rz) f_{c-1,1}(z) dz \leq M_*(r) \leq \int_0^1 M_*(rz) c z^{c-2} dz,$$

for any $r > 0$, where $c > 1$ is a constant.

Proof Since M_* is a nondecreasing concave function with $M_*(0) = 0$ and the expected value of Beta$(c - 1, 1)$ distribution is $(c - 1)/c$, using the Jensen's inequality will yield that

$$\int_0^1 M_*(rz) f_{c-1,1}(z)dz \le M_*\left(r\frac{c-1}{c}\right) \le M_*(r).$$

Furthermore, the concavity of M_* implies that $M_*(rz) \ge zM_*(r)$ for $z \in [0, 1]$ and $r > 0$. Thus,

$$\int_0^1 M_*(rz)cz^{c-2}dz \ge \int_0^1 M_*(r)cz^{c-1}dz = M_*(r)\int_0^1 cz^{c-1}dz = M_*(r).$$

Lemma 13.3 *For $z \in [0, 1]$, let*

$$\ell(z) = \frac{\beta(r)}{p} f_{p,1}(z) + 1 - \frac{\beta(r)}{p},$$

where $\beta(r) = r^2/a$ is considered a parameter. Then $\ell(z)$ is a pdf on $[0, 1]$ when $\beta(r) \le p$. Furthermore, when $\beta(r) \le p$, we have

$$m(r) - \beta(r)\int_0^1 m(rz)z^{p-1}dz \le \left(1 - \frac{\beta(r)}{p}\right)\frac{M_*(r)}{r}$$

for $r > 0$, where m and M_* are defined by (13.4).

Proof When $\beta(r) \le p$, $\ell(z)$ is a pdf on $[0, 1]$ because it is a convex combination of pdfs $f_{p,1}(z)$ and $f_{1,1}(z) = 1$ on $[0, 1]$. Furthermore, since m is a nonincreasing function, we have

$$m(r) \le \int_0^1 m(rz)\ell(z)dz,$$

which leads to

$$\begin{aligned} m(r) - \beta(r)\int_0^1 m(rz)z^{p-1}dz &\le \int_0^1 m(rz)\ell(z)dz - \frac{\beta(r)}{p}\int_0^1 m(rz)f_{p,1}(z)dz \\ &= \int_0^1 m(rz)\left(\ell(z) - \frac{\beta(r)}{p}f_{p,1}(z)\right)dz \\ &= \left(1 - \frac{\beta(r)}{p}\right)\int_0^1 m(rz)dz \\ &= \left(1 - \frac{\beta(r)}{p}\right)\frac{M_*(r)}{r}. \end{aligned}$$

Lemma 13.4 *When $\beta(r) = r^2/a > p$, we have*

$$m(r) - \beta(r)\int_0^1 m(rz)z^{p-1}dz \le \left(1 - \frac{\beta(r)}{p}\right)\frac{c}{r}\int_0^1 M_*(rz)pz^{p-2}dz$$

for $r > 0$, where m and M_* are defined by (13.4) and $c \in [1, p-1]$ is a constant.

Proof Since m is a nonincreasing function, we have

$$m(r) \le \int_0^1 m(rz) f_{p,1}(z) dz,$$

which leads to

$$\begin{aligned} m(r) - \beta(r) \int_0^1 m(rz) z^{p-1} dz &\le \left(1 - \frac{\beta(r)}{p}\right) \int_0^1 m(rz) f_{p,1}(z) dz \\ &\le \left(1 - \frac{\beta(r)}{p}\right) \frac{c}{r} \int_0^1 M_*(rz) p z^{p-2} dz. \end{aligned} \tag{13.10}$$

Here the last inequality in (13.10) follows from the condition (ii) of Theorem 13.1 and $\beta(r) > p$.

Remark 13.5 Lemmas 13.3 and 13.4 can be combined below:

$$m(r) - \beta(r) \int_0^1 m(rz) z^{p-1} dz \le N_1(r) N_2(r), \tag{13.11}$$

where $\beta(r) = r^2/a$ and

$$\begin{aligned} N_1(r) &= \left(1 - \frac{\beta(r)}{p}\right) \frac{1}{r}, \\ N_2(r) &= I[\beta(r) \le p] M_*(r) + cI[\beta(r) > p] \int_0^1 M_*(rz) p z^{p-2} dz. \end{aligned} \tag{13.12}$$

Here $I[A]$ denotes the indicator function of the event A.

Lemma 13.5 *For $r > 0$, let $N_1(r)$ and $N_2(r)$ be defined by (13.12). Then $N_1(r)$ is strictly decreasing in r and $N_2(r)$ is nondecreasing in r. Furthermore, $E_R[N_1(R) N_2(R)] \le 0$ if $a \le \mu_1/(p\mu_{-1})$.*

Proof Since $N_1(r) = 1/r - r/(ap)$, it is a strictly decreasing function of r. Similarly, since both $M_*(r)$ and $\int_0^1 M_*(rz) p z^{p-2} dz$ are nondecreasing in r and $M_*(r) \le \int_0^1 M_*(rz) p z^{p-2} dz$ from Lemma 13.2, we can conclude that $N_2(r)$ is a nondecreasing function of r. Furthermore, applying Lemma 13.1 with $f_1(r) = N_1(r)$, $g_1(r) = N_2(r)$, $f_2(r) = g_2(r) = 1$, and a probability measure $d\alpha = dF$ will yield that

$$E_R[N_1(R) N_2(R)] \le E_R[N_1(R)] E_R[N_2(R)] = \left(\mu_{-1} - \frac{\mu_1}{ap}\right) E[N_2(R)] \le 0$$

if $a \le \mu_1/(p\mu_{-1})$ because $N_2(r) \ge 0$ for $r > 0$.

Proof of Theorem 13.1. When $\mathbf{X} \sim \mathrm{SS}_p(\boldsymbol{\theta}, I_p)$, we have $\mathbf{X} - \boldsymbol{\theta} = \mathbf{Z} \stackrel{d}{=} R\mathbf{U}$, where R and $\mathbf{U}$ are independent, $R \stackrel{d}{=} \|\mathbf{Z}\|$, and $\mathbf{U}$ has a uniform distribution on the surface of the unit sphere. Using the argument of Xu and Izmirlian [24] with a verbatim copy of their (12), we obtain that the difference between the risks of two estimators $\boldsymbol{\delta}_{a,\mathbf{g}}(\mathbf{X})$ and $\mathbf{X}$ is given by

$$\begin{aligned} D_1 &= R\big(\boldsymbol{\delta}_{a,\mathbf{g}}(\mathbf{X}), \boldsymbol{\theta}\big) - R\big(\mathbf{X}, \boldsymbol{\theta}\big) \\ &= a^2 E\left[\|\mathbf{g}(\mathbf{Z}+\boldsymbol{\theta})\|^2\right] + 2aE\left[\mathbf{Z}^T\mathbf{g}(\mathbf{Z}+\boldsymbol{\theta})\right] \\ &= a^2 E\left[\|\mathbf{g}(\mathbf{Z}+\boldsymbol{\theta})\|^2\right] + 2ap^{-1}E\left[R^2\nabla \circ \mathbf{g}(R\mathbf{V}+\boldsymbol{\theta})\right] \\ &\le 2a^2 E\left[-h(R\mathbf{U}+\boldsymbol{\theta})\right] + 2ap^{-1}E\left[R^2 h(R\mathbf{V}+\boldsymbol{\theta})\right] \\ &= 2a^2 E_R\left[E_{\mathbf{U}}\left(-h(R\mathbf{U}+\boldsymbol{\theta})\big|R\right) + (ap)^{-1}R^2 E_{\mathbf{V}}\left(h(R\mathbf{V}+\boldsymbol{\theta})\big|R\right)\right] \\ &= 2a^2 E_R\left[m(R) - (ap)^{-1}R^2 E_{\mathbf{V}}\left(h(R\mathbf{V}+\boldsymbol{\theta})\big|R\right)\right] \\ &= 2a^2 E_R\left[m(R) - \beta(R)\int_0^1 m(Rv)v^{p-1}dv\right] \\ &\le 2a^2 E_R\left[N_1(R)N_2(R)\right] \\ &\le 0 \end{aligned} \tag{13.13}$$

if $a \le \mu_1/(p\mu_{-1})$. Here the first inequality in (13.13) is based on the condition (i); the fifth equality in (13.13) is from the definition of function m; the last equality in (13.13) follows from the definition of m and the fact that $\mathbf{V} \stackrel{d}{=} V\mathbf{U}$, where the random variable $V \sim \mathrm{Beta}(p, 1)$ and $\mathbf{U}$ having a uniform distribution on the surface of the unit sphere are independent; the second-to-last inequality in (13.13) is based on Lemmas 13.3 and 13.4 or (13.11); the last inequality in (13.13) follows from Lemma 13.5. This completes the proof.

Proof of Theorem 13.2. Using the same approach as in Brandwein and Strawderman [4, 6] or Xu and Izmirlian [24], we obtain that the difference between the risks of two estimators $\boldsymbol{\delta}_{a,\mathbf{g}}(\mathbf{X})$ and $\mathbf{X}$ is given by

$$\begin{aligned} D_2 &= R\big(\boldsymbol{\delta}_{a,\mathbf{g}}(\mathbf{X}), \boldsymbol{\theta}\big) - R(\mathbf{X}, \boldsymbol{\theta}) \\ &= E\left[W\left(R^2 + \Delta_a(\mathbf{X})\right)\right] - E\left[W\left(R^2\right)\right], \end{aligned} \tag{13.14}$$

where

$$\Delta_a(\mathbf{X}) = \|\boldsymbol{\delta}_{a,\mathbf{g}}(\mathbf{X}) - \boldsymbol{\theta}\|^2 - \|\mathbf{X} - \boldsymbol{\theta}\|^2.$$

Since W is a nondecreasing concave function,

$$W\left(R^2 + \Delta_a(\mathbf{X})\right) < W\left(R^2\right) + W'\left(R^2\right)\Delta_a(\mathbf{X}).$$

Then we can conclude from (13.14) that

$$\begin{aligned} D_2 &\le E_{\mathbf{X}}\left[W'\left(R^2\right)\Delta_a(\mathbf{X})\right] \\ &\le E_R\left\{W'\left(R^2\right)E_{\mathbf{U}}[\Delta_a(R\mathbf{U}+\boldsymbol{\theta})|R]\right\} \\ &\le 2a^2E_R\left[W'(R^2)N_1(R)N_2(R)\right] \\ &= 2a^2E_{R_*}\left[N_1(R_*)N_2(R_*)\right]E_R\left[W'\left(R^2\right)\right], \end{aligned}$$

where the df G of the random variable R_* is defined by

$$G(t)=\left(\int_0^\infty W'\left(r^2\right)dF(r)\right)^{-1}\int_0^t W'\left(r^2\right)dF(r),\quad t\ge 0,$$

which is a weighted df of F with the weight function $W'\left(r^2\right)$. The result follows immediately from the assumption that $0<E_R\left[W'\left(R^2\right)\right]<\infty$ and the proof of Theorem 13.1 except for a change from the df F to the df G.

Proof of Theorem 13.3. Like Brandwein and Strawderman [7] and Xu and Izmirlian [24], the difference D_3 between the risks of two estimators $\delta^*_{a,\mathbf{g}}(\mathbf{X}_*)$ and $\mathbf{X}$ is equal to

$$\begin{aligned} D_3 &= R\left(\delta^*_{a,\mathbf{g}}(\mathbf{X}_*),\boldsymbol{\theta}\right)-R\left(\mathbf{X},\boldsymbol{\theta}\right) \\ &= \frac{1}{\sigma^2}E\left[a^2(\mathbf{Y}^T\mathbf{Y})^2\|\mathbf{g}(\mathbf{Z}+\boldsymbol{\theta})\|^2+2a\,\mathbf{Y}^T\mathbf{Y}\mathbf{Z}^T\mathbf{g}(\mathbf{Z}+\boldsymbol{\theta})\right] \\ &= \frac{1}{\sigma^2}E\left(a^2D_{31}+2aD_{32}\right), \end{aligned} \tag{13.15}$$

where $\mathbf{Z}=\mathbf{X}-\boldsymbol{\theta}\overset{d}{=}R\mathbf{U}$, and

$$\begin{aligned} D_{31} &= E\left[(\mathbf{Y}^T\mathbf{Y})^2\|\mathbf{g}(\mathbf{Z}+\boldsymbol{\theta})\|^2\big|\,\|\mathbf{Z}\|=R,\|\mathbf{Y}\|=S\right], \\ D_{32} &= E\left[\mathbf{Y}^T\mathbf{Y}\mathbf{Z}^T\mathbf{g}(\mathbf{Z}+\boldsymbol{\theta})\big|\,\|\mathbf{Z}\|=R,\|\mathbf{Y}\|=S\right]. \end{aligned}$$

Using the divergence theorem and condition (i), we obtain that

$$\begin{aligned} D_{32} &= E\left[\mathbf{Y}^T\mathbf{Y}\mathbf{Z}^T\mathbf{g}(\mathbf{Z}+\boldsymbol{\theta})\big|\|\mathbf{Z}\|=R,\|\mathbf{Y}\|=S\right] \\ &= S^2RE_{\mathbf{U}}\left[\mathbf{U}^T\mathbf{g}(R\mathbf{U}+\boldsymbol{\theta})\big|\|\mathbf{Z}\|=R,\|\mathbf{Y}\|=S\right] \\ &= \frac{S^2R^2}{p}E_{\mathbf{V}}\left(\nabla\circ\mathbf{g}(R\mathbf{V}+\boldsymbol{\theta})\big|\|\mathbf{Z}\|=R,\|\mathbf{Y}\|=S\right) \\ &\le -\frac{S^2R^2}{p}\int_0^1 m(Rz)f_{p,1}(z)dz, \end{aligned} \tag{13.16}$$

where m is defined by (13.4). Similarly, using the condition (i) will yield that

$$\begin{aligned} D_{31} &= E\left[(\mathbf{Y}^T\mathbf{Y})^2\|\mathbf{g}(\mathbf{Z}+\boldsymbol{\theta})\|^2\big|\|\mathbf{Z}\| = R, \|\mathbf{Y}\| = S\right] \\ &\le -2S^4 E\left[h(\mathbf{Z}+\boldsymbol{\theta})\big|\|\mathbf{Z}\| = R, \|\mathbf{Y}\| = S\right] \\ &= -2S^4 E\left[h(R\mathbf{U}+\boldsymbol{\theta})\big|\|\mathbf{Z}\| = R, \|\mathbf{Y}\| = S\right] \\ &= 2S^4 m(R). \end{aligned} \tag{13.17}$$

Combining (13.16), (13.17) with (13.15) and using the same argument as the proof of theorem 13.1 will obtain the following inequality

$$\begin{aligned} D_3 &\le \frac{2a}{\sigma^2} E\left(aS^4 m(R) - \frac{S^2R^2}{p}\int_0^1 m(Rz)\, f_{p,1}(z)\, dz\right) \\ &= \frac{2a}{\sigma^2} E\left[(aS^4)\left(m(R) - \frac{R^2}{aS^2}\int_0^1 m(Rz) z^{p-1} dz\right)\right] \\ &\le \frac{2a}{\sigma^2} E\left[(aS^4)\left(1 - \frac{R^2}{apS^2}\right)\frac{1}{R}N_2(R)\right], \end{aligned} \tag{13.18}$$

where the first inequality in (13.18) is based on (13.16) and (13.17), the last inequality of (13.18) follows from (13.11) after replacing a by aS^2, and $N_2(R)$ is defined by (13.12). Let $T^2 = R^2 + S^2$. Then T^2 and $B = R^2/T^2 \sim \text{Beta}(p/2, m/2)$ are independent. Let $C(c,s) = \Gamma(c+s)/[\Gamma(c)\Gamma(s)]$ for $c > 0, s > 0$ and let $C^* = C(p/2, m/2)/C(p/2, (m+2)/2)$. Write $\lambda = a + 1/p$. Then we can see from (13.18) that

$$\begin{aligned} \frac{\sigma^2}{2a} D_3 &\le E\left[aS^4\left(1 - \frac{R^2}{apS^2}\right)\frac{1}{R}N_2(R)\right] \\ &= E\left[(1-B)(a-\lambda B) T^4\left(N_1\left(TB^{1/2}\right) + \frac{N\left(TB^{1/2}\right)}{TB^{1/2}}\right)\right] \\ &= C^* E\left[(a-\lambda B) T^4\left(N_1\left(TB^{1/2}\right) + \frac{N\left(TB^{1/2}\right)}{TB^{1/2}}\right)\right] \\ &= C^* E\left[(a-\lambda B) T^4 N_1\left(TB^{1/2}\right)\right] \\ &\quad + C^* E\left[\left(aB^{-1/2} - \lambda B^{1/2}\right) T^3 N(TB^{1/2})\right] \\ &\le C^*\left(a - \lambda\frac{p}{p+m+2}\right) E\left[T^4 N_1\left(TB^{1/2}\right)\right] \\ &\quad + C^*\left(a\frac{C(p/2,(m+2)/2)}{C((p-1)/2,(m+2)/2)} - \lambda\frac{C(p/2,(m+2)/2)}{C((p+1)/2,(m+2)/2)}\right) \\ &\quad \times E\left[T^3 N\left(TB^{1/2}\right)\right] \\ &\le 0 \end{aligned} \tag{13.19}$$

if

$$\begin{aligned} &a - \lambda \frac{p}{p+m+2} \le 0, \\ &a \frac{C\,(p/2, (m+2)/2)}{C\,((p-1)/2, (m+2)/2)} - \lambda \frac{C\,(p/2, (m+2)/2)}{C\,((p+1)/2, (m+2)/2)} \le 0. \end{aligned} \tag{13.20}$$

Here the second-to-last inequality of (13.19) follows from applications of Lemma 13.1 with the measure $d\alpha = f_{p/2,(m+2)/2}(b)db$ on $[0, 1]$ and $f_1(b) = a - \lambda b$, $g_1(b) = T^4 N_1(Tb^{1/2})$, $f_2(b) = g_2(b) = 1$, and $f_1(b) = ab^{-1/2} - \lambda b^{1/2}$, $g_1(b) = T^3 N(Tb^{1/2})$, $f_2(b) = g_2(b) = 1$, respectively. Simple algebra shows that the first inequality in (13.20) is equivalent to $0 < a \le 1/(m+2)$, while the second inequality in (13.20) is equivalent to $0 < a \le (p-1)/[p(m+2)]$. Therefore, $D_3 \le 0$ if $0 < a \le (p-1)/[p(m+2)]$.

References

1. Bock, M.E.: Minimax estimators that shift towards a hypersphere for location vectors of spherically symmetric distributions. J. Multivariate Anal. **17**, 127–147 (1985)
2. Brandwein, A.C.: Minimax estimation of mean of spherically symmetric distributions under general quadratic loss. J. Multivariate Anal. **9**, 579–588 (1979)
3. Brandwein, A.C., Strawderman, W.E.: Minimax estimation of location parameters for spherically symmetric unimodal distributions under quadratic loss. Ann. Statist. **6**, 377–416 (1978)
4. Brandwein, A.C., Strawderman, W.E.: Minimax estimation of location parameters for spherically symmetric distributions with concave loss. Ann. Statist. **8**, 279–284 (1980)
5. Brandwein, A.C., Strawderman, W.E.: Stein estimation, The spherically symmetric case. Statist. Sci. **5**, 356–369 (1990)
6. Brandwein, A.C., Strawderman, W.E.: Stein estimation for spherically symmetric distributions: recent developments. Statist. Sci. **27**, 11–23 (2012)
7. Brandwein, A.C., Strawderman, W.E.: Generalizations of James-Stein estimators under spherical symmetry. Ann. Statist. **19**, 1639–1650 (1991)
8. Brown, L.D.: On the admissibility of invariant estimators of one or more location parameters. Ann. Math. Statist. **37**, 1087–1136 (1966)
9. Brown, L.D., Zhao, L.H.: A geometrical explanation of Stein shrinkage. Statist. Sci. **27**, 24–30 (2012)
10. Cellier, D., Fourdrinier, D., Robert, C.: Robust shrinkage estimators of the location parameter for elliptically symmetric distributions. J. Multivariate Anal. **29**, 39–52 (1988)
11. Du Plessis, N.: An Introduction to Potential Theory. Hafner, Darien, CT (1970)
12. Fan, J., Fang, K.-T.: Inadmissibility of sample mean and sample regression coefficients for elliptically contoured distributions. In: Fang, K.-T., Anderson, T.W. (eds.) Statistical Inference in Elliptically Contoured and Related Distributions, pp. 275–290. Allerton Press Inc., New York (1990)
13. Fan, J., Fang, K.-T.: Inadmissibility of the usual estimator for the location parameters of spherically symmetric distributions. In: Fang, K.-T., Anderson, T.W. (eds.) Statistical Inference in Elliptically Contoured and Related Distributions, pp. 291–297. Allerton Press Inc., New York (1990)
14. Fan, J., Fang, K.-T.: Shrinkage estimators and ridge regression estimators for elliptically contoured distributions. In: Fang, K.-T., Anderson, T.W. (eds.) Statistical Inference in Elliptically Contoured and Related Distributions, pp. 313–326. Allerton Press Inc., New York (1990)

15. James, W., Stein, C.: Estimation with quadratic loss. Proc. Fourth Berkeley Sympos. Math. Statist. Prob. **1**, 361–379 (1961)
16. Maruyama, Y.: Admissible minimax estimators of a mean vector of scale mixtures of multivariate normal distributions. J. Multivariate Anal. **84**, 274–283 (2003)
17. Miceli, R.J., Strawderman, W.E.: Minimax estimation for certain independent component distributions under weighted squared error loss. Comm. Statist. Theory Methods **15**, 2191–2200 (1986)
18. Shinozaki, N.: Simultaneous estimation of location parameters under quadratic loss. Ann. Statist. **12**, 322–335 (1984)
19. Stein, C.: Inadmissibility of the usual estimator for the mean vector of a multivariate normal distribution. Proc. Third Berkeley Sympos. Math. Statist. Prob. **1**, 197–206 (1956)
20. Stein, C.: Estimation of the mean of a multivariate normal distribution. Ann. Statist. **9**, 1135–1151 (1981)
21. Tosh, C., Dasgupta, S.: Maximum likelihood estimation for mixtures of spherical Gaussians is NP-hard. J. Mach. Learn. Res. **18**, 1–11 (2018)
22. Wijsman, R.A.: A useful inequality on ratios of integrals, with application to maximum likelihood estimation. J. Amer. Statist. Assoc. **80**, 472–475 (1985)
23. Xu, J.-L.: Simultaneous estimation of location parameters for sign-invariant distributions. Ann. Statist. **25**, 2259–2272 (1997)
24. Xu, J.-L., Izmirlian, G.: Estimation of location parameters for spherically symmetric distributions. J. Multivariate Anal. **97**, 514–525 (2006)

Chapter 14
On Equidistant Designs, Symmetries and Their Violations in Multivariate Models

Milan Stehlík, Mirtha Pari Ruiz, Silvia Stehlíková, and Ying Lu

Abstract In this Festschrift to Prof. Kai-Tai Fang 80 birthday we emphasize importance and potential of his results in statistics and general sciences. In particular we concentrate on equidistant designs, symmetric and asymmetric models. We discuss equidistant designs from perspective of optimal designs of experiments with correlated errors. We address symmetry and asymmetry of statistical multivariate models and its recent developments. Several applications are given.

14.1 Introduction

With this contribution we congratulate to Prof. Kai-Tai Fang on occasion of his 80th birthday. We appreciate results of Kai-Tai Fang which inspired several research developments and generation of young statisticians. Uniform distribution plays also

M. Stehlík
Institute of Statistics, University of Valparaíso, Valparaíso Chile and Department of Applied Statistics and Linz Institute of Technology, Johannes Kepler University Linz, Linz, Austria

Department of Statistics and Actuarial Science, The University of Iowa, Iowa City, Iowa, USA
e-mail: milan.stehlik@uv.cl; milan.stehlik@jku.at

M. Pari Ruiz
Departamento de Matemáticas, Universidad de Tarapacá (UTA), Arica, Chile

Universidad de Playa Ancha, Valparaíso, Chile

Institute of Statistics, University of Valparaíso, Valparaíso, Chile
e-mail: mirtha.pari@postgrado.uv.cl; mpari@academicos.uta.cl

S. Stehlíková
Department of Applied Statistics and Linz Institute of Technology, Johannes Kepler University Linz, Linz, Austria
e-mail: silvia.stehlikova@gmail.com

Y. Lu (✉)
Department of Biomedical Data Science, Stanford University, California, USA
e-mail: ylu1@stanford.edu

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_14

principal role in empirical distribution of probability quasi-distances, like we illustrate for Nican Mopohua text, see Sect. 14.3.2. We also provide a brief review of work related to uniform optimal designs, see Sect. 14.2 and the model related asymmetric multivariate distributions, a natural extension of symmetric distribution models.

14.2 On Uniform Optimal Designs

Fang [8] defined uniform designs as a very useful and potential technique for optimal designing for engineering. In 2007 we have proven D-optimality of equidistant design for trend parameter of Ornstein Uhlenbeck (OU) Process, see [17].

That justifies the importance of uniform engineering designs also for models with correlated errors, especially for exponentially decaying covariances with respect to distance of design points. A statistical model we consider is the so called random field, given by

$$Y(x) = \eta(x, \beta) + \varepsilon_\gamma(x) \tag{14.1}$$

with design points (coordinates of monitoring sites) $\xi_n = \{x_1, \ldots, x_n\}$ taken from a compact design space $\mathcal{X} = X^n$, $X = [a, b]$, $-\infty < a < b < \infty$. The parameters β are unknown and the variance-covariance structure of the errors depends on parameters γ. When distribution of errors is known, we can employ the maximum likelihood estimators (MLEs). For the full parameter set $\{\beta, \gamma\}$ the information matrix then exhibits the block diagonal form

$$E\left\{\begin{array}{cc} -\frac{\partial^2 lnL(\beta,\theta)}{\partial\beta\partial\beta^T} & -\frac{\partial^2 lnL(\beta,\gamma)}{\partial\beta\partial\gamma^T} \\ -\frac{\partial^2 lnL(\beta,\gamma)}{\partial\gamma\partial\beta^T} & -\frac{\partial^2 lnL(\beta,\gamma)}{\partial\gamma\partial\gamma^T} \end{array}\right\} = \begin{pmatrix} M_\beta(\xi) & 0 \\ 0 & M_\gamma(\xi) \end{pmatrix}.$$

The theoretical justification for D-optimality in correlated errors was given by [23] and specific issues have been presented in [22]. Pázman [23] also demonstrated that the inverse of the Fisher information matrix may well serve as an approximation of the covariance matrix of MLEs in special cases. Since 2008 we have been addressing a very innovative field in modeling of spatiotemporal random fields with semicontinuous covariance functions. This is a novel and groundbreaking approach which relates to open problems of N.A. Kolmogoroff, some preliminary results have been outlined in [29], where more general framework for equidistant designs arrives naturally by the topological relaxation of distances between individual design points. This fact is not visible from standard optimal design perspective. In [33] we have derived theoretical basis for this approach jointly with groundbreaking results for continuity of covariance and its applications to finances. Uniform designs plays an important role also by D-optimality and optimal prediction for Ornstein-Uhlenbeck sheets, see [2, 4]. Within Chilean project FONDECYT Regular No. 1151441 we obtained optimal designs for mass balance measurements on Chilean mountain glaciers Olivares Alfa and Beta (see [34]). This has clarified old glaciological problem that for small

homogeneous glaciers a moderate number of properly allocated stakes for measurement of mass balance is sufficient and the variance of mass balance estimator can grow with additional number of stakes. In [3] we received non standard design strategies for Kolmogorov's model of Chandler wobble small deviation in the Earth's axis of rotation. We are currently working on extension of these results for spatiotemporal random fields with semicontinuous covariance functions, which is fairly nontrivial task with plenty of important applications to chemometrics, ecological modelling and finances.

Uniform designs play also fundamental role for prediction of random fields. As a classical criterion we can consider Empirical Kriging (EK) prediction error. Here we have to minimize the so-called kriging variance $Var[\widehat{Y}(x|\xi)] = E[(\widehat{Y}(x|\xi) - Y(x))^2]$ (Mean Squared Prediction Error—MSPE), where $\widehat{Y}(x|\xi)$ denotes the best linear unbiased predictor of $Y(x)$ based on the design points in ξ. The EK-optimal design minimizes the criterion function

$$\psi(\xi) = \max_{x \in X} \widehat{Var[\widehat{Y}(x|\xi)]}. \tag{14.2}$$

However, this criterion is difficult to compute. The results for optimality of equidistance designs also for OU sheets have been derived in [2].

14.3 On Symmetric Multivariate Distributions and Beyond

Fang et al. [9, 10] provided basis for study of elliptically symmetric distributions. For geometric measure representation see [25, 26]. Such developments can be generalized by a class of star shaped distributions and their representations, see e.g. [31]. The p-generalized elliptically contoured distributions for $p = 1$ have been derived in [14] for probability mass concentrated on R^{n+}. Such forms of construction motivate the problem of probabilistic quasi-distances which can be both symmetric and asymmetric. Moreover, copulas of elliptically contoured distributions are of interest, see [11], which can be generalized to general geometric and topological constructions of aggregation functions (see e.g. [12, 32]). In the following two subsections we will illustrate importance of proper understanding of symmetry, since in some important instances symmetry need to be replaced by asymmetry. Two illustrative subsections follow, namely Pseudoexponential models for dose finding studies and Asymmetric distance measures of linguistic sequences.

14.3.1 Pseudoexponential Models for Dose Finding Studies

The approach by [6] can lead to bivariate cases of pseudoexponential models (see [12, 13]), where we condition on dose exposure levels, in time to one tumor setup

and delayed toxicities. We can justify the join survival type of model

$$P(T > t, D > d) = \exp(-\theta_1 t - \theta_2 d - \theta_2 A\phi(t)d), \quad \phi(0) = 0. \tag{14.3}$$

Term $\theta_2 d$ in Eq. (14.3) is linear, however, nonlinear dependence on d is possible as the distribution of exposure. The formulation for [6] is $P(T > t|d)$, where d is the constant dose. So the joint distribution should be $\int P(T > t|d)dF(d)$. Thus, the D is the distribution of population toxic exposure. If D is exponentially distributed, we have Eq. (14.3). Otherwise, we may have more general formulation, where T is time to tumor, D is the dose exposure of toxins, $A > 0$ is parameter and $\phi(t)$ is the cumulative remaining toxins in the body up to time t. Following the notation by [6], suppose that individual is first exposed to toxic material amount d at age a and continue to time t. Then in a special case of [6] for every t we can define failure rate $\lambda(t, d)$ for tumor in additive form of the mortality intensity function

$$\lambda(t, d) = -\theta_1 t - \theta_2 d - \theta_2 \frac{\delta}{\nu}(1 - \exp(-\nu t)))d. \tag{14.4}$$

This function is a special case of (14.3), with A being the ratio of absobring coefficient to discharge coefficient and the rational for the given function $\phi(t)$ lies in the remaining toxins $\int_0^t d\delta \exp(-\nu(t - s))ds$. Notice that Eq. (14.4) without $\theta_2 d$ is the conditional cumulative hazard function for time to tumor. The independence of the absorbing coefficient δ and the discharging coefficient ν of time used by [6] is in many instances pretty over-realistic and more general version of cumulative remaining dose is captured by pseudoexponential model (14.3).

14.3.2 Asymmetric Cultural Distance Measures on Linguistic Sequences

If we aim to analyze the changes of at least two different samples of historical linguistic corpuses, we need to develop a feasible statistical model which detects spelling variations. As well we need to define measures of variation of given words/sets of words, in order to define measurable statistics of cultural flow and changes. This provides background for a social-linguistic study through designing an appropriate statistical method that detects linguistic variants by the frequency of letters, words and the presence of key words. Such analysis plays crucial role in conservation of the important cultural heritage of e.g. under-represented cultural groups. We consider two samples of historical corpus *Nican Mopohua* and detect the frequency of selected words in order to aggregate information in probabilistic quasi-distance between both texts. *Nican Mopohua* is orally transmitted in language Náhuatl spoken currently by a minority group in Mexico. Corpus was written by Antonio Valeriano (1556) and we compare two samples of corpus by [18, 27]. The *Nican Mopohua*, historical Nahuatl text, allows us to study and analyze the variations of the Nahuatl language.

Such aggregation of linguistic information in quasi-distance is far from being trivial and it links to topological aggregation introduced in [32]. Here we speak

about quasi distance, since mathematical distance, formalized by concept of metrics is symmetric. However, we have empirically observed asymmetries between both corpuses. Therefore we decided to use Kullback–Leibler (KL) divergences, which can be symmetric and asymmetric, dependently on the underlying distribution. If we compare KL divergences for Náhuatl word "*xochitl*" (*flowers*) from one corpus to another, we receive different divergences, namely 0.618 and 0.272. That empirically underlines the fact that we are not in symmetrized world of distances, but asymmetric divergences. This asymmetry supports the fact, that language will naturally relate much more to the topology (see [28, 32]) than to some metric, since language constructs go hand-in-hand with cognition. Náhuatl uses mereological and topological notions of connection, part, interior, and complement which are central to spatial reasoning and to the semantics of natural language expressions concerning locations and relative positions. Thus the exploration of the phonetic differences by symmetric distances is not satisfactory for reconstruction of the ancestral language, e.g. [24] can over-symmetrize the differences between spoken Romance languages.

We work with divergences between probabilities of co-occurrence of linguistic objects that are an important tool in statistics for studies of natural language processes, see [5, 19]. Reference [15] analyzes the aboriginal words, estimating the relative percentage of words. In addition, it considers that the change of the language is through the own evolution of the country. The comparison of the probability distribution of each letter or word of the corpus is made using the KL divergence, with a limited number of comparison terms for the corpus. In general Φ-divergences to the best knowledge of the authors introduced independently by [1, 7]) are used as asymmetric distances between two probability distributions and KL divergence is a special case for $\phi(t) = t\log(t)$. Reference [30] recalls relationship between Φ-divergences and statistical information in the sense of DeGroot, which was shown in [20]. The definition of ϕ-divergence follows.

Definition The ϕ-divergence between the probability distributions P an Q is defined by

$$D_\phi(P, Q) = \int_\chi Q(x)\phi\left(\frac{P(x)}{Q(x)}\right) d\mu(x),$$

where the function $\phi : (0, \infty) \to [0, \infty)$ is assumed to be continuous, decreasing on $(0, 1)$ and increasing on $(1, \infty)$, with $\phi(1) = 0$. The value $\phi(0) \in (0, \infty]$ is defined by the continuous extension.

Especially because of good robustness properties [16] and direct relation to information theory, we consider KL divergence as the best choice for omnibus asymmetric distance between two linguistic probability distributions in our problem. Reference [19] studies the measures of distributional similarity by usage of the weighted average of the distance and thus its uses a weighted version of the KL-divergence. The probabilities of co-occurrence, based on the frequencies of the co-occurrences themselves plays the most important role. It indicates that the similarity between linguistic objects is determined by the similarity of their corresponding vector characteristics, where these characteristics are e.g. numerical frequencies of co-occurrences. We used the following quasidistances listed in Table 14.1 in order to measure various aspects

Table 14.1 Similarity functions for probability distributions

KL divergence	$D(p \parallel q)$	$= \sum_x p(x)(logp(x) - logq(x))$
Jensen-Shannon	$JS(p,q)$	$= \frac{1}{2}[D(p \parallel avg(p,q)) + D(q \parallel avg(p,q))]$
Skew divergence	$S_\alpha(p,q)$	$= D(q \parallel \alpha p + (1-\alpha)q)$
Euclidean	$euc(p,q)$	$= (\sum_x (p(x) - q(x))^2)^{\frac{1}{2}}$
Cosine	$cos(p,q)$	$= \sum_x p(x)q(x)/\sqrt{\sum_x p(x)^2 \sum_x q(x)^2}$
L_1	$L_1(p,q)$	$= \sum_x \lvert p(x) - q(x) \rvert$
Confusion	$conf(p,q,P(y'))$	$= P(y') \sum_x p(x)q(x)/P(x)$

of differences between two complex linguistic corpuses. In general, evaluation of different quasidistances allows us to see various aspects of differences between corpuses. Reference [19] indicates that the skew divergence is the one that achieves the best performance and that it is closest to a KL-divergence. Also [19] analyzes many functions of similarity such as KL, Jensen-Shannon (see [21]) and Skew divergences, Euclidean, cosine and L_1 (or Manhattan) distances, and also probability confusion, which estimates the substitutability of two given words, based on conditional and marginal probabilities. Table 14.1 displays several similarity functions for probability distributions used in text comparisons.

All the functions of Table 14.1 were used to analyze the data of the historical corpus. We fixed the value $\alpha = 0.7$ for the skew divergence. Parameter α controls the degree to which the function approximates $D(Q||P)$. The highest value yielded the best performance and very small values resulted in the worst error rates. Also in the function Confusion probability we consider the value of the similarity of words as $P(y')/P(x) = 0.5$ in order to fix the word comparisons. Reference [16] examine the two measures of discrepancy, that is, distances and divergences, where the intersection of them is the L_1 distance. These estimates are found directly and without separate estimates of each probability distribution through the Bregman scores method and semi-parametric statistical models. In addition, they find that the L_1 distance of the difference in densities is more robust than the density ratio. The difference in densities $p - q$ is used to calculate the distance L_s between two probability densities

$$d_s(p,q) = \left(\int |p(x) - q(x)|^s dx \right)^{\frac{1}{s}}$$

where $s \geq 1$.

The Numerical Example

The frequency of letters and key words in 9 paragraphs of Sect. 118-126 of corpus.. were analyzed. These paragraphs were selected because they counted the largest number of key words and represented the paragraphs of greatest cultural message.

Table 14.2 Corpus I (Lazo-1649) versus Corpus II (Rojas-1978)

Letter	Frequency in Corpus I (%)	Frequency in Corpus II (%)	$p_{letter} \log\left(\frac{p_{letter}}{q_{letter}}\right)$	$q_{letter} \log\left(\frac{q_{letter}}{p_{letter}}\right)$
a	9.64	10.99	–0.00550442	0.006277879
b	0.10	0.10	7.75194E–06	–7.61113E–06
c	10.13	9.85	0.001231125	–0.001197139
d	0.10	0.10	7.75194E–06	–7.61113E–06
e	3.99	4.21	–0.000906534	0.000955195
g	0.10	0.10	7.75194E–06	–7.61113E–06
h	4.87	4.59	0.001250731	–0.001178891
i	16.75	16.63	0.000492459	–0.000489136
j	0.10	0.10	–1.12663E-05	1.15704E-05
l	6.43	6.21	0.000937741	–0.000906757
m	5.06	4.97	0.000403101	–0.000395779
n	9.93	9.75	0.000790698	–0.000776335
o	8.37	8.41	–0.000169402	0.000170193
p	2.04	2.01	0.000162791	–0.000159834
q	2.04	1.91	0.000596068	–0.000557372
t	7.40	7.36	0.000169028	–0.000168142
u	6.52	6.21	0.001378014	–0.001312595
x	1.95	1.91	0.000155039	–0.000152223
y	1.85	1.91	–0.000264837	0.000273712
z	2.53	2.58	–0.000213397	0.000217579
Total			$D(P\|\|Q) = 0.0005202$	$D(Q\|\|P) = 0.0005891$

Also, the Nahuatl alphabet was set with 20 letters that were the following: a, b, c, d, e, g, h, i, j, l, m, n, o, p, q, t, u, x, y, z. This was obtained by deducting and spelling the historical corpus and calculating the frequency of each of the letters for the nine paragraphs mentioned above. The translations of the key words were carried out with the GDN digital dictionary (Gran Nahuatl dictionary) with the http://www.gdn.unam.mx/termino/search. Kullback–Leibler divergence measure was calculated to demonstrate the asymmetry between the two probability functions. The purpose of this example is to illustrate how asymmetry naturally arises in linguistic studies. Table 14.2 displays the results for numerical experiment, where p_{letter} denotes empirical frequency of occurrence of the given *letter*, e.g. letter "a".

Acknowledgments The work was supported by project Fondecyt Proyecto Regular No. 1151441 and Project LIT-2016-1-SEE-023 modec. We acknowledge support of the Editors and the informative and insightful suggestions of Referee.

References

1. Ali, M.S., Silvey, S.D.: A general class of coefficients of divergence of one distribution from another. J. Royal Stat. Soc. Ser. B **28**, 131–140 (1966)
2. Baran, S., Sikolya, K., Stehlík, M.: On the optimal designs for prediction of Ornstein-Uhlenbeck sheets. Stat. Probab. Lett. **83**(6), 1580–1587 (2013)
3. Baran, S., Szak-Kocsis, C., Stehlík, M.: D-optimal design for complex Ornstein-Uhlenbeck processes. J. Stat. Plan. Inference **197**, 93–106 (2018)
4. Baran, S., Stehlík, M.: Optimal designs for parameters of shifted Ornstein-Uhlenbeck sheets measured on monotonic sets. Stat. Probab. Lett. **99**, 114–124 (2015)
5. Bigi, B.: Using Kullback–Leibler distance for text categorization. In: European Conference on Information Retrieval, pp. 305–319. Springer (2003)
6. Chiang, C.L., Conforti, P.M.: A survival model and estimation of time to tumor. Math. Biosci. **94**, 1–29 (1989)
7. Csiszár, I.: Information-type measures of difference of probability distributions and indirect observations. Studia Sci. Math. Hungarica **2**, 299–318 (1967)
8. Fang, K.T.: Uniform design: application of number-theoretic methods in experimental design. Probab. Stat. Bull. **1**, 56–97 (1978)
9. Fang, K.T., Kotz, S., Ng, K.W.: Symmetric Multivariate and Related Distributions, Monographs on Statistics and Applied Probability, vol. 36. Springer-Science+Business Media (1990). https://doi.org/10.1007/978-1-4899-2937-2
10. Fang, K.T., Anderson, T.W.: Statistical Inference in Elliptical Contoured and Related Distributions, pp. 127–136. Allerton Press, New York (1990)
11. Fang, H.B., Fang, K.T., Kotz, S.: The Meta-elliptical distributions with given marginals. J. Multivar. Anal. **82**, 1–16 (2002)
12. Filus, J., Filus, L., Lu, Y., Jordanova, P., Anrold, B.C., Soza, L.N., Stehlíková, S., Stehlík, M.: On parameter dependence and related topics: the impact of Jerzy Filus from genesis to recent developments (with discussion). In: Vonta, I., Ram, M. (eds.) Reliability Engineering: Theory and Applications, 1st edn., pp. 143–169. CRC Press (2019)
13. Filus, J., Filus, L.: On new multivariate probability distributions and stochastic processes with system reliability and maintenence applications. Method Comput. Appl. Probab. **9**, 425–446 (2007)
14. Henschel, V., Richter, W.D.: Geometric generalization of the exponential law. J. Mult. Anal. **81**, 189–204 (2002)
15. Illert, C.R.: Origins of linguistic zonation in the Australian Alps, part 1. Huygens Principle. J. Appl. Stat. **32**(6), 625–659 (2005)
16. Kanamori, T., Sugiyama, M.: Statistical analysis of distance estimators with density differences and density ratios. Entropy **16**, 921–942 (2014)
17. Kise**ν**lák, J., Stehlík, M.: Equidistant D-optimal designs for parameters of Ornstein-Uhlenbeck process. Stat. Probab. Lett. **78**, 1388–1396 (2008)
18. Lasso, L.: Huey tlamahuizoltica omonexiti ilhuicac tlatoca ihwapilli Sancta Maria. Print Juan Ruyz, Mexico (1649)
19. Lee, L.: On the effectiveness of the skew divergence for statistical language analysis. In: Lee, L. (ed.) Proceedings of Artificial Intelligence and Statistics (AISTATS), pp. 65–72 (2001)
20. Liese, F., Vajda, I.: On divergences and informations in statistics and information theory. IEEE Trans. Inform. Theory **52**(10), 4394–4412 (2006)
21. Lin, J.: Divergence measures based on the Shannon entropy. IEEE Trans. Inform. Theory **37**(1), 145–151 (1991)
22. Müller, W.G., Stehlík, M.: Issues in the optimal design of computer simulation experiments. Appl. Stoch. Models Bus. Ind. **25**, 163–177 (2009)
23. Pázman, A.: Information contained in design points of experiments with correlated observations. Kybernetika **46**(4), 771–783 (2010)

24. Pigoli, D., Hadjipantelis, P.Z., Coleman, J.S., Aston, J.A.D.: The statistical analysis of acoustic phonetic data: exploring differences between spoken Romance languages. Appl. Stat. **67**(4), 1–27 (2018)
25. Richter, W.D.: Laplace-Gauss integrals, Gaussian measure asymptotic behaviour and probabilities of moderate deviations. Z. Anal. Anw. **4**(3), 257–267 (1985)
26. Richter, W.D.: A geometric method in stochastics (in German). Rostock. Math. Kolloqu. **44**, 63–72 (1991)
27. Rojas, M.: Nican Mopohua. Print Ideal, Mexico (1978)
28. Smith, B.: Mereotopology: a theory of parts and boundaries. Data Knowl. Eng. **20**(3), 287–303 (1996)
29. Stehlík, M.: Topological conditions on covariance structures for regression problems. In: Proceedings of 6th St. Petersburg Workshop on Sim, pp. 377–382 (2009)
30. Stehlík, M.: Decompositions of information divergences: recent development, open problems and applications. In: AIP Conference Proceedings, vol. 1493 (2012). https://doi.org/10.1063/1.4765604
31. Stehlík, M., Economou, P., Kiselák, J., Richter, W.D.: Kullback-Leibler life time testing. Appl. Math. Comput. **240**, 122–139 (2014)
32. Stehlík, M.: On convergence of topological aggregation functions. Fuzzy Sets Syst. **287**, 48–56 (2016)
33. Stehlík, M., Helpersdorfer, C., Hermann, P., Supina, J., Grilo, L.M., Maidana, J.P., Fuders, F., Stehlíková, S.: Financial and risk modelling with semicontinuous covariances. Inform. Sci. **394–395C**, 246–272 (2017)
34. Stehlík, M., Hermann, P., Torres, S., Kiselak, J., Rivera, A.: On dynamics underlying variance of mass balance estimation in Chilean glaciers. Ecol. Complex. **31**, 149–164 (2017)

Chapter 15
Estimation of Covariance Matrix with ARMA Structure Through Quadratic Loss Function

Defei Zhang, Xiangzhao Cui, Chun Li, and Jianxin Pan

Abstract In this paper we propose a novel method to estimate the high-dimensional covariance matrix with an order-1 autoregressive moving average process, i.e. ARMA(1,1), through quadratic loss function. The ARMA(1,1) structure is a commonly used covariance structures in time series and multivariate analysis but involves unknown parameters including the variance and two correlation coefficients. We propose to use the quadratic loss function to measure the discrepancy between a given covariance matrix, such as the sample covariance matrix, and the underlying covariance matrix with ARMA(1,1) structure, so that the parameter estimates can be obtained by minimizing the discrepancy. Simulation studies and real data analysis show that the proposed method works well in estimating the covariance matrix with ARMA(1,1) structure even if the dimension is very high.

Keywords ARMA(1,1) structure · Covariance matrix · Quadratic loss function

15.1 Introduction

Covariance matrix estimation is a fundamental problem in multivariate analysis and time series. Especially, the estimation of high-dimensional covariance matrix is rather challenging. In the literature, many research works were proposed to tackle the problem, such as [1, 3, 8, 9] among many others. However, when the covariance matrix has a certain of structures like order-1 autoregressive moving average, i.e. ARMA(1,1) structure or others, the estimation and regularization were hardly [6]. Recently, Lin et al. [7] proposed a new method to estimate and regularize the high-dimensional covariance matrix. Their idea is summarized as follows. Suppose A is a given $m \times m$ covariance matrix, that is, it is symmetric non-negative definite.

D. Zhang · X. Cui · C. Li
Department of Mathematics, Honghe University, Mengzi 661199, China

J. Pan (✉)
Department of Mathematics, University of Manchester, Manchester M13 9PL, UK
e-mail: jianxin.pan@manchester.ac.uk

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_15

Let $\mathscr{S}$ be the set of all $m \times m$ positive definite covariance matrices with structure s, for example, compound symmetry, uniform covariance structure or AR(1). A discrepancy between the given covariance matrix A and the set $\mathscr{S}$ is defined by

$$D(A, \mathscr{S}) = \min_{B \in \mathscr{S}} L(A, B),$$

where $L(A, B)$ is a measure of the discrepancy between the two $m \times m$ matrices A and B. Assume there is a given class of k candidate covariance structures $\{s_1, s_2, \ldots, s_k\}$. Let S_i be the set of all covariance matrices with structure s_i. Denote the set of $m \times m$ covariance matrices with the likely structures by $\Omega = \cup_{i=1}^{k} S_i$. The discrepancy between a given covariance matrix A and the set Ω is then defined by $D(A, \Omega) = \min_{B \in \Omega} L(A, B)$. The point is that, in this set Ω, the structure with which A has the smallest discrepancy can be viewed as the most likely underlying structure behind A, and the minimizer B with this particular structure is considered to be the regularized covariance matrix of A. Obviously, the bigger the class of candidate structures the better the approximation B to the underlying covariance matrix that is estimated by A. The discrepancy considered by [7] is the so-called entropy loss function and the class of the candidates of potential covariance structures they considered include order-1 moving average MA(1), compound symmetry, AR(1) and Toeplitz structures.

Motivated by this, in this paper we focus on the ARMA(1,1) covariance structure because it includes the MA(1), compound symmetry and AR(1) as its special cases. The ARMA(1,1) process is obtained by applying a recursive filter to the white noise, which is given by the model

$$X_t = \phi_1 X_{t-1} + \varepsilon_t + \theta_1 \varepsilon_{t-1} \quad (t = 1, \ldots, m),$$

where ϕ_1 and θ_1 both are parameters, and ε_t is a zero mean white noise process with variance σ_1^2. The covariance matrix of the ARMA(1,1) process (e.g., [2]) is given by

$$\Sigma(\sigma_1, \phi_1, \theta_1) = \frac{(1 + \theta_1^2 + 2\phi_1\theta_1)\sigma_1^2}{1 - \phi_1^2} \begin{bmatrix} 1 & a & a\phi_1 & a\phi_1^2 & \cdots & a\phi_1^{m-2} \\ a & 1 & a & a\phi_1 & \cdots & a\phi_1^{m-3} \\ a\phi_1 & a & 1 & a & \cdots & a\phi_1^{m-4} \\ a\phi_1^2 & a\phi_1 & a & 1 & \cdots & a\phi_1^{m-5} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ a\phi_1^{m-2} & a\phi_1^{m-3} & a\phi_1^{m-4} & \cdots & \cdots & 1 \end{bmatrix}, \tag{15.1}$$

where

$$a := a(\phi_1, \theta_1) = \frac{(1 + \phi_1\theta_1)(\phi_1 + \theta_1)}{1 + \theta_1^2 + 2\phi_1\theta_1}.$$

For simplicity, the covariance matrix in (15.1) can be written as

$$B(\sigma, c, \rho) = \sigma^2 \begin{bmatrix} 1 & c & c\rho & c\rho^2 & \cdots & c\rho^{m-2} \\ c & 1 & c & c\rho & \cdots & c\rho^{m-3} \\ c\rho & c & 1 & c & \cdots & c\rho^{m-4} \\ c\rho^2 & c\rho & c & 1 & \cdots & c\rho^{m-5} \\ \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ c\rho^{m-2} & c\rho^{m-3} & c\rho^{m-4} & \cdots & \cdots & 1 \end{bmatrix}. \tag{15.2}$$

where

$$\sigma^2 = \frac{(1 + \theta_1^2 + 2\phi_1\theta_1)\sigma_1^2}{1 - \phi_1^2}, \quad c = \frac{(1 + \phi_1\theta_1)(\phi_1 + \theta_1)}{1 + \theta_1^2 + 2\phi_1\theta_1} \text{ and } \rho = \phi_1$$

It is clear that there are three special cases for the ARMA covariance matrix (15.2). When $\rho = 0$, the structure (15.2) becomes the MA(1) covariance matrix, namely

$$B(c, \sigma) = \sigma^2 \begin{bmatrix} 1 & c & 0 & \cdots & 0 \\ c & 1 & c & \ddots & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & 1 & c \\ 0 & \cdots & 0 & c & 1 \end{bmatrix}_{m \times m}, \tag{15.3}$$

where $\sigma^2 > 0$ and $-1/\cos(\pi/(m+1)) < c < 1/\cos(\pi/(m+1))$. When $\rho = 1$, it reduces to the compound symmetry structure as

$$B(c, \sigma) = \sigma^2 \begin{bmatrix} 1 & c & c & \cdots & c \\ c & 1 & c & \ddots & \vdots \\ c & \ddots & \ddots & \ddots & c \\ \vdots & \ddots & \ddots & 1 & c \\ c & \cdots & c & c & 1 \end{bmatrix}_{m \times m},$$

where $\sigma^2 > 0$ and $-1/(p-1) < c < 1$ ensure the positive definiteness of the covariance matrix. When $\rho = c$, the structure (15.2) becomes the AR(1) covariance matrix, that is

$$B(c, \sigma) = \sigma^2 \begin{bmatrix} 1 & c & c^2 & \cdots & c^{m-1} \\ c & 1 & c & \cdots & c^{m-2} \\ c^2 & c & 1 & \cdots & c^{m-3} \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ c^{m-1} & c^{m-2} & \cdots & c & 1 \end{bmatrix}_{m \times m}, \tag{15.4}$$

where $\sigma^2 > 0$ and $-1 < c < 1$.

On the other hand, we choose the quadratic loss function rather than the entropy loss function to measure the discrepancy between two matrices. The quadratic loss function was considered by many authors including [3, 9] when estimating covariance matrix. Comparing the entropy loss function, the quadratic loss function avoids the direct calculation of eigenvalues for a likely large covariance matrix with ARMA(1,1) structure. The problem here is that for a given high-dimensional covariance matrix A we aim to find the matrix B with ARMA(1,1) structure such that the discrepancy between A and B is minimized in the domain of the parameters (σ^2, c, ρ). The resulting matrix B is considered to be an approximation to the unknown underlying covariance matrix behind A in terms of structure. The rest of this paper is organized as follows. In Sect. 15.2, we discuss the estimation process under the quadratic loss function and obtain the analytical estimation results. Simulation studies and real data analysis are considered in Sect. 15.3. Conclusions and remarks are provided in Sect. 15.4.

15.2 Estimation Process

We rewrite the covariance matrix of the ARMA(1,1) model as follows,

$$B(c, \rho, \sigma) = \sigma^2 \left(I + c \sum_{i=1}^{m-1} \rho^{i-1} T_i \right), \tag{15.5}$$

where T_i $(1 \leq i \leq m-1)$ is a symmetric matrix with the ith superdiagonal and subdiagonal elements equal to 1 and zeros elsewhere.

As explained in Sect. 15.1, we propose to use the following quadratic loss function

$$L(\Sigma, B) = tr\left(\Sigma^{-1} B - I_m\right)^2 \tag{15.6}$$

to measure the discrepancy between the matrices Σ and B [4, 10]. Our aim is to find the matrix B^* such that

$$L(\Sigma, B^*) = \min_{\{\sigma, c, \rho\} \in R^+ \times [-1,1]^2} L(\Sigma, B)$$

for the underlying population covariance matrix Σ, where $L(\Sigma, B)$ is the function in (15.6). In general, Σ is unknown but can be estimated by an available matrix A such as the sample covariance matrix. Hence, in practice we actually calculate $L(A, B)$ by replacing Σ with A.

Now let $x_0 = \sigma^2$ and $x_i = \sigma^2 c \rho^{i-1}$, $i = 1 : m-1$. The matrix B in (15.5) can be rewritten as

$$B(x) = \sum_{i=0}^{m-1} x_i T_i,$$

where $x = [x_0, x_1, \ldots, x_{m-1}]^T \in \mathbb{R}^m$, $T_0 = I$ and T_i's $(1 \le i \le m-1)$ are already defined in (15.5). We define the set $\Omega \subset \mathbb{R}^m$ by

$$\Omega := \left\{ x \in \mathbb{R}^m : B(x) = \sum_{i=0}^{m-1} x_i T_i \text{ is positive definite} \right\} \tag{15.7}$$

and the function $f(x) : \mathbb{R}^m \mapsto \mathbb{R}$,

$$f(x) := L(\Sigma, B(x)) = tr(\Sigma^{-1}B(x) - I_m)^2.$$

Since Ω is isomorphic to the set of all positive definite matrices, the problem now reduces to minimize the function $f(B)$ over the positive definite matrices B within the set Ω in (15.7).

Since $f(B) := L(\Sigma, B)$ is a strictly convex function of B and $B(x) = \sum_{i=0}^{m-1} x_i T_i$ is an affine map of x, by the fact that a composition with an affine mapping preserves convexity, then function $f(x) := f(B(x))$ is then strictly convex in x. On the other hand, since $\nabla_{x_i} B = T_i$, by applying the chain rule [4, 10] we obtain the gradient of f as

$$\nabla_{x_i} f = 2tr(T_i(\Sigma^{-1}B - I_m)\Sigma^{-1}), \quad i = 0 : m-1,$$

and the Hessian $H = [h_{ij}] \in \mathbb{R}^{m\times m}$ of f where

$$h_{ij} = \nabla^2_{x_i x_j} f = 2tr(T_i \Sigma^{-1} T_j \Sigma^{-1}), \quad i, j = 0 : m-1.$$

Therefore, this is a convex optimization problem so that the function f has a unique minimizer.

The loss function can be now expressed as

$$\begin{aligned} & f(\sigma, c, \rho) \\ & = tr\left(\Sigma^{-1}B - I_m\right)^2 \\ & = \sigma^4 tr\left(\Sigma^{-1} + c\sum_{i=1}^{m-1} \rho^{i-1}\Sigma^{-1}T_i\right)^2 - 2\sigma^2 tr\left(\Sigma^{-1} + c\sum_{i=1}^{m-1} \rho^{i-1}\Sigma^{-1}T_i\right) + m, \end{aligned}$$

where

$$\begin{aligned} & tr\left(\Sigma^{-1} + c\sum_{i=1}^{m-1} \rho^{i-1}\Sigma^{-1}T_i\right)^2 \\ & = c\sum_{i=1}^{m-1} \rho^{i-1} tr(\Sigma^{-2}T_i + \Sigma^{-1}T_i\Sigma^{-1}) + c^2 \sum_{i=1}^{m-1} \rho^{2(i-1)} tr\left((\Sigma^{-1}T_i)^2\right) \\ & \quad + tr(\Sigma^{-2}) + c^2 \sum_{i,j=1, i\neq j}^{m-1} \rho^{i+j-2} tr\left(\Sigma^{-1}T_i\Sigma^{-1}T_j + \Sigma^{-1}T_j\Sigma^{-1}T_i\right). \end{aligned}$$

Therefore, we have

$$
\begin{aligned}
& f(\sigma, c, \rho) \\
&= \sigma^4 tr(\Sigma^{-2}) + c\sigma^4 \sum_{i=1}^{m-1} \rho^{i-1} tr(\Sigma^{-2} T_i + \Sigma^{-1} T_i \Sigma^{-1}) + \sigma^4 c^2 \sum_{i=1}^{m-1} \rho^{2(i-1)} tr\left((\Sigma^{-1} T_i)^2\right) \\
&\quad + \sigma^4 c^2 \sum_{i,j=1, i\neq j}^{m-1} \rho^{i+j-2} tr\left(\Sigma^{-1} T_i \Sigma^{-1} T_j + \Sigma^{-1} T_j \Sigma^{-1} T_i\right) \\
&\quad - 2\sigma^2 tr(\Sigma^{-1}) - 2\sigma^2 c \sum_{i=1}^{m-1} \rho^{i-1} tr(\Sigma^{-1} T_i) + m \\
&= \sigma^4 tr(\Sigma^{-2}) + c\sigma^4 \sum_{i=1}^{m-1} \rho^{i-1} t_i^{(1)} + \sigma^4 c^2 \sum_{i=1}^{m-1} \rho^{2i-2} t_i^{(2)} \\
&\quad + \sigma^4 c^2 \sum_{i,j=1, i\neq j}^{m-1} \rho^{i+j-2} t_{ij}^{(3)} - 2\sigma^2 tr(\Sigma^{-1}) - 2\sigma^2 c \sum_{i=1}^{m-1} \rho^{i-1} t_i^{(4)} + m
\end{aligned}
$$

where $t_i^{(1)} := tr(\Sigma^{-2} T_i + \Sigma^{-1} T_i \Sigma^{-1})$, $t_i^{(2)} := tr((\Sigma^{-1} T_i)^2)$, $t_{ij}^{(3)} := tr\left(\Sigma^{-1} T_i \Sigma^{-1} T_j + \Sigma^{-1} T_j \Sigma^{-1} T_i\right)$, $t_i^{(4)} := tr(\Sigma^{-1} T_i)$.

Note that the first order partial derivative for $f(\sigma, c, \rho)$ is

$$
\nabla f(\sigma, c, \rho) := \begin{bmatrix} \frac{\partial f}{\partial \sigma} \\ \frac{\partial f}{\partial c} \\ \frac{\partial f}{\partial \rho} \end{bmatrix},
$$

where

$$
\begin{aligned}
\frac{\partial f}{\partial \sigma} &:= 4\sigma^3 tr(\Sigma^{-2}) + 4c\sigma^3 \sum_{i=1}^{m-1} \rho^{i-1} t_i^{(1)} + 4\sigma^3 c^2 \sum_{i=1}^{m-1} \rho^{2i-2} t_i^{(2)} \\
&\quad + 4\sigma^3 c^2 \sum_{i,j=1, i\neq j}^{m-1} \rho^{i+j-2} t_{ij}^{(3)} - 4\sigma tr(\Sigma^{-1}) - 4\sigma c \sum_{i=1}^{m-1} \rho^{i-1} t_i^{(4)}, \\
\frac{\partial f}{\partial c} &:= \sigma^4 \sum_{i=1}^{m-1} \rho^{i-1} t_i^{(1)} + 2\sigma^4 c \sum_{i=1}^{m-1} \rho^{2i-2} t_i^{(2)} + 2\sigma^4 c \sum_{i,j=1, i\neq j}^{m-1} \rho^{i+j-2} t_{ij}^{(3)} - 2\sigma^2 \sum_{i=1}^{m-1} \rho^{i-1} t_i^{(4)}, \\
\frac{\partial f}{\partial \rho} &:= c\sigma^4 \sum_{i=1}^{m-1} (i-1) \rho^{i-2} t_i^{(1)} + \sigma^4 c^2 \sum_{i=1}^{m-1} (2i-2) \rho^{2i-3} t_i^{(2)} \\
&\quad + \sigma^4 c^2 \sum_{i,j=1, i\neq j}^{m-1} (i+j-2) \rho^{i+j-3} t_{ij}^{(3)} - 2\sigma^2 c \sum_{i=1}^{m-1} (i-1) \rho^{i-2} t_i^{(4)}.
\end{aligned}
$$

Let $\nabla f(\sigma, c, \rho) = 0$. We then have the estimating equations for (σ^2, c, ρ) as follows,

$$\begin{cases} \sigma^2 tr(\Sigma^{-2}) + c\sigma^2 \sum_{i=1}^{m-1} \rho^{i-1} t_i^{(1)} + \sigma^2 c^2 \sum_{i=1}^{m-1} \rho^{2i-2} t_i^{(2)} + \sigma^2 c^2 \sum_{i,j=1, i\neq j}^{m-1} \rho^{i+j-2} t_{ij}^{(3)} \\ \quad = tr(\Sigma^{-1}) + c \sum_{i=1}^{m-1} \rho^{i-1} t_i^{(4)}, \\ \sigma^2 \sum_{i=1}^{m-1} \rho^{i-1} t_i^{(1)} + 2\sigma^2 c \sum_{i=1}^{m-1} \rho^{2i-2} t_i^{(2)} = -2\sigma^2 c \sum_{i,j=1, i\neq j}^{m-1} \rho^{i+j-2} t_{ij}^{(3)} + 2 \sum_{i=1}^{m-1} \rho^{i-1} t_i^{(4)}, \\ \sigma^2 \sum_{i=1}^{m-1} (i-1)\rho^{i-2} t_i^{(1)} + \sigma^2 c \sum_{i=1}^{m-1} (2i-2)\rho^{2i-3} t_i^{(2)} + \sigma^2 c \sum_{i,j=1, i\neq j}^{m-1} (i+j-2)\rho^{i+j-3} t_{ij}^{(3)} \\ \quad = 2 \sum_{i=1}^{m-1} (i-1)\rho^{i-2} t_i^{(4)}. \end{cases}$$

The Hessian matrix are given by

$$\nabla^2 f := \begin{bmatrix} \frac{\partial^2 f}{\partial \rho^2} & \frac{\partial^2 f}{\partial \rho \partial c} & \frac{\partial^2 f}{\partial \rho \partial \sigma} \\ \frac{\partial^2 f}{\partial c \partial \rho} & \frac{\partial^2 f}{\partial c^2} & \frac{\partial^2 f}{\partial c \partial \sigma} \\ \frac{\partial^2 f}{\partial \sigma \partial \rho} & \frac{\partial^2 f}{\partial \sigma \partial c} & \frac{\partial^2 f}{\partial \sigma^2} \end{bmatrix},$$

where

$$\begin{aligned} \frac{\partial^2 f}{\partial \rho^2} &:= c\sigma^4 \sum_{i=2}^{m-1} (i-1)(i-2)\rho^{i-3} t_i^{(1)} + \sigma^4 c^2 \sum_{i=1}^{m-1} (2i-2)(2i-3)\rho^{2i-4} t_i^{(2)} \\ &\quad + \sigma^4 c^2 \sum_{i,j=1, i\neq j}^{m-1} (i+j-2)(i+j-3)\rho^{i+j-4} t_{ij}^{(3)} - 2\sigma^2 c \sum_{i=2}^{m-1} (i-1)(i-2)\rho^{i-3} t_i^{(4)}, \\ \frac{\partial^2 f}{\partial \rho \partial c} &:= \sigma^4 \sum_{i=1}^{m-1} (i-1)\rho^{i-2} t_i^{(1)} + 2\sigma^4 c \sum_{i=1}^{m-1} (2i-2)\rho^{2i-3} t_i^{(2)} \\ &\quad + 2c\sigma^4 \sum_{i,j=1, i\neq j}^{m-1} (i+j-2)\rho^{i+j-3} t_{ij}^{(3)} - 2\sigma^2 \sum_{i=1}^{m-1} (i-1)\rho^{i-2} t_i^{(4)}. \\ \frac{\partial^2 f}{\partial \rho \partial \sigma} &:= 4c\sigma^3 \sum_{i=1}^{m-1} (i-1)\rho^{i-2} t_i^{(1)} + 4\sigma^3 c^2 \sum_{i=1}^{m-1} (2i-2)\rho^{2i-3} t_i^{(2)} \\ &\quad + 4\sigma^3 c^2 \sum_{i,j=1, i\neq j}^{m-1} (i+j-2)\rho^{i+j-3} t_{ij}^{(3)} - 4\sigma c \sum_{i=1}^{m-1} (i-1)\rho^{i-2} t_i^{(4)}. \\ \frac{\partial^2 f}{\partial c^2} &:= 2\sigma^4 \sum_{i=1}^{m-1} \rho^{2i-2} t_i^{(2)} + 2\sigma^4 \sum_{i,j=1, i\neq j}^{m-1} \rho^{i+j-2} t_{ij}^{(3)}, \end{aligned}$$

$$\frac{\partial^2 f}{\partial c \partial \sigma} := 4\sigma^3 \sum_{i=1}^{m-1} \rho^{i-1} t_i^{(1)} + 8\sigma^3 c \sum_{i=1}^{m-1} \rho^{2i-2} t_i^{(2)} + 8\sigma^3 c \sum_{i,j=1, i\neq j}^{m-1} \rho^{i+j-2} t_{ij}^{(3)} - 4\sigma \sum_{i=1}^{m-1} \rho^{i-1} t_i^{(4)},$$

$$\frac{\partial^2 f}{\partial \sigma^2} := 12\sigma^2 tr(\Sigma^{-2}) + 12c\sigma^2 \sum_{i=1}^{m-1} \rho^{i-1} t_i^{(1)} + 12\sigma^2 c^2 \sum_{i=1}^{m-1} \rho^{2i-2} t_i^{(2)} + 12\sigma^2 c^2 \sum_{i,j=1, i\neq j}^{m-1} \rho^{i+j-2} t_{ij}^{(3)} - 4tr(\Sigma^{-1}) - 4c \sum_{i=1}^{m-1} \rho^{i-1} t_i^{(4)}.$$

Our numerical results including simulation studies and real data analysis show that the determinant $|\nabla^2 f| > 0$ and the theoretical justification is still under investigation.

Theorem 15.1 *Given a positive definite covariance matrix Σ, there exists a unique positive definite matrix $B(\sigma, c, \rho)$ in the form (15.2) such that the quadratic loss function $L(\sigma, c, \rho) := L(\Sigma, B(\sigma, c, \rho))$ in (15.6) is minimized. Furthermore, the minimum must be attained at (σ, c, ρ) that satisfies*

$$\begin{cases} \sigma^2 tr(\Sigma^{-2}) + c\sigma^2 S_1(\rho) + c^2\sigma^2 S_2(\rho) + \sigma^2 c^2 S_3(\rho) = tr(\Sigma^{-1}) + cS_4(\rho), \\ \sigma^2 S_1(\rho) + 2c\sigma^2 S_2(\rho) + 2c\sigma^2 S_3(\rho) = 2S_4(\rho), \\ \sigma^2 S_1'(\rho) + c\sigma^2 S_2'(\rho) + c\sigma^2 S_3'(\rho) = 2S_4'(\rho), \end{cases}$$

where

$$S_1(\rho) := \sum_{i=1}^{m-1} \rho^{i-1} t_i^{(1)}, \qquad S_2(\rho) := \sum_{i=1}^{m-1} \rho^{2i-2} t_i^{(2)},$$

$$S_3(\rho) := \sum_{i,j=1, i\neq j}^{m-1} \rho^{i+j-2} t_{ij}^{(3)}, \qquad S_4(\rho) := \sum_{i=1}^{m-1} \rho^{i-1} t_i^{(4)},$$

and $S_i'(\rho)(i = 1, \ldots, 4)$ is the derivative of $S_i(\rho)$ with respect to ρ, $t_i^{(1)} := tr(\Sigma^{-2} T_i + \Sigma^{-1} T_i \Sigma^{-1})$, $t_i^{(2)} := tr((\Sigma^{-1} T_i)^2)$, $t_{ij}^{(3)} := tr\left(\Sigma^{-1} T_i \Sigma^{-1} T_j + \Sigma^{-1} T_j \Sigma^{-1} T_i\right)$, and $t_i^{(4)} := tr(\Sigma^{-1} T_i)$.

Corollary 15.1 *Given a positive definite covariance matrix Σ, there exists a unique tridiagonal positive definite matrix, i.e. MA(1), $B(c, \sigma)$ in the form (15.3) such that the quadratic loss function $L(c, \sigma) := L(\Sigma, B(\sigma, c))$ in (15.6) is minimized. Furthermore, the minimum must be attained at (σ, c) that satisfies*

$$\begin{cases} \sigma^2 = \dfrac{tr(\Sigma^{-1})tr(\Sigma^{-2}T_1^2) - tr(\Sigma^{-2}T_1)tr(\Sigma^{-1}T_1)}{tr(\Sigma^{-2})tr(\Sigma^{-2}T_1^2) - (tr(\Sigma^{-2}T_1))^2}, \\ c = \dfrac{tr(\Sigma^{-2})tr(\Sigma^{-1}T_1) - tr(\Sigma^{-1})tr(\Sigma^{-2}T_1)}{tr(\Sigma^{-2}T_1^2)tr(\Sigma^{-1}) - tr(\Sigma^{-2}T_1)tr(\Sigma^{-1}T_1)}. \end{cases}$$

Corollary 15.2 *Given a positive definite covariance matrix Σ, there exists a unique AR(1) positive definite matrix $B(c, \sigma)$ in the form (15.4) such that the quadratic loss function $L(c, \sigma) := L(\Sigma, B(\sigma, c))$ in (15.6) is minimized. Furthermore, the minimum must be attained at (σ, c) that satisfies*

$$\begin{cases} \sigma^2 = \dfrac{\sum_{i=0}^{m-1} c^i tr(\Sigma^{-1}T_i)}{\sum_{i=0}^{m-1} c^{2i} tr((\Sigma^{-1}T_i)^2) + 2\sum_{i=0}^{m-2} c^{2i+1} tr(\Sigma^{-1}T_i\Sigma^{-1}T_{i+1})}, \\ \dfrac{\sum_{i=0}^{m-1} ic^{i-1} tr(\Sigma^{-1}T_i)}{\sum_{i=0}^{m-1} c^i tr(\Sigma^{-1}T_i)} = \dfrac{\sum_{i=0}^{m-1} ic^{2i-1} tr((\Sigma^{-1}T_i)^2) + \sum_{i=0}^{m-2}(2i+1)c^{2i} tr(\Sigma^{-1}T_i\Sigma^{-1}T_{i+1})}{\sum_{i=0}^{m-1} c^{2i} tr((\Sigma^{-1}T_i)^2) + 2\sum_{i=0}^{m-2} c^{2i+1} tr(\Sigma^{-1}T_i\Sigma^{-1}T_{i+1})}. \end{cases}$$

Similar results for the compound symmetry structure can be obtained in the same manner but the details are omitted here.

15.3 Numerical Experiments

15.3.1 Simulation Studies

Let m be the dimension of the covariance matrices. We first generate an $m \times n$ data matrix R with columns randomly drawn from the multivariate normal distribution $\mathcal{N}(\mu, \Sigma)$ with a common mean vector $\mu = \sigma^2 e$ with $e = (1, ..., 1)' \in \mathbb{R}^m$ and a common covariance matrix Σ. We then calculate the sample covariance matrix A using the generated random samples R. We assume the true covariance matrix Σ is of ARMA(1,1) structure with dimension m and the parameters (σ^2, c, ρ). We assess the performance of the estimation method by varying dimension m and and values of σ^2, c and ρ. The sample size is chosen as $n = 1000$. We summarize the experimental results in Table 15.1, which is the experiment with the covariance matrix size $m = 100$, and Table 15.2 for $m = 200$. We choose $\sigma^2 \in \{2, 4, 8\}$, $c \in \{0.2, 0.5, 0.75\}$ and $\rho \in \{-0.75, -0.5, -0.2, 0, 0.2, 0.5, 0.75\}$, meaning that Σ may have MA(1), AR(1), and ARMA(1,1) structures, respectively. The notation and abbreviation for the results reported in Tables 15.1 and 15.2 are summarized as follows.

- Σ: the true covariance matrix.
- A: the sample covariance matrix.

- B: the estimated covariance matrix with ARMA(1,1) structure, which minimizes the quadratic loss function $L(A, B)$.
- $L_{\Sigma,A}$, $L_{A,B}$ and $L_{\Sigma,B}$: the quadratic loss functions $L(\Sigma, A)$, $L(A, B)$ and $L(\Sigma, B)$, respectively.

In Tables 15.1 and 15.2, we have the following observations.

(1) When the true covariance structure for Σ is of ARMA(1,1), the resulting matrix B that has the same structure as Σ must satisfy $L_{\Sigma,B} < L_{\Sigma,A}$. It means that the regularized estimator B is much better than the sample covariance matrix A in terms of the quadratic loss function. This is because the sample covariance matrix A contains many noises so that the true ARMA(1,1) structure is blurred if only A is observed. It shows that regularization of the sample covariance matrix A into a proper structure, here ARMA(1,1), is necessary not only for the convenient use of the structure but also for the accuracy of the covariance matrix estimation.
(2) The observations above are the same for differing values of m, σ^2, c and ρ, implying that the findings are consistent and robust against the parameters (σ^2, c, ρ).
(3) Note that it is extremely important to observe the discrepancy $L_{A,B}$ because in practice the true covariance Σ is unknown, and so $L_{\Sigma,B}$ and $L_{\Sigma,A}$ are not possibly known either. The simulation studies presented here aim to assess the performance of the approximation B to the underlying covariance matrix Σ by borrowing information from the sample covariance matrix A. It is concluded that the discrepancy $L_{A,B}$ can be used to identify the true covariance structure of Σ satisfactorily.

Table 15.1 Simulation results with $m = 100$

σ^2	c	ρ	$L_{\Sigma,A}$	$L_{A,B}$	$L_{\Sigma,B}$
2	0.2	−0.75	10.19	27.65	0.23
4	0.2	−0.75	10.22	31.48	0.31
8	0.2	−0.75	10.18	83.64	0.42
2	0.2	−0.5	10.27	34.05	0.96
2	0.5	−0.2	10.01	29.75	0.27
2	0.75	−0.2	9.7	25.03	0.55
2	0.2	0	10.31	26.03	0.25
4	0.5	0	10.01	29.37	0.61
8	0.75	0	10.19	69.48	0.72
2	0.2	0.2	10.19	29.11	0.06
4	0.5	0.5	9.71	29.03	0.93
8	0.75	0.75	10.01	79.43	0.96
2	0.2	−0.2	10.11	78.31	0.98
4	0.5	−0.5	10.03	33.94	0.36
8	0.75	−0.75	10.24	84.21	0.94

Table 15.2 Simulation results with $m = 200$

σ^2	c	ρ	$L_{\Sigma,A}$	$L_{A,B}$	$L_{\Sigma,B}$
2	0.2	−0.75	40.03	46.39	0.38
4	0.2	−0.75	40.56	69.75	0.51
8	0.2	−0.75	40.61	79.01	0.71
2	0.2	−0.5	39.84	72.02	0.31
2	0.5	−0.2	40.09	83.47	0.52
2	0.75	−0.2	39.84	73.57	0.61
2	0.2	0	40.09	84.29	0.54
4	0.5	0	40.69	94.29	0.63
8	0.75	0	40.47	106.44	0.85
2	0.2	0.2	40.02	161.18	0.62
4	0.5	0.5	39.86	83.29	0.72
8	0.75	0.75	39.92	187.27	0.89
2	0.2	−0.2	40.75	92.69	0.55
4	0.5	−0.5	39.36	142.44	0.62
8	0.75	−0.75	40.87	166.73	0.63

15.3.2 Real Data Analysis

15.3.2.1 Cattle Data Analysis

We analyze the Kenward's (1987) [5] cattle data using the proposed approach. The data set involves 60 cattle assigned randomly to two treatment groups 1 and 2, each of which consists of 30 cattle, and received a certain treatment. The cattle in each group were weighed 11 times over a nineteen-week period. The weighing times for all cattle were the same, so that the cattle data is a balanced longitudinal data set. The aim of Kenward's study was to investigate treatment effects on intestinal parasites of the cattle.

Our analysis was made for the cattle data in the same way as in Sect. 15.2 and the results are reported in Table 15.3. We also record, under the column named "Time" in Table 15.3, the time (in seconds) used to find the optimal matrix B for each structure of the possible candidates MA(1), AR(1) and ARMA(1,1).

Table 15.3 Results of experiments for Kenward's cattle data

	MA(1)		AR(1)		ARMA(1,1)	
	$L_{A,B}$	Time	$L_{A,B}$	Time	$L_{A,B}$	Time
Group 1	9.91	2.91	9.46	2.86	9.33	2.82
Group 2	9.53	2.90	9.63	2.76	9.52	2.79

Table 15.4 Results of experiments on Dental data

	MA(1)		AR(1)		ARMA(1,1)	
	$L_{A,B}$	Time	$L_{A,B}$	Time	$L_{A,B}$	Time
Girl group	2.68	0.25	3.43	0.22	2.62	0.21
Boy group	3.01	0.18	3.15	0.18	2.3	0.19

Since the true covariance matrix Σ from the cattle data is unknown, the discrepancies $L_{\Sigma,A}$ and $L_{\Sigma,B}$ are not available and then only the discrepancy $L_{A,B}$ is computed and presented in Table 15.3. From Table 15.3, it is clear that the underlying covariance structures are very likely to be ARMA(1,1) structure for both groups when comparing to other possible candidate structures MA(1) and AR(1), since their discrepancy $L_{A,B}$ has smaller values than other twos.

One may argue that Group 1 is likely to have an AR(1) covariance structure as the values of $L_{A,B}$ for AR(1) and ARMA(1,1) are very close. This should not be surprised because the AR(1) is a special case of the ARMA(1,1) in the sense that c is identical to ρ, see (15.4). This is the case for the Group 1 cattle data analysis due to the fact that the estimates of c and ρ are very close. This conclusion agrees with the finding in [11, 13, 15].

15.3.2.2 Dental Data Analysis

We also did an experiment with dental data (Potthoff and Roy 1964) [12]. Dental measurements were made on 11 girls and 16 boys at ages 8, 10, 12 and 14 years. Each measurement is the distance, in millimeters, from the center of the pituitary to the pterygomaxillary fissure. Similar to the cattle data analysis, the quadratic loss function $L_{A,B}$ is computed for the dental data and presented in Table 15.4.

From Table 15.4, it is clear that the underlying covariance structures are very likely to be ARMA(1,1) for both boy and girl groups, as the discrepancy values of $L_{A,B}$ are smaller than those for both MA(1) and AR(1) structures.

15.4 Conclusions

Motivated by the work of Lin et al. [7], we estimate the underlying covariance structure by minimizing the quadratic loss function between a given covariance matrix and the covariance matrix with ARMA(1,1) structure. Differing from their method, the quadratic loss function is used to replace the entropy loss function where the latter involves the calculation of eigenvalues for a likely large covariance matrix with ARMA(1,1) structure, which is challenging especially for high-dimensional case [14]. Our numerical results including simulation studies and real data analysis

show that the proposed method works well in estimating high-dimensional covariance matrices with an underlying ARMA(1,1) structure and is robust against various choices of the parameters involved in the ARMA(1,1) structure.

Acknowledgments This research is supported by the National Science Foundation of China (11761028 and 11871357). We acknowledge helpful comments and insightful suggestions made by a referee.

References

1. Fan, J., Fan, Y., Lv, J.: High dimensional covariance matrix estimation using a factor model. J. Econ. **147**(1), 186–197 (2008)
2. Francq, C.: Covariance matrix estimation for estimators of mixing weak ARMA models. J. Stat. Plan. Inference **83**(2), 369–394 (2000)
3. Haff, L.R.: Empirical bayes estimation of the multivariate normal covariance matrix. Ann. Statist. **8**(3), 586–597 (1980)
4. Horn, R.A., Johnson, C.R.: Matrix Analysis, 2nd edn. Cambridge University Press, Cambridge, UK (2013)
5. Kenward M. G.: A method for comparing profiles of repeated measurements. *Applied Statistics*, **36**(3), 296–308 (1987)
6. Lin, F. Jovanović, M.R.: Least-squares approximation of structured covariances, IEEE Trans. Automat. Control **54**(7), 1643–1648 (2009)
7. Lin, L., Higham, N.J., Pan, J.: Covariance structure regularization via entropy loss function. Comput. Stat. Data Anal. **72**(4), 315–327 (2014)
8. Ning, L., Jiang, X., Georgiou, T.: Geometric methods for structured covariance estimation. In: American Control Conference, pp. 1877–1882. IEEE (2012)
9. Olkin, I., Selliah, J.B.: Estimating covariance matrix in a multivariate normal distribution. In: Gupta, S.S., Moore, D.S. (eds.) Statistical Decision Theory and Related Topics, vol. II, pp. 313–326. Academic Press, New York (1977)
10. Pan, J., Fang, K.: Growth Curve Models and Statistical Diagnostics. Springer, New York (2002)
11. Pan, J., Mackenzie, G.: On modelling mean-covariance structures in longitudinal studies. Biometrika **90**(1), 239–244 (2003)
12. Potthoff R. F., Roy S. N.: A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika **51**(3-4), 313–326 (1964)
13. Pourahmadi M.: Joint mean–covariance models with applications to longitudinal data: unconstrained parameterisation. Biometrika **86**(3), 677–690 (1999)
14. Xiao, H., Wu, W.: Covariance matrix estimation for stationary time series. Ann. Stat. **40**(1), 466–493 (2012)
15. Ye, H., Pan, J.: Modelling of covariance structures in generalised estimating equations for longitudinal data. Biometrika **93**(4), 927–941 (2006)

Part IV
Data Mining

Chapter 16
Depth Importance in Precision Medicine (DIPM): A Tree and Forest Based Method

Victoria Chen and Heping Zhang

Abstract We propose the novel implementation of a depth variable importance score in a classification tree method designed for the precision medicine setting. The goal is to identify clinically meaningful subgroups to better inform personalized treatment decisions. In the proposed Depth Importance in Precision Medicine (DIPM) method, a random forest of trees is first constructed at each node. Then, a depth variable importance score is used to select the best split variable. This score makes use of the observation that more important variables tend to be selected closer to root nodes of trees. In particular, we aim to outperform an existing method designed for the analysis of high-dimensional data with continuous outcome variables. The existing method uses an importance score based on weighted misclassification of out-of-bag samples upon permutation. Overall, our method is favorable because of its comparable and sometimes superior performance, simpler importance score, and broader pool of candidate splits. We use simulations to demonstrate the accuracy of our method and apply the method to a clinical dataset.

16.1 Introduction

Improving the field of medicine using personalized health data has become a primary focus for researchers. Instead of the traditional focus on average responses to interventions, precision medicine recognizes the heterogeneity that exists between individuals and aims to find the optimal treatment for each person [7, 13]. With the increasing number of large datasets available for analysis, identifying which features are important is a challenge. Ultimately, the development of more sophisticated methodology to match the development of these kinds of data is important to

V. Chen · H. Zhang (✉)
Department of Biostatistics, Yale School of Public Health, New Haven, CT, USA
e-mail: heping.zhang@yale.edu

V. Chen
e-mail: victoria.chen@yale.edu

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_16

help improve the health outcomes and quality of life experienced by each individual patient.

The classification tree is an attractive method for precision medicine due to its flexibility and relatively simple structure. Multiple candidate features may be considered simultaneously, and the final result is an easily interpretable tree. In general, a classification tree is a method that divides the overall sample into smaller and smaller subgroups using optimized subdivisions of the data. The subdivisions, or splits, are based on a predetermined list of candidate split variables. Traditionally, classification trees are used to identify homogenous subgroups of the sample and classify each subject's membership in a predetermined list of categories. In the context of precision medicine, the method is modified to identify subgroups of patients that perform especially well or especially poorly in a treatment group and determine which treatment is best for each subject.

Currently, there are multiple existing tree-based methods designed for the precision medicine setting. Existing methods include: an extension of the RECursive Partition and Amalgamation (RECPAM) algorithm [12], model-based partitioning (MOB) [14, 19], interaction trees (IT) [15–17], the simultaneous threshold interaction modelling algorithm (STIMA) [4], virtual twins (VT) [6], subgroup identification based on differential effect search (SIDES) [9], an extension to SIDES known as SIDEScreen [8], qualitative interaction trees (QUINT) [5], generalized, unbiased, interaction detection and estimation (GUIDE) trees [10, 11], a relative-effectiveness based method [18, 20], and an importance index based method [22].

Although multiple methods already exist, the type of outcome as well as other features of the data determine which subset of methods the user may choose from. For instance, the method developed by Zhang et al. [20] only applies to clinical data with a binary outcome and two treatment groups. Meanwhile, IT, QUINT, STIMA, and the method developed by Zhu et al. [22] apply to data with a continuous outcome. In addition, RECPAM, IT, MOB, SIDES, GUIDE, and the method developed by Zhu et al. [22] have been extended to analyze survival data with right-censored survival times. To date, only IT and GUIDE have an extension for data with longitudinal outcomes. Furthermore, a problem across methods is weakened performance as the number of candidate covariates increases. As noted in Tsai et al. [18], having more candidate covariates decreases the "signal-to-noise ratio" which can lower the chance of finding the most important variables. These concerns are especially problematic given the increased availability of higher dimensional data.

One method of particular interest is the weighted classification tree developed by Zhu et al. [22]. This method aims to achieve better performance in cases of high dimensionality and is designed for data with a continuous outcome variable and two treatment groups. A variable importance score based on weighted misclassification is used to find the best split variable at each node. However, as no method uniformly outperforms all other methods in this setting, there are several drawbacks. In particular, we find that the weighted method's variable importance score misses important signals in the presence of correlations between variables and that the method is unnecessarily complex overall. Instead, we propose the usage of the depth variable importance score developed by Chen et al. [3], and Zhang and Singer [21]. Adapting

this measure for usage within a tree and within the precision medicine framework is novel. Here, we make the case that the proposed Depth Importance in Precision Medicine (DIPM) method is favorable to the aforementioned method because of the proposed method's comparable and sometimes superior performance, simpler importance score, and broader pool of candidate splits. Developing an importance score that is intuitive and convenient to compute that yields comparable or even better results will set the stage for outright superior performance with more complex data scenarios to be demonstrated in future work. The overall goal is to identify variables that are important in the context of precision medicine. Note that the proposed method is an exploratory method as opposed to a confirmatory model. Thus, here we focus on introducing our new importance score and demonstrate its advantage by using datasets with continuous outcome variables for the easy comparison with an existing method.

The remainder of this paper is structured as follows. First, details of the proposed DIPM method and the weighted classification tree method are provided. Then, simulation scenarios assessing and comparing the methods are presented. Next, results of an application to a real-world dataset are described. Lastly, the discussion section includes closing remarks and directions for future work.

16.2 Methods

16.2.1 Overview

We begin with a brief overview of our method. The proposed DIPM method is designed for the analysis of clinical datasets with a continuous outcome variable Y and two treatment assignments A and B. Without loss of generality, higher values of Y denote better health outcomes. Candidate split variables may be binary, ordinal, or nominal. All of the learning data are said to be in the first or root node, and nodes may be split into two child nodes. Borrowing the terminology used in Zhu et al. [22], at each node in the tree, a random forest of "embedded" trees is grown to determine the best variable to split the node. Once the best variable is identified, the best split of the best variable is the split that maximizes the difference in response rates between treatments A and B. Note that "the best variable" is "best" in a narrow sense as defined below. In addition, a flowchart outlining the general steps of our method's algorithm is provided in Fig. 16.1.

16.2.2 Depth Variable Importance Score

The depth variable importance score is used to find the best split variable at a node. In general, the score incorporates two pieces of information: the depth of a node within

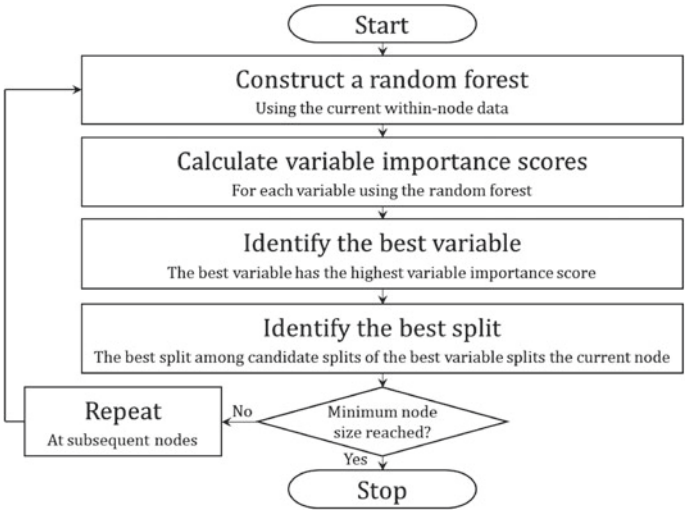


Fig. 16.1 Overview of DIPM method classification tree algorithm. A flowchart outlining the general steps of the proposed method's algorithm is depicted in the figure above

a tree and the magnitude of the relevant effect. Depth information is used because more important variables tend to be selected closer to the root node. Meanwhile, the strength of a split is also taken into account. This second component of the variable importance score is a statistic. The statistic that is used depends on the context of the analysis at hand.

Recall that at each node in the overall classification tree, a random forest is constructed to find the best split variable at the node. Once the forest is fit, for each tree T in this forest, the following sum is calculated for each covariate j:

$$score(T, j) = \sum_{t \in T_j} 2^{-L(t)} G_t. \tag{16.1}$$

T_j is the set of nodes in tree T split by variable j. $L(t)$ is the depth of node t. The root node has depth 1, the left and right child nodes of the root node have depth 2, etc. G_t captures the magnitude of the effect of splitting node t. Since the outcome is continuous, G_t is set equal to the t^2 statistic from testing the significance of β_3 in the model:

$$Y = \beta_0 + \beta_1 * treat + \beta_2 * split + \beta_3 * treat * split + \varepsilon. \tag{16.2}$$

This model is fit using the pertinent within-node data. The test statistic t is squared because the magnitude of the interaction is of interest, while there is no preference in the effect's direction. Note that this t^2 statistic is identical to the statistic used at each node split in the interaction tree method [16].

Next, a "G replacement" feature is implemented that potentially alters the variable importance scores $score(T, j)$. For each tree T in the forest, the G at each split is replaced with the highest G value of any of its descendant nodes if this maximum exceeds the value at the current split. This replacement step is performed because a variable that yields a split with a large effect of interest further down in the tree is certainly important even if its importance is not captured right away. By "looking ahead" at the G values of future splits, a variable's importance is reinforced.

Lastly, the final variable importance scores are averaged across all M trees in the forest f:

$$score(f, j) = \frac{1}{M} \sum_{T \in f} score(T, j). \tag{16.3}$$

The best split variable is the variable with the largest value of $score(f, j)$.

16.2.3 Split Criteria

To identify the best split at a node t, the squared difference in response rates between treatments A and B at node t is first assessed:

$$DIFF(t) = (\bar{Y}_{A,t} - \bar{Y}_{B,t})^2. \tag{16.4}$$

Then, among the list of candidate splits, only splits with child nodes with at least $nmin$ subjects are considered. Of the splits with a sufficient number of subjects, the best split maximizes the weighted sum of the squared difference in response rates of the child nodes:

$$DIFF(t_L, t_R) = \frac{\sum_{s=\{L,R\}} n_s (\bar{Y}_{A,t_s} - \bar{Y}_{B,t_s})^2}{n_L + n_R}. \tag{16.5}$$

Node t is split only when the best split yields two child nodes with a greater difference in treatment response rates than at the current node:

$$DIFF(t_L, t_R) > DIFF(t). \tag{16.6}$$

Splitting stops when there are not enough subjects in any candidate node splits or when no remaining $DIFF(t_L, t_R)$ values exceed $DIFF(t)$.

This split criterion was first proposed by Zhang for data with binary outcomes [18, 20]. Since the proposed method uses continuous outcomes, the mean of Y is used in place of $Pr(Y = 1)$.

16.2.4 Random Forest

A random forest is grown at each node in the overall tree and then used to select the best split variable. Once this variable is identified, all possible splits of the variable are considered, and the best split is found using the criteria described in Sect. 16.2.3.

The forest is constructed as follows. The forest contains a total of M embedded trees, and the recommended value of M is 1000. Each embedded tree is grown using a bootstrap sample. The bootstrap sample contains the same number of subjects as the original sample size at the current node. Then, at each node in the embedded trees, either: (1) all possible splits of all variables are considered, or (2) all possible splits of a certain number, $mtry$, of randomly selected variables are considered. The best split is again found using the criteria described in Sect. 16.2.3.

A recommended value of $mtry$ for a dataset with p variables is $floor(\sqrt{p})$. This value is the default value of $mtry$ used in the `randomForest R` package implementing Breiman's random forest method for classification. The aim is to use a value that balances the strength of each tree by being large enough while minimizing the correlation between trees by being small enough [2].

Also, note that the minimum number of subjects in nodes of the overall classification tree does not have to equal the minimum number of subjects in nodes of the embedded trees. Put another way, $nmin$ is the minimum node size of the overall tree, while $nmin2$ is the minimum node size of trees in the random forest. $nmin$ and $nmin2$ do not have to be equivalent.

16.2.5 Best Predicted Treatment Class

The best predicted treatment class of a node is the treatment group that performs best based on the subjects within the given node. In the proposed method, the means of the response values Y are compared by treatment group. Recall that higher values of Y denote greater benefit for patients. Therefore, if $\bar{Y}_A > \bar{Y}_B$ within a node, then treatment A is the best predicted treatment at that node. If $\bar{Y}_B > \bar{Y}_A$, then treatment B is the best predicted treatment. If $\bar{Y}_A = \bar{Y}_B$, then neither treatment is best.

16.2.6 Splits by Variable Type

The list of possible splits for a candidate split variable depends on the variable's type. For a binary variable, the variable has only two possible values: 0 and 1. Therefore, there is only one possible split: the left child node subsets the data with subjects whose values equal 0, and the right child node contains the rest of the subjects whose values equal 1.

For an ordinal variable, each unique value is a candidate split point. For each candidate split point s, the left child node considers the subjects with values less than or equal to s, and the right child node contains the rest of the subjects with values greater than s. Note that considering the largest unique value is redundant, since every subject takes values less than or equal to the maximum, and no one has values exceeding the maximum.

For a nominal variable, all combinations of all possible subsets of the categories are considered as candidate splits. For example, consider a nominal variable with three categories A, B, and C. One possible split is that the left child node subsets the data with subjects in category A, and the right child node contains the remaining subjects in categories B and C. The two other possible splits are A, B with complement C, and A, C with complement B. In general, for a nominal variable with k categories, the total number of possible splits is $2^{k-1} - 1$ [21].

16.2.7 Comparison Method

The weighted classification tree method by Zhu et al. [22] also uses a forest of M embedded trees to find the best split variable at each node. Again, once the best split variable is found, the best split of all possible splits of that variable is used. However, the forest used in the weighted classification tree method is a forest of extremely randomized trees. These trees select one random split for each variable. The best of these splits is used to split a node, and the split criteria is a weighted Gini impurity score. In addition, each tree uses bootstrap samples that consist of randomly sampling 80% of the node data without replacement.

Before the overall classification tree is constructed, mean estimates for each subject are predicted using a random forest of regression trees. These estimates are used to: (1) construct subject specific weights to be used when calculating the variable importance scores, and (2) perform "treatment flipping". One way to construct the weights is to take the absolute value of the difference between outcome variable Y and the estimated mean for each subject. Next, if a subject's Y value is smaller than the subject's estimated mean, then that subject is placed in the other treatment group. In other words, the treatment is "flipped". Note that treatment flipping does not affect the best predicted treatment at any terminal node of the tree and is done to solve the problem of greater bias for splits near the boundary of a variable.

Once a forest f containing M trees is fit at a node, a weighted variable importance score is calculated for each variable j to find the best split variable. This importance score uses the out-of-bag (OOB) samples at a node to calculate the weighted ratio of misclassified treatments when values are randomly permuted to the amount of misclassification when values are left the same:

$$score^{*}_{cla}(f, j) = \frac{1}{M} \sum_{m=1}^{M} \frac{\sum_{i \in L_{m,o}} w_i I_{\left(A_i \neq \widehat{f}_m(\mathbf{x}_i^{(-j)}, \tilde{x}_i^{(j)})\right)}}{\sum_{i \in L_{m,o}} w_i I_{\left(A_i \neq \widehat{f}_m(\mathbf{x}_i^{(-j)}, x_i^{(j)})\right)}} - 1. \tag{16.7}$$

$\widehat{f}_m$ denotes the predicted best treatment classes of the mth tree in the forest. $L_{m,o}$ is the OOB data for the mth tree. For the OOB samples, w_i is the ith subject's weight, A_i is the ith subject's treatment assignment after treatment flipping, $\mathbf{x}_i^{(-j)}$ is the ith subject's vector of data without variable j, $x_i^{(j)}$ is the ith subject's value of variable j, and $\tilde{x}_i^{(j)}$ is an independent, randomly permuted copy of variable j. I is the indicator function. The best split variable is the variable with the largest value of $score^*_{cla}(f, j)$.

16.2.8 Implementation

The proposed method is implemented using an `R` program. The `R` code calls a `C` program to generate the final classification tree. The `C` backend is used to take advantage of `C`'s higher computational speed in comparison to `R`. Meanwhile, the weighted classification method developed by Zhu et al. [22] is implemented using their `RLT` package on `CRAN`. All computations for the simulation studies and data analysis are implemented in `R`.

16.3 Simulation Studies

16.3.1 Methods

In addition to the weighted classification tree and proposed DIPM methods, two other methods are compared in our simulation studies. These additional methods do not use a random forest at each node. Instead, the additional methods are tree methods that consider all possible splits of all candidate variables at each node. One of these methods uses the weighted Gini impurity score to compare all splits, while the other uses the "DIFF" score described in Sect. 16.2.3. These methods act as controls to further study the effect of using a broader pool of candidate splits.

16.3.2 Scenarios

The following scenarios assess the proposed DIPM method and compare it to the weighted classification tree method. The overall strategy is to design scenarios with known, underlying signals and then measure how often each method accurately detects these signals. This strategy allows us to compare the variable importance scores of the weighted and proposed methods. Recall that the DIPM method is an exploratory method as opposed to a confirmatory model, and the primary goal

is to identify important variables in the context of precision medicine. Therefore, measuring correct variable selection alone is sufficient.

In particular, the simulations are designed to assess how each method performs with increasing amounts of correlation. Altogether, we expect each method to perform worse with greater amounts of correlation, while we are interested in assessing how each method performs in comparison with the others. Note that in all simulations, treatment assignments are randomly generated from $\{A, B\}$ with equal probability. I_A and I_B denote the indicators for assignments to treatments A and B respectively. Furthermore, the error term ε in each scenario is normally distributed, i.e., $\varepsilon \sim N(0, 1)$.

Scenarios 1 through 4 assess method performance as the magnitude of correlation between so-called Z variables and truly important variables increases. In scenarios 1 through 4, there are 250 X variables in the data that are all ordinal. In addition to the X variables, 50 Z variables are part of the data. Each Z is highly correlated with truly important variables as specified for each scenario below. The formulas used to calculate the correlated Z variables include a random term ε_i that is normally distributed, i.e., $\varepsilon_i \sim N(0, sd = \sigma)$. When generating the Z variables, decreasing values of σ are used. As σ decreases, the correlation between the Z variables and the important variables increases. For each value of σ, 1000 simulations are run for sample sizes of 250 subjects. Overall, we expect method performance to decrease as σ decreases, i.e., as the correlation level between each Z variable and a truly important variable increases. When the correlation level is greater, the probability that each method erroneously selects a correlated Z variable instead of a truly important variable is greater as well.

Scenario 1: The first scenario consists of an underlying linear model containing the treatment and one important continuous variable. The formula for the outcome variable Y is:

$$Y = 10.2 - 0.3I_B - 0.1X_1 + 2.9I_B X_1 + \varepsilon.$$

The 250 X variables in the data are all independent and normally distributed, i.e., $N(0, 1)$. The 50 Z variables in the data are each highly correlated with variable X_1 and calculated as follows: $Z_i = 0.8X_1 + 0.1X_2 + 0.1X_3 + \varepsilon_i$, where $\varepsilon_i \sim N(0, sd = \sigma)$.

Scenario 2: The second scenario consists of an underlying model with an exponential term containing the treatment and two important continuous variables. The formula for the outcome variable Y is:

$$Y = 10.2 + 0.1I_B \exp\{(X_2 - 0.3)^2 + (X_{10} - 0.1)^2\} + \varepsilon.$$

The 250 X variables in the data are all independent and normally distributed, i.e., $N(0, 1)$. The first 25 Z variables are highly correlated with variable X_2 and calculated as follows: $Z_i = 0.1\,X_1 + 0.8\,X_2 + 0.1X_3 + \varepsilon_i$, where $\varepsilon_i \sim N(0, sd = \sigma)$.

The last 25 Z variables are highly correlated with X_{10} and calculated as follows: $Z_i = 0.1X_1 + 0.8X_{10} + 0.1X_3 + \varepsilon_i$, where $\varepsilon_i \sim N(0, sd = \sigma)$.

Scenario 3: The third scenario consists of an underlying tree model containing the treatment and two important binary variables. The formula for the outcome variable Y is:

$$\begin{aligned} Y = 10.2 &+ I_A I_{\{X_2 \le 0 \cup X_{10} \le 1\}} + 2.6 I_B I_{\{X_2 > 0 \cup X_{10} > 1\}} \\ &+ 0.3X_{30} + 0.6X_{20} - 0.5X_{11}X_{13} + \varepsilon. \end{aligned}$$

The first 230 X variables in the data are from the Discrete Uniform distribution, i.e., Discrete Uniform[0, 2]. These variables are meant to simulate SNP data that have possible values of 0, 1, or 2. The next 10 X variables are Poisson distributed with mean 1, i.e., Poisson(1). The final 10 X variables are Poisson distributed with mean 2, i.e., Poisson(2). The Poisson distributed variables are meant to simulate ordinal count data that could be collected in a clinical trial. In addition, the first 25 Z variables are highly correlated with variable X_2 and calculated as follows: $Z_i = X_2 + \varepsilon_i$, where $\varepsilon_i \sim N(0, sd = \sigma)$. The last 25 Z variables are highly correlated with X_{10} and calculated as follows: $Z_i = X_{10} + \varepsilon_i$, where $\varepsilon_i \sim N(0, sd = \sigma)$. All 50 Z variables are rounded to the nearest integer. To continue simulating SNP data, values less than 0 are set to 0, and values exceeding 2 are set to 2.

Scenario 4: The fourth scenario consists of an underlying tree model containing the treatment and three important binary variables. The formula for the outcome variable Y is:

$$\begin{aligned} Y = &I_{(X_1 \le 0 \cap X_2 \le 0)}(14I_A + 13I_B) \\ &+ I_{(X_1 \le 0 \cap X_2 > 0)}(12I_A + 16I_B) \\ &+ I_{(X_1 > 0 \cap X_3 \le 0)}(13I_A + 11I_B) \\ &+ I_{(X_1 > 0 \cap X_3 > 0)}(13I_A + 14I_B) + \varepsilon. \end{aligned}$$

The first 230 X variables in the data are from the Discrete Uniform distribution, i.e., Discrete Uniform[0, 2]. These variables are meant to simulate SNP data that have possible values of 0, 1, or 2. The next 10 X variables are Poisson distributed with mean 1, i.e., Poisson(1). The final 10 X variables are Poisson distributed with mean 2, i.e., Poisson(2). The Poisson distributed variables are meant to simulate ordinal count data that could be collected in a clinical trial. In addition, the 50 Z variables in the data are each highly correlated with variable X_1 and calculated as follows: $Z_i = X_1 + \varepsilon_i$, where $\varepsilon_i \sim N(0, sd = \sigma)$. All 50 Z variables are rounded to the nearest integer. To continue simulating SNP data, values less than 0 are set to 0, and values exceeding 2 are set to 2.

Scenarios 5 through 8 assess method performance as the number of variables correlated with truly important variables increases. In scenarios 5 through 8, there are 100 X variables in the data that are all ordinal and independent and normally distributed, i.e., $N(0, 1)$. In addition to the X variables, a varying number of Z vari-

ables are part of the data. Each Z is highly correlated with truly important variables and calculated as specified for each scenario below. For each varying number of Z variables, 1000 simulations are run for sample sizes of 250 subjects. Overall, we expect method performance to decrease as the number of Z variables in the data increases. As the number of Z variables increases, the chance of selecting a correlated Z variable instead of a truly important variable also increases.

Scenario 5: The fifth scenario consists of an underlying linear model containing the treatment and one important continuous variable. The formula for the outcome variable Y is:

$$Y = 10.2 - 0.3I_B - 0.1X_1 + 2.9I_B X_1 + \varepsilon.$$

Each Z variable in the data is highly correlated with X_1 and calculated as follows: $Z_i = 0.8X_1 + 0.1X_2 + 0.1X_3 + \varepsilon_i$, where $\varepsilon_i \sim N(0, sd = 0.5)$.

Scenario 6: The sixth scenario consists of an underlying model with an exponential term containing the treatment and two important continuous variables. The formula for the outcome variable Y is:

$$Y = 10.2 + 0.1I_B \exp\{(X_2 - 0.3)^2 + (X_{10} - 0.1)^2\} + \varepsilon.$$

Each Z variable in the data is highly correlated with X_2 and calculated as follows: $Z_i = 0.1X_1 + 0.8X_2 + 0.1X_3 + \varepsilon_i$, where $\varepsilon_i \sim N(0, sd = 0.5)$.

Scenario 7: The seventh scenario consists of an underlying tree model containing the treatment and two important binary variables. The formula for the outcome variable Y is:

$$\begin{aligned} Y = 10.2 &+ I_A I_{\{X_2 \le 0 \cup X_{10} \le 1\}} + 2.6 I_B I_{\{X_2 > 0 \cup X_{10} > 1\}} \\ &+ 0.3X_{30} + 0.6X_{20} - 0.5X_{11}X_{13} + \varepsilon. \end{aligned}$$

Each Z variable in the data is highly correlated with X_2 and calculated as follows: $Z_i = 0.1X_1 + 0.8X_2 + 0.1X_3 + \varepsilon_i$, where $\varepsilon_i \sim N(0, sd = 0.5)$.

Scenario 8: The final scenario consists of an underlying tree model containing the treatment and three important binary variables. The formula for the outcome variable Y is:

$$\begin{aligned} Y = &I_{(X_1 \le 0 \cap X_2 \le 0)}(14I_A + 13I_B) \\ &+ I_{(X_1 \le 0 \cap X_2 > 0)}(12I_A + 16I_B) \\ &+ I_{(X_1 > 0 \cap X_3 \le 0)}(13I_A + 11I_B) \\ &+ I_{(X_1 > 0 \cap X_3 > 0)}(13I_A + 14I_B) + \varepsilon. \end{aligned}$$

Each Z variable in the data is highly correlated with X_1 and calculated as follows: $Z_i = 0.8X_1 + 0.1X_2 + 0.1X_3 + \varepsilon_i$, where $\varepsilon_i \sim N(0, sd = 0.5)$.

16.3.3 Results

All simulation results are presented in Table 16.1. As expected, across all of the simulation scenarios, as the amount of correlation between the Z variables and truly important variables increases, method performance decreases. Method performance is assessed by measuring each method's ability to select the correct relevant variables at early splits. In general, the forest based methods tend to outperform the control methods in scenarios with non-tree models, i.e., scenarios 1 and 5 which contain a linear term and scenarios 2 and 6 which contain an exponential term. Meanwhile,

Table 16.1 Results of simulation scenarios. Proportions of correct variable selection are displayed for each method

Scenario	S.D.	Forest			Tree	
		Weighted	DIPM $mtry$	DIPM no $mtry$	Weighted	DIFF
1. Linear	0.5	0.998	0.993	0.946	0.972	0.913
	0.4	0.950	0.926	0.759	0.876	0.742
	0.3	0.751	0.722	0.457	0.586	0.480
2. Exponential term	0.5	0.028	0.083	0.042	0.051	0.034
	0.4	0.013	0.061	0.020	0.034	0.024
	0.3	0.009	0.037	0.013	0.018	0.014
3. Tree of depth 2	0.5	0.618	0.412	0.293	0.595	0.438
	0.4	0.307	0.236	0.062	0.311	0.207
	0.3	0.020	0.058	0.001	0.033	0.014
4. Tree of depth 3	0.5	0.038	0.090	0.048	0.110	0.289
	0.4	0.012	0.027	0.000	0.037	0.093
	0.3	0.000	0.002	0.000	0.000	0.001
		# of Z Vars.				
5. Linear	0	1.000	0.999	1.000	1.000	1.000
	10	1.000	0.995	0.995	0.996	0.979
	100	0.980	0.978	0.886	0.956	0.872
6. Exponential term	0	0.297	0.599	0.661	0.463	0.329
	10	0.138	0.335	0.352	0.215	0.154
	100	0.067	0.191	0.243	0.066	0.073
7. Tree of depth 2	0	1.000	0.997	0.994	1.000	0.990
	10	0.882	0.870	0.866	0.949	0.886
	100	0.548	0.530	0.497	0.707	0.593
8. Tree of depth 3	0	0.221	0.270	0.245	0.194	0.168
	10	0.032	0.072	0.168	0.192	0.164
	100	0.002	0.007	0.044	0.180	0.141

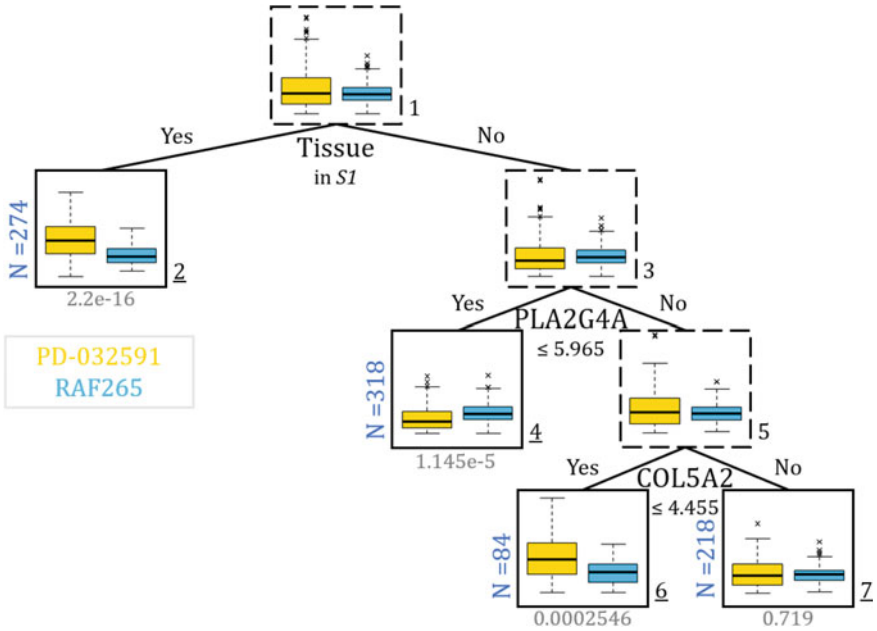


Fig. 16.2 Results of CCLE data application from Zhu et al. [22]. Boxplots comparing the two treatments are in each node, and paired t-test p-values are beneath terminal nodes. For the first split using tissue, $S1$ is the set of categories: autonomic ganglia, large intestine, pancreas, skin, biliary tract, oesophagus, stomach, thyroid, and urinary tract

the control methods tend to outperform the forest based methods in scenarios with underlying tree models, i.e., scenarios 3, 4, 7, and 8.

When comparing the DIPM method that selects $mtry$ variables at each node in embedded trees with the weighted classification tree method, the weighted method slightly outperforms the DIPM method in scenarios 1, 5, and 7. However, in scenarios 2, 4, 6, and 8, the DIPM method outperforms the weighted method. Finally, in scenario 3, the weighted method outperforms the DIPM method until $\sigma = 0.3$. Based on these simulation scenarios, the DIPM method demonstrates comparable and sometimes superior performance in comparison to the more complex weighted method. Although the DIPM method does not consistently outperform the weighted method, recall that our goal is to demonstrate how our intuitive and easy-to-compute importance score can still yield generally comparable performance to the weighted method. These initial developments will then set the stage for consistently better performance in data of greater complexity in future work.

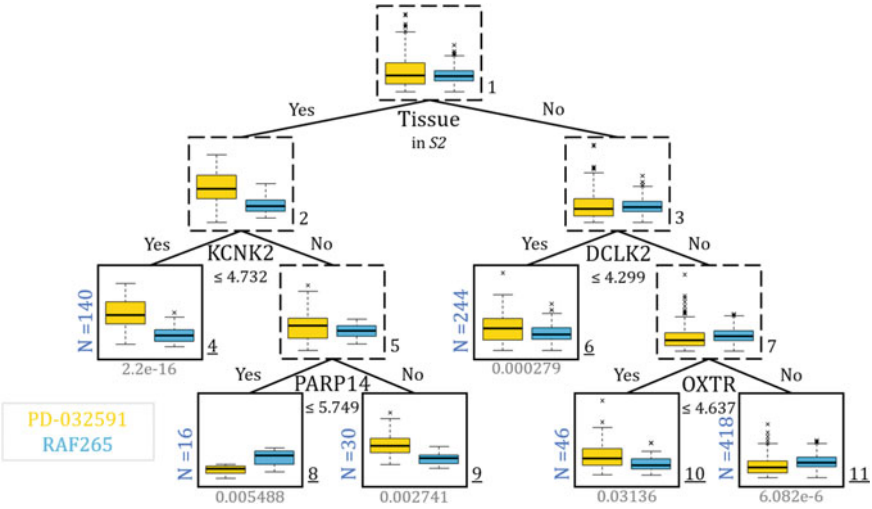


Fig. 16.3 Results of CCLE data application using the DIPM method. Boxplots comparing the two treatments are in each node, and paired t-test p-values are beneath terminal nodes. For the first split using tissue, $S2$ is the set of categories: autonomic ganglia, large intestine, pancreas, and skin

16.4 Analysis of CCLE Data

The DIPM method is applied to a real-world dataset. The data used are a product of the Cancer Cell Line Encyclopedia (CCLE) project by the Broad Institute and the Novartis Institutes for Biomedical Research [1]. The data consist of genetic information and pharmacologic outcomes for more than 1,100 human cancer cell lines. The data are publicly available online (https://portals.broadinstitute.org/ccle/) and are also used by Zhu et al. in their paper [22].

Drug activity measures of multiple drugs are recorded for each cell line. Following the analysis by Zhu et al., two drugs, RAF265 and PD-0325901, are selected for the present analysis. Although Zhu et al. pre-screen the gene expressions and use only the top 500 genes, we use all available gene expressions. For the two selected drugs, there are 447 cell lines, 18,988 gene expressions, and 3 clinical variables available for analysis. The clinical variables are gender, tissue type, and histology. Since the outcome variable is measured for each cell line for each of the two treatments, the final dataset contains 894 observations and 18,991 candidate split variables. All in all, the application of the proposed method to these data produce useful insights. We can use the DIPM method to search for genetic and/or clinical subgroups with varying drug activity levels across the two selected drugs. Moreover, the application presents us with the opportunity to apply the proposed method to a dataset with a large number of candidate split variables.

The constructed tree for the DIPM method is compared to the final tree presented by Zhu et al. [22]. The two trees are depicted in Figs. 16.2 and 16.3. Since their

final tree has a maximum depth of 4, we also present the results with a maximum depth of 4. Terminal node pairs with different optimal treatments are pruned. This simple pruning strategy removes redundant splits and is proposed in Tsai et al. [18]. Furthermore, paired t-test p-values comparing the mean drug activity levels of each treatment are reported beneath the terminal nodes of the two trees. This is done to help quantify how different the two drugs are with respect to drug activity levels within each subgroup. Note that the paired t-test is used since the outcome variable is recorded for both drugs for each cell line.

Both methods identify tissue type as the best split variable at the root node. Though the first split variable is the same, the split values are slightly different. The weighted method places tissue categories autonomic ganglia, large intestine, pancreas, skin, biliary tract, oesophagus, stomach, thyroid, and urinary tract in the child node that identifies PD-0325901 as the optimal treatment. Meanwhile, the DIPM method places only autonomic ganglia, large intestine, pancreas, and skin in the child node that identifies PD-0325901 as the optimal treatment. Despite these differences, ultimately, the t-test p-values comparing the two treatments in these nodes are both approximately equal, i.e., p-value = 2.2e-16.

Meanwhile, the subsequent splits of the DIPM method tree differ from those in Zhu et al.'s final tree. The other splits in Zhu et al.'s final tree use the PLA2G4A and COL5A2 genes. By contrast, the other splits in the proposed method's tree use KCNK2, DCLK2, PARP14, and OXTR. Although neither method clearly outperforms the other in this data application, overall, these results point to the robustness of the effect of tissue type as a potentially useful subgroup indicator. The identified gene expression variables by both methods are also potentially useful subgroup indicators that would have to be examined further for true biological relevance.

16.5 Discussion

In this article, we present the novel DIPM method. The DIPM method is an exploratory method designed to search through existing clinical data for variables that are important in the context of precision medicine. We demonstrate how the proposed method performs well and, in particular, how it compares to the weighted classification tree developed by Zhu et al. [22]. In our simulations, the depth variable importance score demonstrates comparable and sometimes better performance than the variable importance score of the weighted method. The DIPM method achieves this level of performance as a simpler method overall. The DIPM method has no subject specific weights, has no treatment flipping, and considers all possible splits instead of one random split per variable at the nodes of embedded trees. Searching through all splits strengthens the proposed method and better ensures that signals are not missed by sheer chance as in the weighted method. Furthermore, calculating the depth variable importance score is simpler than randomly permuting each variable and counting the misclassifications of out-of-bag samples at each node. In short, the proposed method is less complicated and easier to understand.

Although the presently proposed DIPM method is restricted to the analysis of datasets with continuous outcome variables, the flexibility of the depth variable importance score makes the method readily extendable to other outcome variable types. One useful extension of the DIPM method will be the application to censored survival outcomes. To achieve this application, we will redefine the split criteria and the G statistic in the depth variable importance score accordingly. Note that Zhu et al. have already extended their weighted classification tree method to the analysis of right-censored survival data. It would be interesting to discover whether our markedly simpler method can in fact outperform the weighted method for data with survival endpoints. Also, it would be useful to extend the DIPM method to data with longitudinal outcomes. As mentioned in the introduction, to date, only IT and GUIDE have an extension for data with longitudinal outcomes. It would be interesting to create and assess the performance of the DIPM method when adapted to longitudinal data as well.

A topic of interest for future consideration is covariate selection bias. When searching for the best split at a node, covariates with a greater number of possible splits tend to be selected more often than covariates with fewer splits. The concern is that this phenomenon occurs even when the covariate is not relevant. In this research setting, only Loh et al. have directly addressed this bias by developing a two-step approach [10, 11]. Though we are aware of this bias, we do not directly address covariate selection bias with the proposed method. We aim to continue to consider this issue while developing tree-based methodology moving forward.

Acknowledgements This work was supported in part by NIH Grants T32MH14235, R01 MH116527, and NSF grant DMS1722544. We thank an anonymous referee for their invaluable comments. The Cancer Cell Line Encyclopedia (CCLE) data used in this article are obtained from the CCLE of the Broad Institute. Their database is available publicly online, and they did not participate in the analysis of the data or the writing of this report.

References

1. Barretina, J., Caponigro, G., Stransky, N., Venkatesan, K., Margolin, A.A., et al.: The cancer cell line encyclopedia enables predictive modelling of anticancer drug sensitivity. Nature **483**, 603–607 (2012)
2. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
3. Chen, X., Liu, C.T., Zhang, M., Zhang, H.: A forest-based approach to identifying gene and gene-gene interactions. Proc. Natl. Acad. Sci. U.S.A. **104**, 19199–19203 (2007)
4. Dusseldorp, E., Conversano, C., Van Os, B.J.: Combining an additive and tree-based regression model simultaneously: STIMA. J. Comput. Graph. Stat. **19**, 514–530 (2010)
5. Dusseldorp, E., Van Mechelen, I.: Qualitative interaction trees: a tool to identify qualitative treatment-subgroup interactions. Stat. Med. **33**, 219–237 (2014)
6. Foster, J.C., Taylor, J.M.G., Ruberg, S.J.: Subgroup identification from randomized clinical trial data. Stat. Med. **30**, 2867–2880 (2011)
7. Hamburg, M.A., Collins, F.S.: The path to personalized medicine. N. Engl. J. Med. **363**, 301–304 (2010)

8. Lipkovich, I., Dmitrienko, A.: Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using sides. J. Biopharm. Stat. **24**, 130–153 (2014)
9. Lipkovich, I., Dmitrienko, A., Denne, J., Enas, G.: Subgroup identification based on differential effect search-a recursive partitioning method for establishing response to treatment in patient subpopulations. Stat. Med. **30**, 2601–2621 (2011)
10. Loh, W.Y., Fu, H., Man, M., Champion, V., Yu, M.: Identification of subgroups with differential treatment effects for longitudinal and multiresponse variables. Stat. Med. **35**, 4837–4855 (2016)
11. Loh, W.Y., He, X., Man, M.: A regression tree approach to identifying subgroups with differential treatment effects. Stat. Med. **34**, 1818–1833 (2015)
12. Negassa, A., Ciampi, A., Abrahamowicz, M., Shapiro, S., Boivin, J.F.: Tree-structured subgroup analysis for censored survival data: validation of computationally inexpensive model selection criteria. Stat. Comput. **15**, 231–239 (2005)
13. Ruberg, S.J., Chen, L., Wang, Y.: The mean does not mean as much anymore: finding subgroups for tailored therapeutics. Clin. Trials **7**, 574–583 (2010)
14. Seibold, H., Zeileis, A., Hothorn, T.: Model-based recursive partitioning for subgroup analyses. Int. J. Biostat. **12**, 45–63 (2016)
15. Su, X., Meneses, K., McNees, P., Johnson, W.O.: Interaction trees: Exploring the differential effects of an intervention programme for breast cancer survivors. J. R. Stat. Soc. (Appl. Stat.) **60**, 457–474 (2011)
16. Su, X., Tsai, C.L., Wang, H., Nickerson, D.M., Li, B.: Subgroup analysis via recursive partitioning. J. Mach. Learn. Res. **10**, 141–158 (2009)
17. Su, X., Zhou, T., Yan, X., Fan, J., Yang, S.: Interaction trees with censored survival data. Int. J. Biostat. **4**, 1–26 (2008)
18. Tsai, W.M., Zhang, H., Buta, E., O'Malley, S., Gueorguieva, R.: A modified classification tree method for personalized medicine decisions. Stat. Interface **9**, 239–253 (2016)
19. Zeileis, A., Hothorn, T., Hornik, K.: Model-based recursive partitioning. J. Comput. Graph. Stat. **17**, 492–514 (2008)
20. Zhang, H., Legro, R.S., Zhang, J., Zhang, L., Chen, X., et al.: Decision trees for identifying predictors of treatment effectiveness in clinical trials and its application to ovulation in a study of women with polycystic ovary syndrome. Hum. Reprod. **25**, 2612–2621 (2010)
21. Zhang, H., Singer, B.: Recursive Partitioning and Applications. Springer, New York (2010)
22. Zhu, R., Zhao, Y.Q., Chen, G., Ma, S., Zhao, H.: Greedy outcome weighted tree learning of optimal personalized treatment rules. Biometrics **73**, 391–400 (2017)

Chapter 17
Bayesian Mixture Models with Weight-Dependent Component Priors

Elaheh Oftadeh and Jian Zhang

Abstract In the conventional Bayesian mixture models, independent priors are often assigned to weights and component parameters. This may cause bias in estimation of missing group memberships due to the domination of these priors for some components when there is a big variation across component weights. To tackle this issue, we propose weight-dependent priors for component parameters. To implement the proposal, we develop a simple coordinate-wise updating algorithm for finding empirical Bayesian estimator of allocation or labelling vector of observations. We conduct a simulation study to show that the new method can outperform the existing approaches in terms of adjusted Rand index. The proposed method is further demonstrated by a real data analysis.

17.1 Introduction

Finite mixture models are a popular tool for modelling unobserved heterogeneity in many applications including biology, medicine, economics and engineering among many others (e.g., [3, 4]). Suppose that we sample $\mathbf{y} = (y_1, \cdots, y_N)$ from a population with K groups, described by mixture distribution

$$p(y_i|\boldsymbol{\theta}, \boldsymbol{\eta}) = \sum_{k=1}^{K} \eta_k p(y_i|\boldsymbol{\theta}_k),$$

with unknown component parameters $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_K)$ and unknown weights $\boldsymbol{\eta} = (\eta_1, ..., \eta_K)$. Given the dataset $\mathbf{y} = (y_1, \cdots, y_N)$, we want to infer parameters $(\boldsymbol{\theta}, \boldsymbol{\eta})$

E. Oftadeh (✉) · J. Zhang
School of Mathematics, Statistics and Actuarial Science, University of Kent, Canterbury CT2 7FS, UK
e-mail: eo217@kentforlife.net

J. Zhang
e-mail: jz79@kent.ac.uk

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_17

as well as unobserved component origins of these observations, labelled by allocation (or labelling) vector $\boldsymbol{S} = (s_1, ..., s_N)$. In Bayesian inference, we often adopt the following hierarchical setting:

$$p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{S}) = \prod_{k=1}^{K} \prod_{S_i=k} p(y_i|\boldsymbol{\theta}_k),\ p(\boldsymbol{S}|\boldsymbol{\eta}) = \prod_{k=1}^{K} \eta_k^{\sum_{i=1}^{N} I(S_i=k)},$$

$$p(\boldsymbol{\eta}) = \frac{\Gamma(Ke_0)}{\Gamma(e_0)^K} \prod_{k=1}^{K} \eta_k^{e_0-1},\ e_0 > 0, \quad (\boldsymbol{\theta}, \boldsymbol{\eta})) \sim p(\boldsymbol{\theta})p(\boldsymbol{\eta}),$$

where $I(\cdot)$ is an indicator function, $p(\mathbf{y}|\boldsymbol{\theta}, \boldsymbol{\eta}, \boldsymbol{S})p(\boldsymbol{S}|\boldsymbol{\eta})$ is the complete likelihood and $\boldsymbol{\theta}$ and $\boldsymbol{\eta}$ are of independent priors. The above setting is useful for fitting finite mixture models to data, because they enable the uncertainty in the model parameters to be directly quantified by the posterior distribution. However, it is difficult to make an objective prior setting for the component parameters (such as the component means and variances, in univariate Gaussian mixtures), when there is no subjective information available on which a prior could be based. For example, when some component weights are small, only a small proportion of observations are expected to obtain from these components. In such a situation, the priors can easily dominate the data for these components. Such a prior domination in the inference can cause a bias. To reduce the bias, we need to set these priors compatible to the available information from the data. Ideally, the priors are set to be close to non-informative. On the other hand, standard non-informative priors such as the Jeffreys prior generally cannot be used here, because placing independent improper priors on the component parameters will cause the posterior to be improper as well [9]. This motivates us to explore the advantage of the weight-dependent component priors. In this paper, we propose a new weight-dependent prior specification for finite mixture models in the form $(\boldsymbol{\theta}, \boldsymbol{\eta}) \sim p(\boldsymbol{\theta}|\boldsymbol{\eta})p(\boldsymbol{\eta})$. We develop a coordinate-wise updating algorithm for conducting Bayesian inference: First, given the data, derive a marginal posterior distribution for allocation vector $\boldsymbol{S}$ and optimize it over the labelling space to obtain an optimal allocation estimate $\widehat{\boldsymbol{S}}$. Then, conditional on $\widehat{\boldsymbol{S}}$, calculate the posterior distribution of parameters $(\boldsymbol{\theta}, \boldsymbol{\eta})$. We conduct a simulation study to show that the new approach can outperform the existing methods in terms of adjusted Rand index. The proposed method is further demonstrated by a real data analysis.

The remaining of the paper is organized as follows. The details of the proposed methodology and algorithm are provided in Sect. 17.2. A comparison to the existing methods are made through a simulation study in Sect. 17.3. A real data application is presented in Sect. 17.4. The conclusion is made in Sect. 17.5.

17.2 Methodology

In Bayesian inference, the main task is to calculate the posterior distribution of unknown parameters by combining the prior information about the parameters of interest with the data. Let $\boldsymbol{\vartheta} = (\boldsymbol{\theta}, \boldsymbol{\eta})$. By augmenting the missing allocation vector $\boldsymbol{S}$ into the finite mixture model and letting $p(\boldsymbol{S}|\boldsymbol{\vartheta}) = \prod_{k=1}^{K} \eta_k^{N_k(\boldsymbol{S})}$ with $N_k(\mathbf{S})$ being the size of group k, we can link the incomplete likelihood to the complete likelihood as follows:

$$p(\mathbf{y}|\boldsymbol{\vartheta}) = \int p(\mathbf{y}|\boldsymbol{\vartheta}, \boldsymbol{S}) p(\boldsymbol{S}|\boldsymbol{\vartheta}) d\boldsymbol{S}.$$

Denote the complete data by $(\mathbf{y}, \mathbf{S})$ and the complete-data likelihood by

$$p(\mathbf{y}, \boldsymbol{S}|\boldsymbol{\vartheta}) = p(\mathbf{y}|\mathbf{S}, \boldsymbol{\vartheta}) p(\mathbf{S}|\boldsymbol{\vartheta}) = \prod_{i=1}^{N} p(y_i|\boldsymbol{\vartheta}, S_i) p(S_i|\boldsymbol{\vartheta}).$$

Note that $p(y_i|S_i = k, \boldsymbol{\vartheta}) = p(y_i|\boldsymbol{\theta}_k)$ and $P(S_i = k|\boldsymbol{\vartheta}) = \eta_k$. Therefore, the complete-data likelihood function can be rewritten as

$$p(\mathbf{y}, \mathbf{S}|\boldsymbol{\vartheta}) = \prod_{i=1}^{N} \prod_{k=1}^{K} (p(y_i|\boldsymbol{\theta}_k)\eta_k)^{I(S_i=k)} = \left(\prod_{k=1}^{K} \eta_k^{N_k(\mathbf{S})} \right) \prod_{k=1}^{K} \left(\prod_{i:S_i=k} p(y_i|\boldsymbol{\theta}_k) \right). \tag{17.1}$$

We assign a Dirichlet prior to the weights with the concentration parameter e_0 in the form

$$p(\boldsymbol{\eta}) = \frac{\Gamma(Ke_0)}{\Gamma(e_0)^K} \prod_{k=1}^{K} \eta_k^{e_0 - 1}.$$

By integrating out $\boldsymbol{\eta}$ in $p(\boldsymbol{S}|\boldsymbol{\eta}) p(\boldsymbol{\eta}|e_0)$, we obtain the marginal prior on $\boldsymbol{S}$ and posterior of $\boldsymbol{\eta}$ as follows

$$p(\mathbf{S}) = \frac{\Gamma(Ke_0)}{\Gamma(e_0)^K} \int \prod_{k=1}^{K} \eta_k^{N_k(S)+e_0-1} d\eta_k, \quad p(\boldsymbol{\eta}|\mathbf{S}) = \frac{p(\mathbf{S}|\boldsymbol{\eta}) p(\boldsymbol{\eta})}{p(\mathbf{S})}.$$

Once we have an estimate of $\boldsymbol{S}$, using the above formulas we are able to calculate the posterior of $\boldsymbol{\eta}$. So, in the following, we focus on Bayesian clustering, i.e., Bayesian estimation of allocation vector $\boldsymbol{S}$.

One of the pioneering works in Bayesian clustering was done by [1], where the problem was formulated in a Bayesian decision theoretic framework with a loss function $R(\mathbf{S}, \widehat{\mathbf{S}})$. This loss function measures the difference between the estimate $\widehat{\mathbf{S}}$ and the true grouping $\boldsymbol{S}$. Here, we take an empirical Bayesian method by optimizing the marginal posterior of allocation vector of $\boldsymbol{S}$, $p(\boldsymbol{S}|\mathbf{y})$. In the simulation study,

we evaluate the accuracy of the clustering by calculating the similarity between the estimated and the true labelling by the so-called adjusted Rand index [5, 7]. We consider two different sets of hierarchical priors and derive the corresponding posteriors. In the Bayesian inference for Gaussian mixtures, it is common to choose the component parameter priors to be independent of weights. We derive the posteriors for Bayesian mixture models with independent priors in Sect. 17.2.1.1 and for the models with dependent priors in Sect. 17.2.1.2 below. Although from now on we focus on univariate normal mixtures, the method can be extended to other mixtures such as multivariate normal or non-normal mixtures. For simplicity, we assume that K is known. Otherwise, we can take a Poisson distribution as a prior for K.

17.2.1 Mixture of Univariate Normals

Suppose that $y_i \sim N(\mu_k, \sigma_k^2), i = 1, \cdots, N$, with $\boldsymbol{\theta}_k = (\mu_k, \sigma_k^2), k = 1, 2, ..., K$. For the univariate normal mixtures, we first derive the posterior distribution for mean μ_k and variance $\sigma_k^2, k = 1, ..., K$, given the complete data $(\mathbf{y}, \boldsymbol{S})$. We then work out the formulas for calculating and optimizing $p(\boldsymbol{S}|\mathbf{y})$.

17.2.1.1 Weight-Independent Component Priors

We start with a review of the conventional hierarchical model with weight-independent priors on (μ_k, σ_k^2) in [2, 3]:

$$y_i \sim N(\mu_k, \sigma_k^2), \quad \mu_k \sim N(\mu_{k0}, \sigma_{k0}^2), \quad \sigma_k^2 \sim IG(a_0, b_0),$$

where $IG(a_0, b_0)$ is an inverse Gamma density with hyperparameters (a_0, b_0).

The posterior probability of μ_k given the complete data $(\mathbf{S}, \mathbf{y})$ and σ_k^2 can be written as

$$\begin{aligned} p(\mu_k|\mathbf{y}, \mathbf{S}, \sigma_k^2) &\propto p(\mathbf{y}|\mu_k, \sigma_k^2, \mathbf{S}) p(\mu_k) \\ &\propto \exp\left\{-\frac{1}{2}\left(\frac{N_k(\mathbf{S})}{\sigma_k^2} + \frac{1}{\sigma_{k0}^2}\right)\left(\mu_k - \frac{\sum y_i}{\sigma_k^2} + \frac{\mu_{k0}}{\sigma_{k0}^2}\right)^2\right\}. \end{aligned}$$

Thus the posterior distribution of μ_k is the following normal distribution

$$\begin{aligned} p(\mu_k|\mathbf{y}, \boldsymbol{S}, \sigma_k^2) \sim \mathcal{N}(b_k(\boldsymbol{S}), B_k(\boldsymbol{S})), \quad B_k(\boldsymbol{S})^{-1} &= \sigma_{k0}^{-2} + \sigma_k^{-2} N_k(\boldsymbol{S}) \\ b_k(\boldsymbol{S}) \quad &= B_k(\boldsymbol{S})\left(\sigma_k^{-2} N_k(\boldsymbol{S}) \bar{y}_k(\boldsymbol{S}) + \sigma_{k0}^{-2} \mu_{k0}\right), \end{aligned}$$

where the sample mean and variance in the kth group are denoted by

$$\bar{y}_k(\boldsymbol{S}) = \frac{1}{N_k(\boldsymbol{S})} \sum_{i:S_i=k} y_i, \quad S^2_{y,k} = \frac{1}{N_k(\boldsymbol{S})} \sum_{i:S_i=k} (y_i - \bar{y}_k(\boldsymbol{S}))^2.$$

Similarly, for σ_k^2 we have

$$\sigma_k^2 | \mathbf{y}, \boldsymbol{S}, \mu_k \sim \mathscr{G}^{-1}(c_k(\boldsymbol{S}), C_k(\boldsymbol{S})), \quad c_k(\boldsymbol{S}) = a_0 + \frac{1}{2} N_k(\boldsymbol{S}),$$
$$C_k(\boldsymbol{S}) = b_0 + \frac{1}{2} \sum_{i:S_i=k} (y_i - \mu_k)^2.$$

If we are able to calculate the maximum marginal posterior estimator of the allocation vector, $\widehat{\boldsymbol{S}} = \text{argmax}_S p(\boldsymbol{S}|\mathbf{y})$, then we can directly calculate posterior distribution of $(\boldsymbol{\theta}, \boldsymbol{\eta})$. To derive the marginal posterior distribution of allocations, we integrate out $(\boldsymbol{\theta}, \boldsymbol{\eta})$ from the model, i.e., consider the following integration

$$\begin{aligned} p(\mathbf{S}|\mathbf{y}) &= \iint p(\mathbf{y}|\boldsymbol{\eta}, \mathbf{S}, \boldsymbol{\theta}) p(\mathbf{S}|\boldsymbol{\eta}) p(\boldsymbol{\eta}) p(\boldsymbol{\theta}) d\theta d\boldsymbol{\eta} \\ &= 2^{K a_0} N_0^{K/2} \left(\frac{1}{\pi}\right)^{N/2} \left(\frac{b_0^{a_0}}{\Gamma(a_0)}\right)^K \frac{\Gamma(K e_0)}{\Gamma(e_0)^K} \prod_{k=1}^K \sqrt{\frac{1}{N_k(\mathbf{S}) + N_0}} \\ &\times \frac{\prod_{k=1}^K \Gamma(N_k(S) + e_0)}{\Gamma(K e_0 + N)} \prod_{k=1}^K \mathscr{B}^{-(a_0 + \frac{N_k(\mathbf{S})}{2})} \prod_{k=1}^K \Gamma(a_0 + \frac{N_k(\mathbf{S})}{2}), \end{aligned}$$

where

$$\begin{aligned} S^2_{y,k} &= \frac{1}{N_k(\boldsymbol{S})} \sum_{i:S_i=k} (y_i - \bar{y}_k(\boldsymbol{S}))^2 \\ \mathscr{B} &= N_k(\mathbf{S}) S^2_{y,k}(\mathbf{S}) + 2b_0 + \frac{N_k(\mathbf{S}) N_0}{N_k(\mathbf{S}) + N_0} (\bar{y}(\mathbf{S}) - \mu_{k0})^2. \end{aligned}$$

Taking logarithm, we have

$$\begin{aligned} \log(p(\mathbf{S}|\mathbf{y})) \propto & \sum_{k=1}^K \log \Gamma(N_k(\mathbf{S}) + e_0) \sum_{k=1}^K \log \Gamma(a_0 + \frac{N_k(\mathbf{S})}{2}) \\ & - \sum_{k=1}^K \frac{1}{2} \log(N_k(\mathbf{S}) + N_0) - \sum_{k=1}^K (a_0 + N_k(\mathbf{S})/2) \log \mathscr{B}. \quad (17.2) \end{aligned}$$

17.2.1.2 Weight-Dependent Component Priors

Although we consider the same normal mixture model as in the previous section, we allow certain dependency of the hierarchical priors on component weights as follows:

$$\mu_k|\sigma_k^2, \eta_k \sim N(\mu_{k0}, \frac{\sigma_{k0}^2}{N_0\eta_k}), \quad \sigma_k^2 \sim IG(a_0, b_0), k = 1, ..., K. \quad \boldsymbol{\eta} \sim D(e_0, \cdots, e_0),$$

where $D(e_0, \cdots, e_0)$ is a Dirichlet density with concentration parameter e_0. Since N_0 is the total number of prior units we assign to the model, $N_0\eta_k$ is the number of prior units we assign to μ_k. Unlike the weight-independent priors, the prior of μ_k is adaptive to η_k in the sense that the amount of priors will be varying in η_k, in particular, it will be nearly non-informative when η_k tends to zero. The posterior of μ_k given $(\boldsymbol{S}, \mathbf{y})$, σ_k^2 and η_k can then be written as

$$\begin{aligned} p(\mu_k|\mathbf{y}, \boldsymbol{S}, \sigma_k^2, \eta_k) &\propto p(\mathbf{y}|\mu_k, \sigma_k^2, \mathbf{S})p(\mu_k|\eta_k) \\ &\propto \prod_{k=1}^{K} (\frac{1}{\sigma_k^2})^{-N_k(\mathbf{S})/2} \exp\left\{-\frac{1}{2\sigma_k^2}\sum_{i:S_i=k}(y_i - \mu_k)^2\right\} \\ &\times (\frac{1}{\sigma_{k0}^2}\eta_k)^{1/2} \exp\left\{-\frac{1}{2\sigma_{k0}^2\eta_k}(\mu_k - \mu_{k0})^2\right\} \\ &\propto \exp\left\{-\frac{1}{2}\left(\frac{N_k(\mathbf{S})}{\sigma_k^2} + \frac{1}{\sigma_{k0}^2\eta_k}\right)\left(\mu_k - \frac{\sum y_i}{\sigma_k^2} + \frac{\mu_{k0}}{\sigma_{k0}^2\eta_k}\right)^2\right\}. \end{aligned}$$

Thus the posterior distribution of μ_k is the following normal distribution

$$\begin{aligned} p(\mu_k|\mathbf{y}, \boldsymbol{S}, \sigma_k^2, \eta_k) &\sim \mathcal{N}(b_k(\boldsymbol{S}), B_k(\boldsymbol{S})), \\ B_k(\boldsymbol{S})^{-1} &= \sigma_{k0}^{-2}\eta_k^{-1} + \sigma_k^{-2}N_k(\boldsymbol{S}) \\ b_k(\boldsymbol{S}) &= B_k(\boldsymbol{S})\left(\sigma_k^{-2}N_k(\boldsymbol{S})\bar{y}_k(\boldsymbol{S}) + \eta_k^{-1}\sigma_{k0}^{-2}\mu_{k0}\right), \end{aligned}$$

where the sample mean and variance in the kth group are denoted by

$$\bar{y}_k(\boldsymbol{S}) = \frac{1}{N_k(\boldsymbol{S})}\sum_{i:S_i=k} y_i, \quad s_{\bar{y},k}^2 = \frac{1}{N_k(\boldsymbol{S})}\sum_{i:S_i=k}(y_i - \bar{y}_k(\boldsymbol{S}))^2.$$

Similarly, for σ_k^2 we have

$$\sigma_k^2|\mathbf{y}, \boldsymbol{S}, \mu_k \sim \mathscr{G}^{-1}(c_k(\boldsymbol{S}), C_k(\boldsymbol{S})),$$
$$c_k(\boldsymbol{S}) = a_0 + \frac{1}{2}N_k(\boldsymbol{S}),$$
$$C_k(\boldsymbol{S}) = b_0 + \frac{1}{2}\sum_{i:S_i=k}(y_i - \mu_k)^2.$$

According to the above hierarchical prior setting, the joint distribution of the data and the model parameters can be expressed as

$$\begin{aligned}
&p(\mathbf{y}|\mathbf{S}, \boldsymbol{\vartheta})p(\mathbf{S}|\boldsymbol{\eta})p(\boldsymbol{\theta}|\boldsymbol{\eta})p(\boldsymbol{\eta}) \\
&= \prod_{i=1}^{N}\prod_{k=1}^{K}\left(p(\mathbf{y}_i|\mu_k, \sigma_k^2)\eta_k\right)^{I_{S_i=k}}\prod_{k=1}^{K}p(\mu_k|\sigma_k^2, \eta_k)p(\sigma_k^2)p(\eta_k) \\
&= \prod_{k=1}^{K}\left(\prod_{i:S_i=k}\frac{1}{\sqrt{2\pi\sigma^2}}\exp\left\{-\frac{\sum_{i:S_i=k}(y_i - \mu_k)^2}{2\sigma_k^2}\right\}\right)\left(\prod_{k=1}^{K}\eta_k^{\sum_{i=1}^{N}I_{S_i=k}}\right) \\
&\quad \times \prod_{k=1}^{K}\left(\frac{N_0\eta_k}{2\pi\sigma_k^2}\right)^{1/2}\exp\left\{-\frac{N_0\eta_k}{2\sigma_k^2}(\mu_k - \mu_{k0})^2\right\} \\
&\quad \times \prod_{k=1}^{K}\frac{b_0^{a_0}}{\Gamma(a_0)}(\sigma_k^2)^{-a_0-1}\exp\left\{-b_0/\sigma_k^2\right\} \times \frac{\Gamma(\sum_{k=1}^{K}e_0)}{\prod_{k=1}^{K}\Gamma(e_0)}\prod_{k=1}^{K}\eta_k^{e_0-1}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&p(\mathbf{y}|\mathbf{S}, \boldsymbol{\vartheta})p(\mathbf{S}|\boldsymbol{\eta})p(\boldsymbol{\theta}|\boldsymbol{\eta})p(\boldsymbol{\eta}) \\
&= \left(\frac{1}{2\pi}\right)^{\frac{\sum_{k=1}^{K}N_k(S)}{2}}\left(\frac{N_0}{2\pi}\right)^{K/2}\left(\frac{b_0^{a_0}}{\Gamma(a_0)}\right)^{K}\frac{\Gamma(Ke_0)}{\Gamma(e_0)^K} \\
&\quad \times \prod_{k=1}^{K}\exp\left\{-\frac{\sum_{i:S_i=k}(y_i - \mu_k)^2 + N_0\eta_k(\mu_k - \mu_0)^2 + 2b_0}{2\sigma_k^2}\right\} \\
&\quad \times \prod_{k=1}^{K}\eta_k^{e_0+N_k(S)-1/2}\prod_{k=1}^{K}\frac{1}{\sigma_k^{2(a_0+1)+N_k(S)+1}}
\end{aligned}$$

After doing some simple algebra we get

$$
\begin{aligned}
&p(\mathbf{y}|\mathbf{S},\boldsymbol{\vartheta})\,p(\mathbf{S}|\boldsymbol{\eta})\,p(\boldsymbol{\theta}|\boldsymbol{\eta})\,p(\boldsymbol{\eta})\\
&=\left(\frac{1}{2\pi}\right)^{N/2}\left(\frac{N_0}{2\pi}\right)^{K/2}\left(\frac{b_0^{a_0}}{\Gamma(a_0)}\right)^{K}\frac{\Gamma(Ke_0)}{\Gamma(e_0)^K}\\
&\times\prod_{k=1}^{K}\exp\left\{-\frac{(N_k(\mathbf{S})+N_0\eta_k)\left[\mu_k-\frac{N_k(\mathbf{S})\bar{y}_k(\mathbf{S})+N_0\eta_k\mu_{k0}}{N_k(\mathbf{S})+N_0\eta_k}\right]^2}{2\sigma_k^2}\right\}\\
&\times\prod_{k=1}^{K}\exp\left\{-\frac{N_k(\mathbf{S})S_{y,k}^2(\mathbf{S})+2b_0+\frac{N_k(\mathbf{S})N_0\eta_k}{N_k(\mathbf{S})+N_0\eta_k}(\bar{y}_k(\mathbf{S})-\mu_{k0})^2}{2\sigma_k^2}\right\}\\
&\times\prod_{k=1}^{K}\eta_k^{e_0+N_k(S)-1/2}\prod_{k=1}^{K}\frac{1}{\sigma_k^{2(a_0+1)+N_k(S)+1}}
\end{aligned}
$$

Now we are going to find the marginal posterior distribution of the allocation vector $p(\mathbf{S}|\mathbf{y})$ by integrating out all parameters. We first integrate out μ_k from the above expression and we get

$$
\begin{aligned}
&\prod_{k=1}^{K}\int p(\mathbf{y},\mathbf{S}|\boldsymbol{\vartheta})\,p(\mathbf{S}|\boldsymbol{\eta})\,p(\boldsymbol{\theta})\,p(\boldsymbol{\eta})\,d\mu_k\\
&=N_0^{K/2}\left(\frac{1}{2\pi}\right)^{N/2}\left(\frac{b_0^{a_0}}{\Gamma(a_0)}\right)^{K}\frac{\Gamma(Ke_0)}{\Gamma(e_0)^K}\prod_{k=1}^{K}\sqrt{\frac{1}{N_k(\mathbf{S})+N_0\eta_k}}\\
&\times\prod_{k=1}^{K}\exp\left\{-\frac{N_k(\mathbf{S})S_{y,k}^2(\mathbf{S})+2b_0+\frac{N_k(\mathbf{S})N_0\eta_k}{N_k(\mathbf{S})+N_0\eta_k}(\bar{y}_k(\mathbf{S})-\mu_{k0})^2}{2\sigma_k^2}\right\}\\
&\times\prod_{k=1}^{K}\eta_k^{e_0+N_k(\boldsymbol{S})-1/2}\prod_{k=1}^{K}\frac{1}{\sigma_k^{2(a_0+1)+N_k(\boldsymbol{S})+2}}
\end{aligned}
$$

Finally integrating out σ_k and η_k, the posterior $p(\mathbf{S}|\mathbf{y})$ is obtained as

$$
\begin{aligned}
p(\mathbf{S}|\mathbf{y})&=\int_0^1\iint p(\mathbf{y}|\boldsymbol{\eta},\mathbf{S},\boldsymbol{\theta})\,p(\mathbf{S}|\eta)\mathbf{p}(\boldsymbol{\eta})\mathbf{p}(\boldsymbol{\theta})\mathbf{d}\boldsymbol{\theta}\mathbf{d}\boldsymbol{\eta}\\
&=N_0^{K/2}\left(\frac{1}{\pi}\right)^{N/2}\left(\frac{b_0^{a_0}}{\Gamma(a_0)}\right)^{K}\frac{\Gamma(Ke_0)}{\Gamma(e_0)^K}2^{Ka_0}\\
&\times\frac{\prod_{k=1}^{K}\Gamma(N_k(\mathbf{S})+e_0+1/2)}{\Gamma(N+Ke_0+K/2)}\prod_{k=1}^{K}\Gamma(N_k(\mathbf{S})/2+a_0)\\
&\times\prod_{k=1}^{K}\int_0^1\frac{\mathscr{B}(\eta_k)^{-a_0-\frac{N_k(\mathbf{S})}{2}}}{(N_k(\mathbf{S})+N_0\eta_k)^{1/2}}d\eta_k,
\end{aligned}
\tag{17.3}
$$

where

$$\mathscr{B}(\eta_k) = N_k(\mathbf{S})S^2_{y,k}(\mathbf{S}) + 2b_0 + \frac{N_k(\mathbf{S})N_0\eta_k}{N_k(\mathbf{S}) + N_0\eta_k}(\bar{y}_k(\mathbf{S}) - \mu_{k0})^2.$$

As we can see for the case with dependent hierarchical priors there is no explicit form for the posterior $p(\boldsymbol{S}|\mathbf{y})$. Due to this formulation, we faced some challenges in calculating the integration in the expression (17.3). Calculating this integration is not always possible in a usual way as a result of overflow or underflow, depending on simulation settings. To address this issue we calculate this definite integral by calculating Riemann sums over a partition of [0, 1].

Note that

$$\int_0^1 f(\eta_k)d\eta_k = \int_0^1 \frac{(N_k(\mathbf{S})S^2_{y,k}(\mathbf{S}) + 2b_0 + \frac{N_k(\mathbf{S})N_0\eta_k}{N_k(\mathbf{S})+N_0\eta_k}(\bar{y}_k(\mathbf{S}) - \mu_{k0})^2)^{-a_0-\frac{N_k(\mathbf{S})}{2}}}{(N_k(\mathbf{S}) + N_0\eta_k)^{1/2}}.$$

We rearrange the above integrand as follows:

$$f(\eta_k) = \frac{(N_k(\mathbf{S})S^2_{y,k}(\mathbf{S}))^{-a_0-N_k(\mathbf{S})/2}D(\eta_k)^{-a_0-N_k(\mathbf{S})/2}}{N_k(\mathbf{S})^{1/2}\left(1 + \frac{N_0\eta_k}{N_k(\mathbf{S})}\right)^{1/2}}, \tag{17.4}$$

where

$$D(\eta_k) = 1 + \frac{1}{N_k(\mathbf{S})S^2_{y,k}(\mathbf{S})}\left(2b_0 + \frac{N_k(\mathbf{S})N_0\eta_k}{N_k(\mathbf{S}) + N_0\eta_k}(\bar{y}_k(\mathbf{S}) - \mu_{k0})^2\right) \tag{17.5}$$

We partition [0, 1] into subintervals $[x_0, x_1], [x_1, x_2], \cdots, [x_{n-1}, x_n]$ with $\Delta x_i = x_i - x_{i-1} = 1/n$ and $x_i^* = i\Delta x_i$. This leads to

$$\int_0^1 f(\eta_k)d\eta_k \approx \frac{1}{n}\sum_{i=1}^n f_k(x_i^*).$$

Even using the above approximation did not completely solve the problem of overflow and underflow and we still got some infinity values in numerical calculations. To tackle this problem we divide all summands by the largest element which is $f_k(x_n^*) = f_k(1)$. Therefore we calculate

$$\int_0^1 f(\eta_k)d\eta_k \approx \frac{f_k(x_n^*)}{n}\sum_{i=1}^n \frac{f_k(x_i^*)}{f_k(x_n^*)}, \tag{17.6}$$

where according to the equation (17.4) we have

$$\frac{f_k(x_i^*)}{f_k(x_n^*)} = \frac{(1+\frac{N_0}{N_k(\mathbf{S})})^{1/2}}{(1+\frac{N_0 x_i^*}{N_k(\mathbf{S})})^{1/2}(\frac{D_k(x_i^*)}{D_k(1)})^{a_0+N_k(\mathbf{S})/2}}.$$

Now according to the expression (17.5) we have

$$\begin{aligned}\frac{D_k(x_i^*)}{D_k(1)} &\approx \frac{N_k(\mathbf{S})S_{y,k}^2(\mathbf{S}) + 2b_0 + \frac{N_k(\mathbf{S})N_0 x_i^*}{N_k(\mathbf{S})+N_0 x_i^*}(\bar{y}_k(\mathbf{S})-\mu_{k0})^2}{N_k(\mathbf{S})S_{y,k}^2(\mathbf{S}) + 2b_0 + \frac{N_k(\mathbf{S})N_0}{N_k(\mathbf{S})+N_0}(\bar{y}_k(\mathbf{S})-\mu_{k0})^2} \\ &= \frac{S_{y,k}^2(\mathbf{S}) + \frac{2b_0}{N_k(\mathbf{S})} + \frac{N_0 x_i^*}{N_k(\mathbf{S})+N_0 x_i^*}(\bar{y}_k(\mathbf{S})-\mu_{k0})^2}{S_{y,k}^2(\mathbf{S}) + \frac{2b_0}{N_k(\mathbf{S})} + \frac{N_0}{N_k(\mathbf{S})+N_0}(\bar{y}_k(\mathbf{S})-\mu_{k0})^2}.\end{aligned}$$

In order to use the latter expression in computational programming and avoid any possible underflow issue, we further rearrange the latter expression to get

$$\begin{aligned}\frac{D_k(x_i^*)}{D_k(1)} &= 1 - \frac{(\bar{y}_k(\mathbf{S})-\mu_{k0})^2\left(\frac{N_0}{N_k(\mathbf{S})+N_0} - \frac{N_0 x_i^*}{N_k(\mathbf{S})+N_0 x_i^*}\right)}{S_{y,k}^2(\mathbf{S}) + \frac{2b_0}{N_k(\mathbf{S})} + \frac{N_0}{N_k(\mathbf{S})+N_0}(\bar{y}_k(\mathbf{S})-\mu_{k0})^2} \\ &= 1 - \frac{\frac{N_0}{N_k(\mathbf{S})+N_0}\left(1 - \frac{(N_k(\mathbf{S})+N_0)x_i^*}{N_k(\mathbf{S})+N_0 x_i^*}\right)(\bar{y}_k(\mathbf{S})-\mu_{k0})^2}{S_{y,k}^2(\mathbf{S}) + \frac{2b_0}{N_k(\mathbf{S})} + \frac{N_0}{N_k(\mathbf{S})+N_0}(\bar{y}_k(\mathbf{S})-\mu_{k0})^2}.\end{aligned}$$

Now the integration in (17.6) can be approximated by the following summation

$$\begin{aligned}\frac{f_k(x_n^*)}{n}\sum_{i=1}^{n}\frac{f_k(x_i^*)}{f_k(x_n^*)} &= \frac{1}{n}\frac{(N_k(\mathbf{S})S_{y,k}^2(\mathbf{S}))^{-a_0-\frac{N_k(\mathbf{S})}{2}}D(1)^{-a_0-N_k(\mathbf{S})/2}}{N_k(\mathbf{S})^{1/2}} \\ &\quad\times\sum_{i=1}^{n}\frac{1}{(1+\frac{N_0 x_i^*}{N_k(\mathbf{S})})^{1/2}(\frac{D_k(x_i^*)}{D_k(1)})^{a_0+N_k(\mathbf{S})/2}}\end{aligned}$$

Substituting the above expression in the allocation posterior results in

$$\begin{aligned}p(\mathbf{S}|\mathbf{y}) &\approx N_0^{K/2}\left(\frac{1}{\pi}\right)^{N/2}\left(\frac{b_0^{a_0}}{\Gamma(a_0)}\right)^K\frac{\Gamma(Ke_0)}{\Gamma(e_0)^K}2^{Ka_0} \\ &\quad\times\frac{\prod_{k=1}^{K}\Gamma(N_k(\mathbf{S})+e_0+1/2)}{\Gamma(N+Ke_0+K/2)}\prod_{k=1}^{K}\Gamma(N_k(\mathbf{S})/2+a_0) \\ &\quad\times\frac{1}{n}\frac{(N_k(\mathbf{S})S_{y,k}^2(\mathbf{S}))^{-a_0-N_k(\mathbf{S})/2}D(1)^{-a_0-N_k(\mathbf{S})/2}}{N_k(\mathbf{S})^{1/2}} \\ &\quad\times\sum_{i=1}^{n}\frac{1}{(1+\frac{N_0 x_i^*}{N_k(\mathbf{S})})^{1/2}(\frac{D_k(x_i^*)}{D_k(1)})^{a_0+N_k(\mathbf{S})/2}}\end{aligned}$$

Taking the logarithm, we have

$$\begin{aligned}\log(p(\mathbf{S}|\mathbf{y})) \approx\ & K/2\log(N_0) - N/2\log(\pi) + Ka_0\log(b_0) - K\log\Gamma(a_0) \\ & + \log\Gamma(Ke_0) - K\log\Gamma(e_0) + Ka_0\log(2) - \Gamma(N + Ke_0 + K/2) \\ & + \sum_{k=1}^{K}\log\Gamma(N_k(\mathbf{S}) + e_0 + 1/2) + \sum_{k=1}^{K}\log\Gamma(a_0 + \frac{N_k(\mathbf{S})}{2}) \\ & - \sum_{k=1}^{K}(a_0 + N_k(\mathbf{S})/2)\left[\log(N_k(\mathbf{S})) + \log(S^2_{y,k}) + \log(D(1))\right]\end{aligned}$$

$$- \sum_{k=1}^{K}\log(n) + \sum_{k=1}^{K}\log\sum_{i=1}^{n}\frac{1}{(1+\frac{N_0 x_i^*}{N_k(\mathbf{S})})^{1/2}(\frac{D_k(x_i^*)}{D_k(1)})^{a_0+N_k(\mathbf{S})/2}} \tag{17.7}$$

17.3 A Simulation Study

In this section, we conduct simulations to compare the classification accuracy of Bayesian normal mixture model with that of normal mixture models. We implement the Bayesian normal mixture model based on both independent priors and dependent priors. In order to compare the performance of the Bayesian mixture model to the frequentist model, we use the Mclust software, where the optimal allocation estimate is obtained by using the Expectation-Maximization algorithm.

17.3.1 Adjusted Rand Index

We review one of the widely used methods called adjusted Rand index for quantifying the degree of the agreement between partitions derived from different methods. Suppose we have n objects to classify and $P_1 = \{C_1, \cdots, C_r\}$ is a partition that assigns these objects into r classes and $P_2 = \{C_1, \cdots, C_s\}$ assigns them into s classes. Each pair of objects, either have the same class label or a different one. Since the number of classified objects is n, we have the total number of $n(n-1)/2$ pairs to compare. Let a be the number of pairs that the two partitions agree by assigning the elements to the same classes and b be the number of pairs that partitions agree by assigning them to different classes. Considering all pairs, the proportion of agreement between P_1 and P_2 is evaluated by the following Rand index (RI)

$$\mathrm{RI}(P_1, P_2) = \frac{a+b}{n(n-1)/2}$$

Since the expectation of Rand index for two random partitions is not a constant, [5] proposed a normalized Rand index which is defined by

$$\text{ARI} = \frac{\text{Rand index} - \text{Expected value of Rand index}}{\text{Maximum value of Rand index} - \text{Expected value of Rand index}}.$$

When two partitions completely agree, the adjusted Rand index reaches the maximum value 1. The higher ARI value, the greater degree of agreement between two partitions is.

17.3.2 Simulated Data

We generated data from a normal mixture model with three components. We used the same setting as used in one of the examples in [3] to generate the data. The underlying weights (0.3, 0.2, 0.5). The underlying component means and variances are $(-3, 0, 2)$ and $(1, 0.5, 0.8)$ respectively.

17.3.3 Results

We utilized the Bayesian mixture model under the following hierarchical priors where the component mean depends on the weight corresponding to that component

$$\mu_k|\sigma_k^2, \eta_k \sim N(\mu_{k0}, \frac{\sigma_{k0}^2}{N_0\eta_k}), \quad \sigma_k^2 \sim IG(a_0, b_0), \quad \eta_k \sim D(e_0, \cdots, e_0),$$

which results in the log-allocation posterior in equation (17.7). We also implemented the Bayesian mixture model with hierarchical priors where the mean of each component was independent of the weight as following

$$\mu_k|\sigma_k^2 \sim N(\mu_{k0}, \frac{\sigma_{k0}^2}{N_0}), \quad \sigma_k^2 \sim IG(a_0, b_0), \quad \eta_k \sim D(e_0, \cdots, e_0).$$

The allocation posterior can be regarded as a function of hyperparameters $N_0, a_0, b_0, e_0, \mu_{k0}$. Following [8], we set μ_{k0} to the median of the data. The hyperparameters are chosen as $a_0 = 2$ and $e_0 = 1$ and for the parameter b_0 they consider the prior $b_0 \sim G(0.2, 10/R^2)$ where R^2 is the length of the interval of the variation of the data. In order to choose N_0, following [6], we set $N_0 = 2.6/(y_{\max} - y_{\min})$.

We found the optimal classification by maximizing the logarithm of the allocation posterior. The optimization was carried out by the following iterative algorithm: We updated the coordinates of the allocation vector in one-by-one and calculated the corresponding posterior. The algorithm started with an initial allocation vector $\boldsymbol{S}^{(0)} = \boldsymbol{S}_{\text{current}}$ by the result derived from MClust. For example, to update the coordinate S_1 corresponding to y_1 while other coordinates were fixed, we generated a random number U from the uniform distribution $\mathscr{U}[0, 1]$. If $U < \eta_1$, assign the observation y_1 to the first component. If $\eta_1 \leq U < \eta_1 + \eta_2$, assign the observation to the second component. Otherwise, assign y_1 to the third component. This resulted in an updated allocation vector $\boldsymbol{S}^{(1)} = \boldsymbol{S}_{\text{updated}}$. The number of elements in each component changed. If $S_1^{\text{new}} = S_1 = k$, then no moving occurred whereas, if the observation moved to another component, say l, then the number of observations in each component was updated as

$$N_k(S_1^{\text{new}}, \boldsymbol{S}_{-1}) = N_k(\boldsymbol{S}) - 1, \quad N_l(S_1^{\text{new}}, \boldsymbol{S}_{-1}) = N_l(\boldsymbol{S}) + 1,$$

where $\boldsymbol{S}_{-1} = (S_2, ..., S_N)$. Correspondingly, the mean $\bar{y}_k(\boldsymbol{S})$ and the variance $S_{y,k}(\boldsymbol{S})$ of each component were updated. Then the log-posterior $p((S_1^{\text{new}}, \boldsymbol{S}_{-1})|\mathbf{y})$ of the updated allocation vector was calculated using the expression (17.7). The updated allocation for the first observation was accepted if the updated posterior was greater than the current posterior, i.e. $p((S_1^{\text{new}}, \boldsymbol{S}_{-1})|\mathbf{y}) > p(\boldsymbol{S}^{(0)}|\mathbf{y})$. If the new allocation was accepted, then this updated allocation was used as the current allocation in the next iteration $\boldsymbol{S}_{\text{current}} = \boldsymbol{S}_{\text{updated}}$ and the observation was moved to the component l. Otherwise, the observation was kept in the current component k and the algorithm moved to the next observation y_2. These steps were repeated until all observations $i = 1 \cdots, N$ were updated and until the posterior reaches a local maximum. Then this optimal allocation vector was recorded and compared with Mclust by computing their adjusted Rand index.

We simulated 300 datasets from a three component mixture of normals. We applied the above algorithm to find the optimal grouping for each of these data. We applied both the weight-dependent (17.7) and weight-independent (17.2) prior approaches.

Results displayed in Fig. 17.1 show that the Bayesian clustering outperformed the Mclust particularly when the component priors were weight-dependent. The results illustrated that imposing dependency of component priors on weights can reduce the bias of clustering due to the effect of weight heterogeneity. Note that if we used a more refined optimization algorithm such as evolutionary Markov chain Monte Carlo algorithms rather than a simple coordinate-wise updating optimization, then the result would be further improved. See [10, 11].

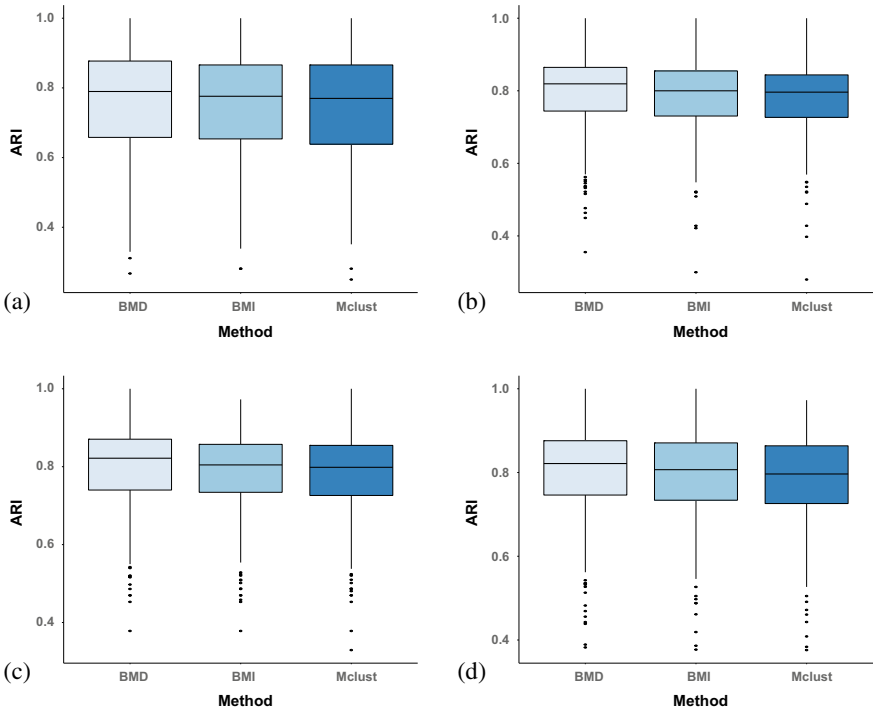


Fig. 17.1 Box plots of adjusted Rand index values corresponding to the classifications performed by applying Bayesian mixture model with weight-dependent priors (BMD), Bayesian mixture model with weight-independent priors (BMI) and the non-Bayesian mixture of normals (Mclust), where $N_0 = 2.6/(y_{\max} - y_{\min})$ and $b_0 \sim G(0.2, 10/R^2)$ where R^2 is the length of the interval of the variation of the data. Other hyperparameters and sample size are chosen as follows: (a) $N = 50$, $a_0 = 2$, $e_0 = 1$, (b) $N = 100$, $a_0 = 2$, $e_0 = 1$, (c) $N = 100$, $a_0 = 5$, $e_0 = 1$, (d) $N = 100$, $a_0 = 5$, $e_0 = 2$

17.4 Application to a Real Dataset

We applied to the so-called 'acidity data', which concerns an acidity index measured in a sample of 155 lakes in north-central Wisconsin and was previously analysed using a Bayesian mixture of Gaussian distributions on the log-scale by [8]. These authors calculated the posterior for K (the number of components) favours $3 \sim 5$ components. Here, letting $K = 3$, we applied the BMD, BMI and Mclust to the dataset respectively. The three clustering results presented in Fig. 17.2 reveal that BMD performed better in dealing with outliers in the dataset: Unlike BMD, both Cluster 2 derived from BMI or Mclust contained 3 outliers which should belong to Cluster 1.

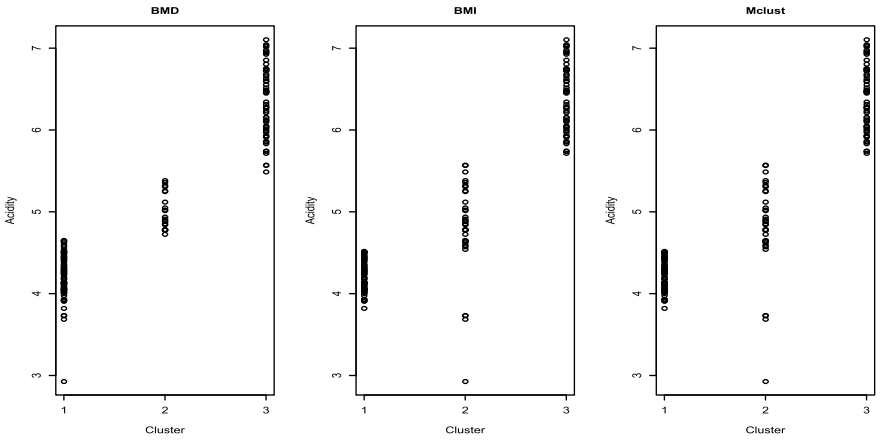


Fig. 17.2 From left to right, the panel presented the clusters derived from BMD, BMI and Mclust. We used the same approach to set hyperparameters as in our simulation study

17.5 Conclusion

In this paper, we have developed a novel prior scheme for Bayesian mixture models. Unlike the classical prior specification, we allow the component priors to depend on their weights (i.e., mixing proportions). This help us tackle the effect of varying weights on estimation of hidden group memberships of the observations. We have conducted a simulation study to compare the proposed method to the existing approaches. The simulation results have shown that the new method can performed better than its competitors in terms of adjusted Rand index. A real data application has suggested that the proposal method is more robust to outliers than the existing methods.

Acknowledgements The research of Elaheh Oftadeh is supported by the 50th anniversary PhD scholarship from the University of Kent.

References

1. Binder, D.A.: Bayesian cluster analysis. Biometrika **65**, 31–38 (1978)
2. Diebolt, J., Robert, C.P.: Estimation of finite mixture distributions through Bayesian sampling. J. R. Stat. Soc. Series B **56**, 363–375 (1994)
3. Frühwirth-Schnatter, S.: Finite Mixture and Markov Switching Models: Modeling and Applications to Random Processes. Springer, New York (2006)
4. Fraley, C., Raftery, A.E.: Model-based clustering, discriminant analysis, and density estimation. J. Am. Stat. Assoc. **971**, 611–631 (2002)
5. Hubert, L., Arabie, P.: Comparing partitions. J. Classif. **2**, 193–218 (1985)

6. Raftery, A.E.: Hypothesis testing and model selection. In: Gilks, W.R., Spiegelhalter, D.J., Richardson, S. (eds.) Markov Chain Monte Carlo in Practice, pp. 163–188. Chapman and Hall, London (1996)
7. Rand, W.M.: Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc. **66**, 846–850 (1971)
8. Richardson, S., Green, P. J.: On Bayesian analysis of mixtures with an unknown number of components (with discussions), J. R. Stat. Soc. Series B (statistical methodology) **5**9, 731–792 (1997)
9. Roeder, K., Wasserman, L.: Practical Bayesian density estimation using mixtures of normals. J. Am. Stat. Assoc. **92**, 894–902 (1997)
10. Zhang, J.: A Bayesian model for biclustering with applications. JRSSC (Applied Statistics) **59**, 635–656 (2010)
11. Zhang, J.: Genralized plaid models. Neurocomputing **79**, 95–104 (2012)

Chapter 18
Cosine Similarity-Based Classifiers for Functional Data

Tianming Zhu and Jin-Ting Zhang

Abstract In many situations, functional observations in a class are also similar in shape. A variety of functional dissimilarity measures have been widely used in many pattern recognition applications. However, they do not take the shape similarity of functional data into account. Cosine similarity is a measure that assesses how related are two patterns by looking at the angle instead of magnitude. Thus, we generalize the concept of cosine similarity between two random vectors to the functional setting. Some of the main characteristics of the functional cosine similarity are shown. Based on it, we define a new semi-distance for functional data, namely, functional cosine distance. Combining it with the centroid and k-nearest neighbors (kNN) classifiers, we propose two cosine similarity-based classifiers. Some theoretical properties of the cosine similarity-based centroid classifier are also studied. The performance of the cosine similarity-based classifiers is compared with some existing centroid and kNN classifiers based on other dissimilarity measures. It turns out that the proposed classifiers for functional data perform well in our simulation study and a real-life data example.

18.1 Introduction

Functional data consists of functions. In recent decades, it is prevalent in many fields such as economics, biology, finance, and meteorology (for an overview, see [14]). The goals of the functional data analysis (FDA) are essentially the same as those of any other branch of statistics [13]. References [5, 13] provided broad overviews of the techniques of FDA. In this paper, we are interested in supervised classification for functional data.

T. Zhu · J.-T. Zhang (✉)
Department of Statistics and Applied Probability, National University of Singapore, Singapore 117546, Singapore
e-mail: stazjt@nus.edu.sg

T. Zhu
e-mail: stazt@nus.edu.sg

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_18

Supervised classification is one of the oldest statistical problems in experimental science. We have a training sample and a test sample whose class memberships are known. The aim of classification is to create a method for assigning a new coming observation to one of the predefined classes based on the training sample. Its classification accuracy can be assessed via the misclassification error rate (MER) of the test sample. Many supervised classification methods for functional data have been developed in recent years. A number of studies have extended the traditional classification methods for multivariate data to the context of functional data. For instance, [1] proposed to filter the training samples of functional observations using the Fourier basis so that the classical kNN classifier can be applied to the resulting Fourier coefficients. [15] extended the methodology based on support vector machine for functional data. In addition, a centroid method for classifying functional observations has been adopted by [3]. They used the project of each functional observation onto a given direction instead of the functional observation itself so that a functional data classification problem becomes a one-dimensional classification problem. Further, [11] extended linear discriminant analysis to functional data. References [8–10, 17] proposed classifiers based on functional principal components while [4] developed functional classifiers based on shape descriptors.

The concepts of similarity and distance are fundamentally important in almost every scientific field. Similarity and distance measures are also an essential requirement in almost all pattern recognition applications including classification, clustering, outlier detection, regression and so on. There exist a large number of similarity measures in the literature and the performance of any pattern recognition technique largely depends on the choice of the similarity measures. In the recent literature on functional data, some authors have proposed semi-distances well adapted for sample functions such as the semi-distances based on functional principal components [5] and the functional Mahalanobis semi-distance [7]. However, most of the similarity measures are used in multivariate data and have not been extended to the functional framework. Our first contribution is to extend the cosine similarity to functional settings and define a new semi-distance for functional data.

The cosine similarity measure can be defined between two functional observations. If these two functional observations are similar in shape, this functional cosine similarity measure will be close to 1; if they are not similar or even opposite in shape, the associated cosine similarity measure will be small or even be negative. Therefore, it can be used to classify functional data. Our second contribution is that by combining the new functional semi-distance with the centroid and kNN classifiers, we propose the cosine similarity-based classifiers for functional data.

The rest of this work is organized as follows. We review a number of dissimilarity measures for functional data in Sect. 18.2. Section 18.3 introduces the concept of functional cosine similarity (FCS) and shows its main characteristics. Based on FCS, we define functional cosine distance (FCD). Section 18.4 develops the FCD-based centroid and kNN classifiers for functional data. In particular, the asymptotic MER of the FCD-based centroid classifier for functional data is derived. A simulation study for comparing the proposed cosine similarity-based centroid and kNN classifiers against other existing centroid and kNN classifiers is presented in Sect. 18.5.

Applications of the proposed cosine similarity-based centroid and kNN classifiers to a real-life data example is given in Sect. 18.6. Some concluding remarks are given in Sect. 18.7. The proofs of the main theoretical results are given in the Appendix.

18.2 Functional Dissimilarity Measures

In this section, we review some dissimilarity measures for functional data. In practice, functional data are obtained via observing some measure over time, and we assume the sample of functional observations was generated from a stochastic process.

Let $\mathscr{T}$ be some compact set. Let $x(t), t \in \mathscr{T}$ be a stochastic process having mean function $\eta(t), t \in \mathscr{T}$ and covariance function $\gamma(s,t), s,t \in \mathscr{T}$. We write $x(t) \sim \mathrm{SP}(\eta, \gamma)$ for simplicity. Throughout this work, let $\mathscr{T}$ be a finite interval, and we use $\|x\|_p$ to denote the L^p-norm of a function $x(t), t \in \mathscr{T}$: $\|x\|_p = \left(\int_{\mathscr{T}} |x(t)|^p dt\right)^{1/p}$, for $p = 1, 2, \ldots$. When $p = 2$, we may use $\|\cdot\|$ to denote the L^2-norm for simplicity. If $\|x\|_p < \infty$, we say $x(t), t \in \mathscr{T}$ is L^p-integrable. In this case, we write $x \in \mathscr{L}^p(\mathscr{T})$ where $\mathscr{L}^p(\mathscr{T})$ denotes the Hilbert space formed by all the L^p integrable functions over $\mathscr{T}$. In particular, $\mathscr{L}^2(\mathscr{T})$ denotes the Hilbert space formed by all the squared integrable functions over $\mathscr{T}$, which is an inner product space. The associated inner-product for any two functions in $\mathscr{L}^2(\mathscr{T})$ is defined as $< x, y >= \int_{\mathscr{T}} x(t)y(t)dt,\ x(t), y(t) \in \mathscr{L}^2(\mathscr{T})$. The above L^p-norm and inner-product definitions can be used to define various dissimilarity measures. Let $x(t)$ and $y(t)$ be two functional observations defined over $\mathscr{T}$, which are L^p integrable. The L^p-distance between $x(t)$ and $y(t)$ is then defined as:

$$d_p(x, y) = \|x - y\|_p,$$

for $p = 1, 2, \ldots$. We often use L^1, L^2, and L^∞-distances. It is well known that $d_\infty(x, y) = \|x - y\|_\infty = \sup_{t \in \mathscr{T}} |x(t) - y(t)|$.

The L^p-distances can be implemented easily in supervised classification but they do not take the correlation of a functional observation into account. To partially address this issue, [7] proposed the so-called functional Mahalanobis semi-distance so that the correlation structure of functional observations can be taken into account partially. The functional Mahalanobis semi-distance is defined using a number of the largest eigenvalues and the associated eigenfunctions. Note that when the covariance function $\gamma(s,t)$ has a finite trace, i.e., $\mathrm{tr}(\gamma) = \int_{\mathscr{T}} \gamma(t,t)dt < \infty$, it has the following singular value decomposition ([18], p. 3): $\gamma(s,t) = \sum_{r=1}^{\infty} \lambda_r \phi_r(s)\phi_r(t)$, where $\lambda_r, r = 1, 2, \ldots$ are the decreasing-ordered eigenvalues of $\gamma(s,t)$, and $\phi_r(t)$, $r = 1, 2, \ldots$ are the associated orthonormal eigenfunctions.

Let $y(t) \sim \mathrm{SP}(\eta, \gamma)$. By assuming $\gamma(s,t)$ has a finite trace, we have the following Karhunen-Loéve expansion: $y(t) = \sum_{r=1}^{\infty} \xi_r \phi_r(t)$, where $\xi_r =< y, \phi_r >, r = 1, 2, \ldots$ denote the associated principal component scores of $y(t)$. Let $x(t)$ be another functional observation whose covariance function is also $\gamma(s,t)$. Then we can also

expand $x(t)$ in terms of the eigenfunctions of $\gamma(s,t)$ as $x(t) = \sum_{r=1}^{\infty} \zeta_r \phi_r(t)$, where $\zeta_r =< x, \phi_r >, r = 1, 2, \ldots$ denote the associated principal component scores of $x(t)$. Then, the functional Mahalanobis (FM) semi-distance between $x(t)$ and $y(t)$ is given by

$$d_{FM,q}(x, y) = \left(\sum_{r=1}^{q} \lambda_r^{-1} (\zeta_r - \xi_r)^2 \right)^{1/2}.$$

Based on the principal component scores of $x(t)$ and $y(t)$, [5] defined the so-called functional principal components (FPC) based semi-distance which can be used as a dissimilarity measurement:

$$d_{FPC,q}(x, y) = \left(\sum_{r=1}^{q} (\zeta_r - \xi_r)^2 \right)^{1/2}.$$

Based on these dissimilarity measures, a number of classifiers are adopted for functional data. However, all these dissimilarity measures do not take the shape similarity of the functional data into account. Note that in many situations, functional observations in one class are also similar in shape. To take this information into account, in the next section, we introduce the cosine similarity measure for functional data.

18.3 Functional Cosine Similarity

The main goal of this section is to generalize the cosine similarity measure between two random vectors to the functional settings. The cosine similarity measure between two n-dimensional vectors $\mathbf{x}$ and $\mathbf{y}$ is defined as: $\mathrm{CS}(\mathbf{x}, \mathbf{y}) = \frac{<\mathbf{x},\mathbf{y}>}{\|\mathbf{x}\|\|\mathbf{y}\|}$, where $\|\cdot\|$ and $<\cdot,\cdot>$ denote the usual Euclidean norm and the usual inner product in $\mathbf{R}^n$. It is seen that the cosine similarity measure is the ratio of the inner product between the two vectors to the product of their Euclidean norms. The main characteristic of the cosine similarity measure is that it measures the closeness or similarity between two vectors using the cosine value of the angle between the two vectors, which takes value between $[-1, 1]$. It is thus a judgment of orientation and not magnitude. If two vectors have the same orientation, they have a cosine similarity measure of 1; if two vectors are orthogonal, they have a cosine similarity measure of 0; if two vectors have exactly opposite orientations, they have a cosine similarity measure of -1. When two vectors are similar, this similarity measure will take larger values.

We now extend the above cosine similarity measure to for functional data. Let $x(t), t \in \mathscr{T}$ and $y(t), t \in \mathscr{T}$ be any two functions in $\mathscr{L}^2(\mathscr{T})$. Then the functional cosine similarity (FCS) measure between $x(t)$ and $y(t)$ can be defined as follows:

$$\mathrm{FCS}(x, y) = \frac{< x, y >}{\|x\|\|y\|},$$

where $\|\cdot\|$ and $<\cdot,\cdot>$ denote the usual L^2-norm and the usual inner product in $\mathscr{L}^2(\mathscr{T})$ as defined before. It is seen that $\mathrm{FCS}(x, y)$ measures the similarity or closeness between $x(t)$ and $y(t)$ using the cosine value of the angle between the two functions $x(t)$ and $y(t)$ which was proposed by [13]. It has the following properties: (1) $-1 \leq \mathrm{FCS}(x, y) \leq 1$, normalization; (2) $\mathrm{FCS}(x, y) = \mathrm{FCS}(y, x)$, symmetry or commutativity; (3) $x(t) = y(t) \Rightarrow \mathrm{FCS}(x, y) = 1$, reflexivity; and (4) $\mathrm{FCS}(x, y) =< \tilde{x}, \tilde{y} >= 1 - \|\tilde{x} - \tilde{y}\|^2/2$ where $\tilde{x}(t) = x(t)/\|x\|$ denotes the normalization version of $x(t)$ and $\tilde{y}(t)$ is similarly defined.

Item (1) says that $\mathrm{FCS}(x, y)$ ranges from -1 (when $x(t)$ is exactly opposite to $y(t)$) to 1 (when $x(t)$ and $y(t)$ are proportional, that is, when $x(t) = ay(t)$) and takes value 0 when $x(t)$ and $y(t)$ are orthogonal. It is due to the fact that $-\|x\|\|y\| \leq < x, y > \leq \|x\|\|y\|$ by the well-known Cauchy-Schwarz inequality between two squared-integrable functions. Items (2) and (3) are obviously held. Item (4) can be shown via some simple algebra. It says that the cosine similarity measure between $x(t)$ and $y(t)$ is exactly 1 minus half of the squared L^2-norm of the difference between their normalization versions $\tilde{x}(t)$ and $\tilde{y}(t)$. Note that $\tilde{x}(t)$ is also called the spatial sign function of $x(t)$ [16], which can be interpreted as the direction of $x(t)$. Thus, the functional cosine similarity measure $\mathrm{FCS}(x, y)$ also can be interpreted as the similarity measure between the directions of $x(t)$ and $y(t)$. If $\tilde{x}(t) = \tilde{y}(t)$, that is, $x(t)$ and $y(t)$ have the same direction, the associated $\mathrm{FCS}(x, y)$ takes value 1.

Note that FCS is not a distance or semi-distance since it is not nonnegative and its value is not 0 when the two functions $x(t)$ and $y(t)$ are exactly the same. However, this can be easily corrected. For this purpose, we define the following functional cosine distance (FCD) between two functions $x(t), t \in \mathscr{T}$ and $y(t), t \in \mathscr{T}$:

$$\mathrm{FCD}(x, y) = [2 - 2\mathrm{FCS}(x, y)]^{1/2} = \left(2 - 2\frac{< x, y >}{\|x\|\|y\|}\right)^{1/2} = \|\tilde{x} - \tilde{y}\|. \quad (18.1)$$

It is obvious that $\mathrm{FCD}(x, y) = 0$ if $x(t)$ and $y(t)$ are exactly the same. Further, we have (1) $0 \leq \mathrm{FCD}(x, y) \leq 2$; (2) $\mathrm{FCD}(x, y) = \mathrm{FCD}(y, x)$, symmetry; and (3) $\mathrm{FCD}(x, y) \leq \mathrm{FCD}(x, z) + \mathrm{FCD}(y, z)$ for any three functions $x(t), t \in \mathscr{T}$, $y(t), t \in \mathscr{T}$ and $z(t), t \in \mathscr{T}$, triangle inequality.

Using the properties of $\mathrm{FCS}(x, y)$, it is easy to verify the first two properties of $\mathrm{FCD}(x, y)$ above. Item (3) can be shown by the well-known Minkowski inequality. Consequently, FCD is a functional semi-distance since $\mathrm{FCD}(x, y) = 0$ cannot imply $x(t) = y(t), t \in \mathscr{T}$. Nevertheless, we can define some classifiers based on FCD for functional data.

18.4 Cosine Similarity-Based Classifiers for Functional Data

Let $G \geq 2$ be an integer. Suppose we have G training functional samples

$$x_{i1}(t), x_{i2}(t), \ldots, x_{in_i}(t) \overset{\text{i.i.d.}}{\sim} \text{SP}(\eta_i, \gamma_i), \quad i = 1, \ldots, G, \tag{18.2}$$

where $\eta_i(t)$'s are the unknown group mean functions and $\gamma_i(s, t)$'s are the unknown group covariance functions. Note that throughout this work, we assume that the functional observations of the same group are i.i.d. and functional observations of different groups are also independent. For a new coming functional observation $x(t)$, our aim is to determine the class membership of $x(t)$ based on the above G training samples.

In this section, our aim is to propose new nonparametric classifiers via combining the centroid and kNN classifiers with FCD. The resulting classifiers are called the FCD-based centroid and kNN classifiers, respectively.

18.4.1 FCD-Based Centroid Classifier

There are many different approaches which can design a nonparametric classifier. The first one, also the simplest one, is based on the concept of similarity. Observations that are similar should be assigned to the same class. Thus, once the similarity measure is established, the new coming observation can be classified accordingly. The choice of the similarity measure is crucial to the success of this approach. The first representative of this approach is the nearest mean classifier, also called nearest centroid classifier. Each class is represented by its mean of all the training patterns in that class. A new observation will be assigned to the class whose mean is closest to the new observation.

For functional data, the class center is the group mean function which can be estimated using its usual group sample mean function. For the G training functional samples (18.2), the G class centers can be estimated as $\bar{x}_i(t) = n_i^{-1} \sum_{j=1}^{n_i} x_{ij}(t)$, $i = 1, \ldots, G$. Then the FCDs between the new coming functional observation $x(t)$ and the above class centers $\bar{x}_i(t), i = 1, \ldots, G$ can be expressed as $\text{FCD}(x, \bar{x}_i), i = 1, \ldots, G$. The FCD-based centroid classifier for functional data then puts $x(t)$ into Class g where

$$g = \text{argmin}_{1 \leq i \leq G} \text{FCD}^2(x, \bar{x}_i). \tag{18.3}$$

18.4.2 FCD-Based kNN Classifier

The classical kNN classifier was first proposed by [6]. Due to its simplicity and efficiency, it is widely used to perform supervised classification in multivariate settings. The classical kNN classifier consists of the following steps: given a training sample with known class labels, classify a new observation into a class by examining its k nearest neighbors and applying the majority vote rule.

For the G training functional samples (18.2), the FCDs between the coming functional observation $x(t)$ and all the training functional observations can be computed as $\mathrm{FCD}(x, x_{ij}),\ j = 1, \ldots, n_i;\ i = 1, \ldots, G$. Let k be some given integer. The $x_{ij}(t)$'s associated with the k smallest values of the above FCDs are the k nearest neighbors of $x(t)$ from the whole training functional sample. Let m_i denote the number of the nearest neighbors from Class i where $i = 1, \ldots, G$. Then $\sum_{i=1}^{G} m_i = k$. Note that some of m_i's are equal to each other and some are equal to 0. The FCD-based kNN classifier for functional data then puts $x(t)$ into Class g where $g = \mathrm{argmax}_{1 \le i \le G} m_i$.

18.4.3 Theoretical Properties of the FCD-Based Centroid Classifier

In this subsection, we study the theoretical properties of the FCD-based centroid classifier. That is, we shall derive its asymptotic misclassification error rate (MER) and show some of its good properties. Recall that $x(t)$ denotes the new coming functional observation. As mentioned in the previous subsection, for the G training functional samples (18.2), the FCD-based centroid classifier will put $x(t)$ to Class g determined by (18.3). For each class $g = 1, \ldots, G$, we have a classification vector function $\mathbf{T}_g(x)$ based on the G-class FCD-based centroid classifier which can be expressed as $\mathbf{T}_g(x) = [T_{g,1}(x), \ldots, T_{g,g-1}(x), T_{g,g+1}(x), \ldots, T_{g,G}(x)]^T$, where $T_{g,i}(x) = \mathrm{FCD}^2(x, \bar{x}_i) - \mathrm{FCD}^2(x, \bar{x}_g)$ for $i = 1, \ldots g-1, g+1, \ldots, G$. Then the G-class FCD-based centroid classifier for functional data assigns $x(t)$ to class g if $\mathbf{T}_g(x) > \mathbf{0}$, where $\mathbf{0}$ denotes the zero vector.

Let π_i denote the probability that $x(t)$ from Class i for $i = 1, \ldots, G$. Assuming that $\mathrm{tr}(\gamma_i) < \infty$, $i = 1, 2, \ldots, G$, we can show that as $n_i, i = 1, 2, \ldots, G$ tend to infinity with $n_i/n \to \tau_i > 0$ where $n = n_1 + n_2 + \cdots + n_G$, we have $\bar{x}_i(t) \to \eta_i(t),\ i = 1, 2, \ldots, G$ uniformly over the compact set $\mathscr{T}$ so that the classification vector functions $\mathbf{T}_g(x)$, $g = 1, \ldots, G$ will tend to

$$\mathbf{T}_g^*(x) = [T_{g,1}^*(x), \ldots, T_{g,g-1}^*(x), T_{g,g+1}^*(x), \ldots, T_{g,G}^*(x)]^T, \tag{18.4}$$

where $T_{g,i}^*(x) = \mathrm{FCD}^2(x, \eta_i) - \mathrm{FCD}^2(x, \eta_g)$ for $i = 1, \ldots, g-1, g+1, \ldots, G$.

For further discussion, let $\mathscr{C}_i$ denote Class i for $i = 1, \ldots, G$. The prior probabilities of Class i can then be expressed as $\pi_i = \Pr(x \in \mathscr{C}_i)$. For a G-class classification problem, a mistake is made when $x \in \mathscr{C}_g$, by using the classifier, we

assign it to Class i, $i \neq g$. Therefore, the MER of the G-class FCD-based centroid classifier can then be expressed as $\text{MER} = \sum_{g=1}^{G} \pi_g \Big(1 - \Pr\{\mathbf{T}_g(x) > \mathbf{0} | x \in \mathscr{C}_g\}\Big) = 1 - \sum_{g=1}^{G} \pi_g \Pr\{\mathbf{T}_g(x) > \mathbf{0} | x \in \mathscr{C}_g\}$. The asymptotic MER of the FCD-based centroid classifier is presented in Theorem 18.1.

Theorem 18.1 *Assume the G training functional samples (18.2) are independent with $\text{tr}(\gamma_i) < \infty, i = 1, \ldots, G$. In addition, as $n \to \infty$, we have $n_i/n \to \tau_i > 0$. Then as $n \to \infty$, we have the following asymptotic MER of the FCD-based centroid classifier:*

$$MER \to MER^* = 1 - \sum_{g=1}^{G} \pi_g F_{\mathbf{R}_g}(\boldsymbol{\Sigma}_g^{-1/2} \boldsymbol{\mu}_g), \tag{18.5}$$

where for $g = 1, \ldots, G$, $\boldsymbol{\mu}_g = [\mu_{g,1}, \ldots, \mu_{g,g-1}, \mu_{g,g+1}, \ldots, \mu_{g,G}]^T$, and $\boldsymbol{\Sigma}_g = (\sigma^2_{g_i,g_l}) : (G-1) \times (G-1)$, with $\mu_{g,i} = \|\eta_g\| FCD^2(\eta_i, \eta_g), i = 1, \ldots, g-1, g+1, \ldots, G$, and $\sigma^2_{g_i,g_l} = 4 \int_{\mathscr{T}} \int_{\mathscr{T}} [\tilde{\eta}_i(s) - \tilde{\eta}_g(s)] \gamma_g(s,t) [\tilde{\eta}_l(t) - \tilde{\eta}_g(t)] ds dt, i, l \in \{1, \ldots, g-1, g+1, \ldots, G\}$. In addition, $F_{\mathbf{R}_g}(\cdot), g = 1, \ldots, G$ denotes the cumulative distribution functions of some random variable $\mathbf{R}_g$ which has zero mean vector $\mathbf{0}$ and identity covariance matrix $\mathbf{I}$.

Remark 18.2 The expression (18.5) indicates that the asymptotic MER may not tend to 0 even when the group sample sizes tend to infinity. Note that when MER is 0, there is a perfect classification. However, whether we can have a perfect classification is determined by the data information. If the data are not separable, we cannot have a perfect classification even when the sizes of training samples diverge.

When $G = 2$, the G-class FCD-based centroid classifier reduces to a two-class one. In this case, the results in Theorem 18.1 can be simplified. In addition, we can give an upper error bound of the associated MER. We now denote $\pi_1 = \pi$ and $\pi_2 = 1 - \pi$. The classification function of the two-class FCD-based centroid classifier can then be simply expressed as

$$T(x) = \text{FCD}^2(x, \bar{x}_2) - \text{FCD}^2(x, \bar{x}_1). \tag{18.6}$$

As $n_i, i = 1, 2$ tend to infinity with $n_1/n \to \tau > 0$, $T(x)$ will tend to

$$T^*(x) = \text{FCD}^2(x, \eta_2) - \text{FCD}^2(x, \eta_1). \tag{18.7}$$

Therefore, the MER of the two-class FCD-based centroid classifier $T(x)$ can then be expressed as $\text{MER} = \pi \Pr\{T(x) \leq 0 | x \in \mathscr{C}_1\} + (1 - \pi) \Pr\{T(x) > 0 | x \in \mathscr{C}_2\}$. By Theorem 18.1, we present the asymptotic MER of the two-class FCD-based centroid classifier and its upper bound in Theorem 18.3 below.

Theorem 18.3 *Assume the $(G = 2)$ training functional samples (18.2) are independent with $\text{tr}(\gamma_i) < \infty, i = 1, 2$. In addition, as $n \to \infty$, we have $n_1/n \to \tau > 0$. Then as $n \to \infty$, when we use the FCD-based centroid classifier, we have*

$$MER \to MER^* = \pi F_{R_1}(-\mu_1/\sigma_1) + (1-\pi)[1 - F_{R_2}(\mu_2/\sigma_2)], \tag{18.8}$$

where $\mu_i = \|\eta_i\| FCD^2(\eta_1, \eta_2)$, *and* $\sigma_i^2 = 4\int_{\mathscr{T}}\int_{\mathscr{T}}[\tilde{\eta}_1(s) - \tilde{\eta}_2(s)]\gamma_i(s,t)[\tilde{\eta}_1(t) - \tilde{\eta}_2(t)]dsdt, i = 1, 2$. $F_{R_i}(\cdot)$ *is the cumulative distribution function of some random variable* R_i *which has mean* 0 *and variance* 1. *Further, the upper bound of the asymptotical MER (18.8) is given by the following expression*

$$MER^* \le \pi F_{R_1}\left(-\frac{\|\eta_1\| FCD(\eta_1, \eta_2)}{2\sqrt{\lambda_{1,\max}}}\right) + (1-\pi)\left[1 - F_{R_2}\left(\frac{\|\eta_2\| FCD(\eta_1, \eta_2)}{2\sqrt{\lambda_{2,\max}}}\right)\right], \tag{18.9}$$

where $\lambda_{i,\max}$ *denote the largest eigenvalue of* $\gamma_i(s,t)$ *for* $i = 1, 2$. *In particular, when the functional data are Gaussian,* $F_{R_1}(\cdot)$ *and* $F_{R_2}(\cdot)$ *should also be replaced with* $\Phi(\cdot)$, *the cumulative distribution function of the standard normal distribution.*

Remark 18.4 The asymptotic MER (18.8) will become smaller if the group mean functions $\eta_1(t)$ and $\eta_2(t)$ become less similar from each other, that is, $FCD(\eta_1, \eta_2)$ becomes larger. This is reasonable. If the group mean functions are not similar, it is easy to classify the new coming observation correctly. In addition, the upper bound of the asymptotic MER (18.9) indicates the smaller the value of $\lambda_{i,\max}, i = 1, 2$ are, the smaller the value of MER^*. This is also reasonable since when $\lambda_{i,\max}, i = 1, 2$ are small, the data are less noisy. Thus, it is easier to classify the new coming functional observation $x(t)$ correctly.

Remark 18.5 If the data are Gaussian, the expression (18.9) indicates that for Gaussian functional data, we always have $MER^* < 1/2$ as long as $FCD(\eta_1, \eta_2) > 0$. That is, the worse case of this two-class FCD-based centroid classifier is better than of the random guessing.

18.5 A Simulation Study

To demonstrate the good performance of the proposed cosine similarity-based classifiers for functional data, we conduct a simulation study in this section. The results of the simulation study allow us to compare the proposed FCD-based centroid and kNN classifiers against some existing centroid and kNN classifiers based on other dissimilarity measures. The centroid and kNN classifiers are defined similarly to the FCD-based centroid and kNN classifiers for functional data as in Sects. 18.4.1 and 18.4.2 except replacing the FCD with one of the dissimilarity measures reviewed in Sect. 18.2. These dissimilarity measures include the L^p-distances for $p = 1, 2$, and ∞, the functional Mahalanobis (FM) semi-distance assuming a common covariance function, and the functional principal components (FPC) semi-distance assuming a common covariance function, as defined in Sect. 18.2. The resulting centroid or kNN classifiers are labeled with L^1, L^2, L^∞, FPC, and FM respectively.

We consider generating functional data for a two-class classification problem under four different scenarios. In the first scenario, two functional samples are generated from two Gaussian processes defined over $I = [0, 1]$, with different group mean functions $\eta_1(t) = 25t^{1.1}(1-t)$ and $\eta_2(t) = 25t(1-t)^{1.1}$ but their covariance functions $\gamma_1(s,t)$ and $\gamma_2(s,t)$ are the same, denoted as $\gamma(s,t)$ whose eigenfunctions are given by $\phi_r(t) = \sqrt{2}\sin(r\pi t)$, $r = 1, 2, \ldots$ and the associated eigenvalues are given by $\lambda_r = 1/(r\pi)^2$, for $r = 1, 2, \ldots$. The generated functions are evaluated at 1000 equidistant time points over $I = [0, 1]$. In the second scenario, the functions are generated in a similar way except that the two covariance functions $\gamma_1(s,t)$ and $\gamma_2(s,t)$ are not the same although their eigenfunctions are the same as those defined in the first scenario but their eigenvalues are given by $\lambda_{1r} = 1/(r\pi)^2$ and $\lambda_{2r} = 2/(r\pi)^2$, for $r = 1, 2, \ldots$ respectively. In the third and fourth scenarios, the functions are generated in a similar way as in the first and second scenarios respectively except the two Gaussian processes are replaced with two standardized exponential processes with rate 1 with the same group mean functions and the group covariance functions.

Under each scenario, two functional samples of equal sizes 100 are generated. The training sample is formed via selecting 50 functions from each sample so that the whole training sample consists of 100 functional observations. The remaining functional observations from the two functional samples form the test sample. The training sample is used to determine the tuning parameters. In particular, we use the 10-fold cross-validation approach. For a kNN classifier, the possible number of nearest neighbors k ranges from 1 to 25. In order to avoid ties, we also set k to be odd numbers only. Similarly, the number of principal components q used in the centroid or kNN classifiers ranges from 1 to q_0 where q_0 may be chosen such that the sum of the first q_0 eigenvalues of the pooled sample covariance function $\widehat{\gamma}(s,t)$ is about 95% of the total variation given by $\mathrm{tr}(\widehat{\gamma})$. Note that the accuracy of a centroid or kNN classifier is measured by its MER which is estimated using the test sample. We repeat the process 1000 times so that we have 1000 MERs. The boxplots of the 1000 MERs of the test samples under the four scenarios are shown in Fig. 18.1.

In view of this figure, it is seen that under the fourth scenario, FCD-based centroid classifier outperforms other centroid classifiers and the FCD-based kNN classifier outperforms other kNN classifiers as well. In the third scenario, the best performance is attended by the proposed FCD-based centroid classifier. Therefore, Gaussianity is not necessarily an advantage for the FCD-based classifiers and they perform well for non-Gaussian data. In practice, it is usually very difficult to check the Gaussianity, hence our proposed classifiers may work well in real problems. In addition, in the first and second scenarios, our FCD-based classifiers perform the second best and FM-based classifiers perform best. However, the FM semi-distance is a rather complicated dissimilarity measure and consumes time in programing and computing.

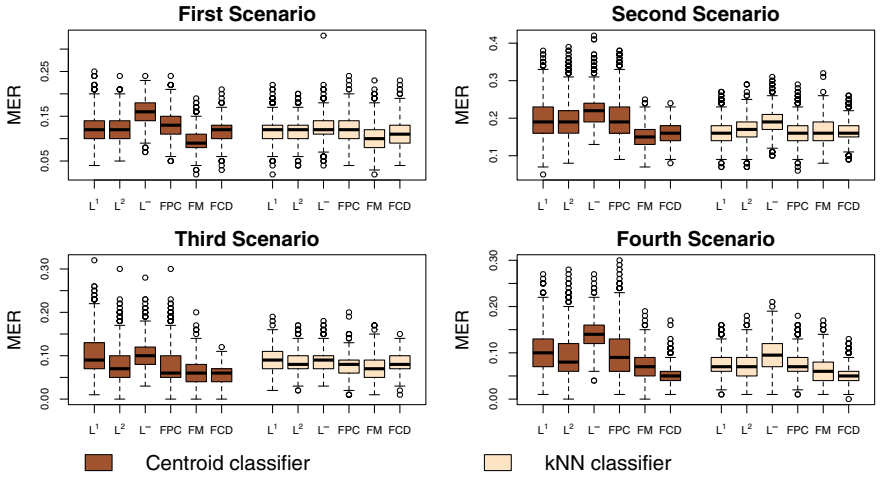


Fig. 18.1 MERs achieved by various centroid and kNN classifiers under all four scenarios

18.6 Application to Australian Rainfall Data

The Australian rainfall data set is available at https://rda.ucar.edu/datasets/ds482.1/. It has been analyzed by [2, 12] respectively to illustrate their classification methodologies. The data set consists of daily rainfall measurements between January 1840 and December 1990, at each of 191 Australian weather stations. The daily rainfall measurements of a station form a rainfall curve. We then have $N = 191$ rainfall curves. Among the 191 weather stations, $N_1 = 43$ of them are located at the northern Australia and the remaining ones are located at the southern Australia. For each station, for simplicity, we just consider the rainfall over a year, i.e., over $t \in [1, 365]$. As in [2], a rainfall curve for a station is obtained via taking the average of the rainfall at each time point $t \in [1, 365]$ over the years which the station had been operating. The resulting raw rainfall curves are then smoothed using a B-spline basis of order 6. The order of B-spline basis is chosen by leave-one-out cross-validation so that the raw rainfall curves can be well represented by the smoothed rainfall curves as shown in Fig. 18.2. From this figure, we can see that some of the weather stations, although geographically located in the north, have a rainfall pattern that is typical of the south. Thus, it is not so easy to distinguish the northern rainfall curves from the southern rainfall curves.

To apply the centroid and kNN classifiers for the Australian rainfall data, we randomly split the 191 rainfall curves into a training sample of size n and a test sample of size $191 - n$ and we take $n = 50$. The number of nearest neighbors is bounded by the smaller sample size of the two classes and the maximum number of eigenfunctions is limited to 20. This process is repeated 1000 times so that we have 1000 MERs for each classifier. Figure 18.3 presents the boxplots of the 1000 MERs of the various centroid and kNN classifiers. It is seen that the FCD-based

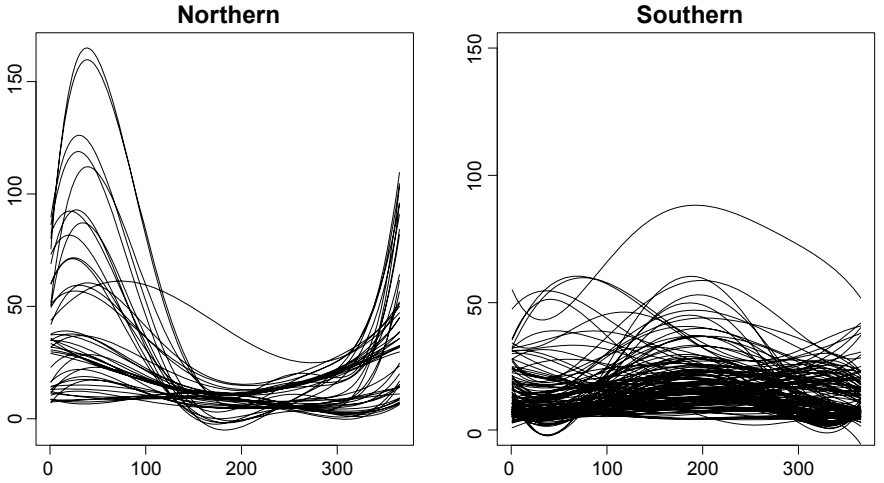


Fig. 18.2 Smoothed Australian rainfall curves for the northern weather stations (left panel) and the southern weather stations (right panel)

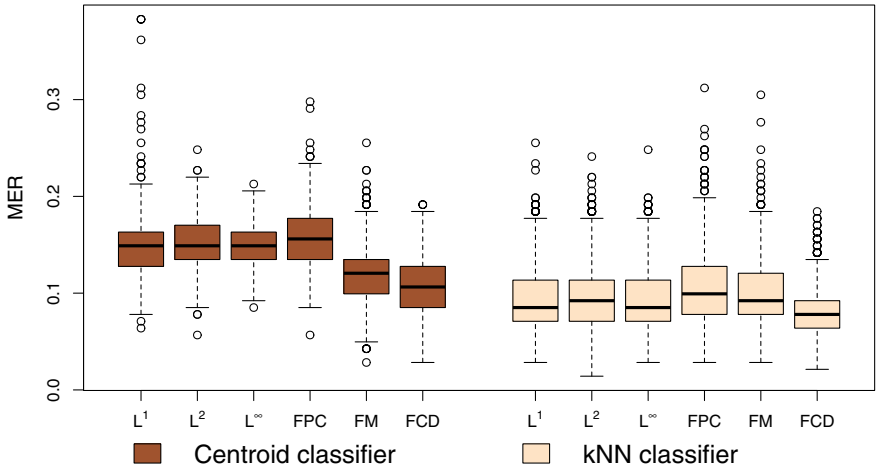


Fig. 18.3 MERs achieved by various centroid and kNN classifiers for the Australian rainfall data

kNN classifier performs best and it obtained mean MER of 0.079. Moreover, the FCD-based centroid classifier outperforms other centroid classifiers which obtained mean MER of 0.106. It is also seen that the kNN classifiers are generally better than the centroid classifiers with the same dissimilarity measures. Using a similar experiment, [3] obtained mean MERs of 0.103 by the centroid classifier which was proposed by [3].

18.7 Concluding Remarks

In this work, we extend the cosine similarity measure for functional data. Based on the FCS, we introduce a new semi-distance for functional data named FCD. This functional semi-distance is simple and can be implemented easily in supervised classification. By combining with the centroid and kNN classifiers, we propose a FCD-based centroid classifier and a FCD-based kNN classifier for functional data. We also study the theoretical properties of the FCD-based centroid classifier. It turns out the cosine similarity-based classifiers for functional data perform well in our simulation study and a real-life data example. As mentioned previously, the range of applications for the new similarity measure or the new functional semi-distance is wide and includes clustering, hypothesis testing, and outlier detection, among others. However, since the proposed FCD does not take the magnitude of the functional data into account, it is expected that the proposed FCD-based classifiers will not perform well for classifying functional data which are different only in their magnitudes. It is then interesting and warranted to study how both the magnitude and shape of the data can be taken into account in FCD-based classifiers so that their performance can be further improved.

18.8 Appendix

Proof (Proof of Theorem 18.1). Under the given conditions, since $\mathrm{tr}(\gamma_i) < \infty$, $i = 1, 2, \ldots, G$, as $n \to \infty$ with $n_i/n \to \tau_i > 0$, we have

$$\mathrm{MER} \to \mathrm{MER}^* = 1 - \sum_{g=1}^{G} \pi_g \Pr\{\mathbf{T}_g^*(x) > \mathbf{0} | x \in \mathscr{C}_g\}, \tag{18.10}$$

where $\mathbf{T}_g^*(x)$ is given in (18.4). Set $\mathbf{S}_g^*(x) = \|x\| \mathbf{T}_g^*(x)$, then we have

$$\mathbf{S}_g^*(x) = [S_{g,1}^*(x), \ldots, S_{g,g-1}^*(x), S_{g,g+1}^*, \ldots, S_{g,G}^*(x)]^T,$$

where $S_{g,i}^*(x) = 2 < x, \tilde{\eta}_g - \tilde{\eta}_i >$, $i = 1, \ldots, g-1, g+1, \ldots, G$. Since $\|x\| > 0$, we have

$$\mathrm{MER}^* = 1 - \sum_{g=1}^{G} \pi_g \Pr\{\mathbf{S}_g^*(x) > \mathbf{0} | x \in \mathscr{C}_g\}. \tag{18.11}$$

When $x \in \mathscr{C}_g$, for $i = 1, \ldots, g-1, g+1, \ldots, G$, we have

$$\mu_{g,i} = \mathbb{E}\left\{S_{g,i}^*(x) | x \in \mathscr{C}_g\right\} = 2 < \eta_g, \tilde{\eta}_g - \tilde{\eta}_i >= \|\eta_g\| \mathrm{FCD}^2(\eta_i, \eta_g),$$

and for any $i, l \in \{1, \ldots, g-1, g+1, \ldots, G\}$,

$$\sigma^2_{g_i,g_l} = \mathrm{Cov}\{S^*_{g,i}(x), S^*_{g,l}(x)|x \in \mathscr{C}_g\}$$
$$= 4\int_{\mathscr{T}}\int_{\mathscr{T}} [\tilde{\eta}_i(s) - \tilde{\eta}_g(s)]\gamma_g(s,t)[\tilde{\eta}_l(t) - \tilde{\eta}_g(t)]dsdt.$$

Therefore, we have

$$\boldsymbol{\mu}_g = \mathbb{E}\{\mathbf{S}^*_g(x)|x \in \mathscr{C}_g\} = [\mu_{g,1}, \ldots, \mu_{g,g-1}, \mu_{g,g+1}, \ldots, \mu_{g,G}]^T,$$
$$\boldsymbol{\Sigma}_g = \mathrm{Cov}\{\mathbf{S}^*_g(x)|x \in \mathscr{C}_g\} = (\sigma^2_{g_i,g_l}) : (G-1) \times (G-1).$$

We can then write $\Pr\{\mathbf{S}^*_g(x) > \mathbf{0}|x \in \mathscr{C}_g\} = \Pr\{\mathbf{R}_g < \boldsymbol{\Sigma}_g^{-1/2}\boldsymbol{\mu}_g|x \in \mathscr{C}_g\}$, where

$$\mathbf{R}_g = \boldsymbol{\Sigma}_g^{-1/2}(-\mathbf{S}^*_g(x) + \boldsymbol{\mu}_g),$$

which is a random variable with mean vector $\mathbf{0}$ and covariance matrix $\mathbf{I}$. Therefore,

$$\mathrm{MER}^* = 1 - \sum_{g=1}^{G} \pi_g F_{\mathbf{R}_g}(\boldsymbol{\Sigma}_g^{-1/2}\boldsymbol{\mu}_g), \tag{18.12}$$

as desired where $F_{\mathbf{R}_g}(\cdot)$, $g = 1, \ldots, G$ denote the cumulative distribution functions of $\mathbf{R}_g$, $g = 1, \ldots, G$. □

Proof (Proof of Theorem 18.3) Under Theorem 18.1, when $G = 2$, the classification function of the two-class FCD-based centroid classifier can be simply expressed as (18.6). As $n_i, i = 1, 2$ tend to infinity with $n_1/n \to \tau > 0$, $T(x)$ will tend to (18.7). Thus, the corresponding $S^*(x) = 2 < x, \tilde{\eta}_1 - \tilde{\eta}_2 >$ is a one-dimensional random variable.

When $x \in \mathscr{C}_1$, we have

$$\mu_1 = \mathbb{E}\left\{S^*(x)|x \in \mathscr{C}_1\right\} = 2 < \eta_1, \tilde{\eta}_1 - \tilde{\eta}_2 >= \|\eta_1\|\mathrm{FCD}^2(\eta_1, \eta_2),$$
$$\sigma_1^2 = \mathrm{Var}\left\{S^*(x)|x \in \mathscr{C}_1\right\} = 4\int_{\mathscr{T}}\int_{\mathscr{T}} [\tilde{\eta}_1(s) - \tilde{\eta}_2(s)]\gamma_1(s,t)[\tilde{\eta}_1(t) - \tilde{\eta}_2(t)]dsdt.$$

We can then write $\Pr\{S^*(x) \le 0|x \in \mathscr{C}_1\} = \Pr(R_1 \le -\mu_1/\sigma_1)$ where

$$R_1 = (S^*(x) - \mu_1)/\sigma_1,$$

which is a random variable with mean 0 and variance 1. Similarly, we can show that $\Pr\{S^*(x) > 0|x \in \mathscr{C}_2\} = \Pr(R_2 > \mu_2/\sigma_2)$ where

$$R_2 = (S^*(x) + \mu_2)/\sigma_2,$$

which is a random variable with mean 0 and variance 1, and

$$\begin{aligned}\mu_2 &= -\mathbb{E}\left\{S^*(x)|x \in \mathscr{C}_2\right\}\\ &= -2 < \eta_2, \tilde{\eta}_1 - \tilde{\eta}_2 >= \|\eta_2\|\mathrm{FCD}^2(\eta_1, \eta_2),\\ \sigma_2^2 &= \mathrm{Var}\left\{S^*(x)|x \in \mathscr{C}_2\right\}\\ &= 4\int_{\mathscr{T}}\int_{\mathscr{T}}[\tilde{\eta}_1(s) - \tilde{\eta}_2(s)]\gamma_2(s,t)[\tilde{\eta}_1(t) - \tilde{\eta}_2(t)]dsdt.\end{aligned}$$

Therefore

$$\begin{aligned}\mathrm{MER}^* &= \pi\,\Pr\{S^*(x) \le 0|x \in \mathscr{C}_1\} + (1-\pi)\Pr\{S^*(x) > 0|x \in \mathscr{C}_2\}\\ &= \pi\,\Pr(R_1 \le -\mu_1/\sigma_1) + (1-\pi)\Pr(R_2 > \mu_2/\sigma_2)\\ &= \pi F_{R_1}(-\mu_1/\sigma_1) + (1-\pi)\left[1 - F_{R_2}(\mu_2/\sigma_2)\right],\end{aligned} \tag{18.13}$$

as desired where $F_{R_i}(\cdot), i = 1, 2$ denote the cumulative distribution functions of $R_i, i = 1, 2$.

Let $\lambda_{i,\max}$ denote the largest eigenvalue of $\gamma_i(s,t)$ for $i = 1, 2$. Then we have

$$\begin{aligned}\sigma_i^2 &= 4\int_{\mathscr{T}}\int_{\mathscr{T}}[\tilde{\eta}_1(s) - \tilde{\eta}_2(s)]\gamma_i(s,t)[\tilde{\eta}_1(t) - \tilde{\eta}_2(t)]dsdt\\ &\le 4\lambda_{i,\max}\|\tilde{\eta}_1 - \tilde{\eta}_2\|^2 = 4\lambda_{i,\max}\mathrm{FCD}^2(\eta_1, \eta_2), \quad i = 1, 2.\end{aligned}$$

It follows that

$$\mu_i/\sigma_i \ge \frac{\|\eta_i\|\mathrm{FCD}^2(\eta_1,\eta_2)}{\sqrt{4\lambda_{i,\max}\mathrm{FCD}^2(\eta_1,\eta_2)}} = \frac{\|\eta_i\|\mathrm{FCD}(\eta_1,\eta_2)}{2\sqrt{\lambda_{i,\max}}},\ i = 1, 2.$$

Therefore, by (18.13), we have

$$\mathrm{MER}^* \le \pi F_{R_1}\left(-\frac{\|\eta_1\|\mathrm{FCD}(\eta_1,\eta_2)}{2\sqrt{\lambda_{i,\max}}}\right) + (1-\pi)\left[1 - F_{R_2}\left(\frac{\|\eta_2\|\mathrm{FCD}(\eta_1,\eta_2)}{2\sqrt{\lambda_{i,\max}}}\right)\right]. \tag{18.14}$$

When the functional data are Gaussian, we have $R_i \sim N(0,1), i = 1, 2$. Therefore, we should replace $F_{R_i}(\cdot), i = 1, 2$ in the expressions (18.13) and (18.14) with $\Phi(\cdot)$, the cumulative distribution of the standard normal distribution. □

References

1. Biau, G., Bunea, F., Wegkamp, M.H.: Functional classification in hilbert spaces. IEEE Trans. Inf. Theory **51**(6), 2163–2172 (2005). https://doi.org/10.1109/TIT.2005.847705
2. Delaigle, A., Hall, P.: Defining probability density for a distribution of random functions. Ann. Stat. **38**(2), 1171–1193 (2010)
3. Delaigle, A., Hall, P.: Achieving near perfect classification for functional data. J. R. Stat. Soc.: Ser. B (Statistical Methodology) **74**(2), 267–286 (2012)

4. Epifanio, I.: Shape descriptors for classification of functional data. Technometrics **50**(3), 284–294 (2008)
5. Ferraty, F., Vieu, P.: Nonparametric Functional Data Analysis: Theory and Practice. Springer Series in Statistics. Springer, New York (2006)
6. Fix, E., Hodges Jr, J.L.: Discriminatory analysis: nonparametric discrimination: consistency properties. US Air Force School of Aviation Medicine. Technical report, vol. 4(3), 477+ (1951)
7. Galeano, P., Joseph, E., Lillo, R.E.: The mahalanobis distance for functional data with applications to classification. Technometrics **57**(2), 281–291 (2015)
8. Glendinning, R.H., Herbert, R.: Shape classification using smooth principal components. Pattern Recogn. Lett. **24**(12), 2021–2030 (2003)
9. Hall, P., Poskitt, D.S., Presnell, B.: A functional datałanalytic approach to signal discrimination. Technometrics **43**(1), 1–9 (2001)
10. Huang, D.S., Zheng, C.H.: Independent component analysis-based penalized discriminant method for tumor classification using gene expression data. Bioinformatics **22**(15), 1855–1862 (2006)
11. James, G.M., Hastie, T.J.: Functional linear discriminant analysis for irregularly sampled curves. Journal of the Royal Statistical Society: Series B (Statistical Methodology) **63**(3), 533–550 (2001)
12. Lavery, B., Joung, G., Nicholls, N.: A historical rainfall data set for Australia. Australian Meteorol. Mag. **46** (1997)
13. Ramsay, J., Ramsay, J., Silverman, B.: Functional Data Analysis. Springer Series in Statistics. Springer, New York (2005)
14. Ramsay, J.O., Silverman, B.W.: Applied Functional Data Analysis: Methods and Case Studies. Springer Series in Statistics. Springer, New York (2002)
15. Rossi, F., Villa, N.: Support vector machine for functional data classification. Neurocomputing **69**(7), 730–742 (2006)
16. Sguera, C., Galeano, P., Lillo, R.: Spatial depth-based classification for functional data. Test **23**(4), 725–750 (2014)
17. Song, J.J., Deng, W., Lee, H.J., Kwon, D.: Optimal classification for time-course gene expression data using functional data analysis. Comput. Biol. Chem. **32**(6), 426–432 (2008)
18. Wahba, G.: Spline models for observational data. Soc. Ind. Appl. Math. **59** (1990)

Part V
Hypothesis Test and Statistical Models

Chapter 19
Projection Test with Sparse Optimal Direction for High-Dimensional One Sample Mean Problem

Wanjun Liu and Runze Li

Abstract Testing whether the mean vector from some population is zero or not is a fundamental problem in statistics. In the high-dimensional regime, where the dimension of data p is greater than the sample size n, traditional methods such as Hotelling's T^2 test cannot be directly applied. One can project the high-dimensional vector onto a space of low dimension and then traditional methods can be applied. In this paper, we propose a projection test based on a new estimation of the optimal projection direction $\Sigma^{-1}\mu$. Under the assumption that the optimal projection $\Sigma^{-1}\mu$ is sparse, we use a regularized quadratic programming with nonconvex penalty and linear constraint to estimate it. Simulation studies and real data analysis are conducted to examine the finite sample performance of different tests in terms of type I error and power.

19.1 Introduction

One-sample mean vector test or two-sample test on the equality of two means is a fundamental problem in high-dimensional data analysis. These tests are commonly encountered in genome-wide association studies. For instance, [6] performed a hypothesis testing to identify sets of genes which are significant with respect to certain treatments in a genetics research. Reference [21] applied various tests to the bipolar disorder dataset from a genome-wide association study collected by [7] in which one would like to test whether there is any association between a disease and a large number of genetic variants. In these applications, the dimension of the data p is often much larger than the sample size n. Traditional methods such as Hotelling's T^2 test [13] either cannot be directly applied or have low power against the alternative.

W. Liu (✉) · R. Li
Department of Statistics, Pennsylvania State University,
University Park 16802-2111, USA
e-mail: wxl204@psu.edu

R. Li
e-mail: rzli@psu.edu

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_19

Suppose that a random sample $x_1, \ldots, x_n$ from a p-dimensional population x with finite mean $\mathrm{E}(x) = \mu$ and positive definite covariance matrix $\mathrm{cov}(x) = \Sigma$. Of interest is to test the following hypothesis

$$H_0 : \mu = \mu_0 \quad \text{versus} \quad H_1 : \mu \neq \mu_0, \tag{19.1}$$

for some known vector μ_0. This problem is typically referred to as the one-sample hypothesis testing problem in multivariate analysis and has been extensively studied when $p < n$ and p is fixed. Without loss of generality, we assume $\mu_0 = 0$ and the one-sample problem (19.1) becomes

$$H_0 : \mu = 0 \quad \text{versus} \quad H_1 : \mu \neq 0. \tag{19.2}$$

In most cases, the test statistic constructed for one-sample problem can be easily extended to two-sample problem and the theories hold as well. For this reason, we only focus on the one-sample problem (19.2) and assume $\mu_0 = 0$. Let $\bar{x}$ and S be the sample mean vector and the sample covariance matrix respectively,

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad S = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(x_i - \bar{x})^\top. \tag{19.3}$$

The Hotelling's T^2 statistic [13] for problem (19.2) is $T^2 = n\bar{x}^\top S^{-1}\bar{x}$. If $x_1, \ldots, x_n$ are normally distributed, under H_0, then we have $(n-p)/\{(n-1)p\}T^2$ follows $F_{p,n-p}$, the F distribution with degrees of freedom p and $n-p$. The Hotelling's T^2 requires that the sample covariance matrix S is invertible and cannot be directly used in high-dimensional setting where $p > n$. Despite the singularity of S, it has been observed that the power of the Hotelling's T^2 test can be adversely affected even when $p < n$, if S is nearly singular; see [1, 17].

Several one-sample tests for high-dimensional data have been proposed recently. These tests can be roughly classified into three types. The first type is the sum-of-squares-type test which is based on the sum-of-squares of the sample mean and can be regarded as modified versions of Hotelling's T^2 test. These tests simply replace S by some invertible matrix such as identity matrix I or diagonal matrix, leading to a sum-of-squares test statistic. Bai and Saranadasa [1] proposed the following test statistic for one-sample problem, in which S is substituted by identity matrix I,

$$T_{BS} = \bar{x}^\top\bar{x} - \mathrm{tr}S/n.$$

The test statistic T_{BS} can be regarded as unscaled distance $\bar{x}^\top\bar{x}$ with offset $\mathrm{tr}S/n$. Bai and Saranadasa [1] established its asymptotic normal null distribution when $p/n \to c$ for some $c > 0$. Chen and Qin [6] further studied an equivalent form of T_{BS}:

$$T_{CQ} = \frac{1}{n(n-1)}\sum_{i\neq j}^{n} x_i^\top x_j.$$

under a different set of assumptions on population. Neither T_{BS} nor T_{CQ} is invariant under different scales. To get rid of the unit effect, [19] replaced S with diagonal matrix D, where $D = \text{diag}(S)$ is a diagonal matrix with diagonal elements from the sample covariance matrix S. The test statistic in [19] is defined as

$$T_{SD} = n\bar{x}^\top D^{-1}\bar{x} - (n-1)p/(n-3),$$

which is also asymptotically normally distributed under null hypothesis.

The second type is the maximum-type test. Cai et al. [4] introduced a test that is based on a linear transformation of the data by the precision matrix $\Omega = \Sigma^{-1}$ which incorporates the correlations among the variables. Given that the precision matrix $\Omega = (\omega_{ij})_{p\times p}$ is known, the test statistic is defined as

$$T_{CLX} = n \max_{1\le j\le p} (\Omega\bar{x})_j^2/\omega_{jj}. \tag{19.4}$$

If Ω is known to be sparse, then the CLIME estimator [3] can be used to estimate Ω directly. Otherwise, Ω can be estimated by the inverse of the adaptive thresholding estimator of Σ [2]. Under H_0, the test statistic T_{CLX} converges to the type I extreme value distribution. Chen et al. [5] proposed a test that removes components that are estimated to be zero via thresholding. The motivation is that zero components are expected to contribute little to the squared sample mean and those smaller than a given threshold can be ignored. The test statistic with index s is defined as

$$T_{CLZ}(s) = \sum_{j=1}^{p} \left\{ \frac{n\bar{x}_j^2}{\sigma_{jj}} - 1 \right\} I\left\{ \frac{n\bar{x}_j^2}{\sigma_{jj}} > \lambda_p(s) \right\},$$

where the threshold level is set to be $\lambda_p(s) = 2s \log p$ for some $s \in (0, 1)$. Since the optimal choice of the threshold is unknown, [5] further proposed using s that results in the largest value of $T_{CLZ}(s)$ as the final test statistic,

$$T_{CLZ} = \max_{s\in(0,1-\eta)} \{T_{CLZ}(s) - \widehat{\mu}_{CLZ,0}(s)\}/\widehat{\sigma}_{CLZ,0}(s),$$

for some $\eta \in (0, 1)$, where $\widehat{\mu}_{CLZ,0}(s)$ and $\widehat{\sigma}_{CLZ,0}(s)$ are estimates of the mean and standard deviation of $T_{CLZ}(s)$ under H_0. The asymptotic null distribution of T_{CLZ} is the Gumbel distribution.

The third type is the projection test. The idea is to project the high-dimensional vector x onto a space of low dimension and then traditional methods such as Hotelling's T^2 can be applied. Lauter [14] proposed the following procedure for the one-sample normal mean problem based on left-spherical distribution theory [11, 12]. Consider the linear score $z = (z_1, \ldots, z_n)^\top = Xd$, where d is a $p \times 1$ weight

vector depending on X only through $X^\top X$ and $d \neq 0$ with probability 1. Then one can perform the one-sample t-test using $z_1, \ldots, z_n$. Lauter [14] also proposed different ways to obtain the weight vector d. For example, d can take the form of $d = (\text{diag}(X^\top X))^{-1/2}$, or be the eigenvector corresponding to the largest eigenvalue $\lambda_{\max}$ for the following eigenvalue problem $(X^\top X)d = \text{diag}(X^\top X)d\lambda_{\max}$. Lopes et al. [16] proposed a test based on random projection. Let P_k be a $p \times k$ random matrix whose entries are randomly drawn from the $N(0, 1)$ distribution. Define $y_i = P_k^\top x_i, i = 1, \ldots, n$. The random projection test T_{RP} in [16] is defined as

$$T_{RP} = n\bar{y}^\top S_y^{-1}\bar{y} = n\bar{x}^\top P_k(P_k^\top S P_k)^{-1}P_k^\top \bar{x},$$

where $\bar{y}$ and S_y are the sample mean and sample covariance matrix of $y_1, \ldots, y_n$. As a result, this random projection test is the Hotelling's T^2 test with $y_1, \ldots, y_n$ and is an exact test if x_i's are normally distributed. Lopes et al. [16] also proposed a test that utilizes multiple projection to improve the power of random projection test. The idea is generating the projection matrix P_k multiple times and using their average as the final projection matrix.

These types of tests are powerful only against certain alternatives. For example, if the true mean μ is dense in the sense that there is a large proportion of small to moderate nonzero components, then sum-of-squares-type test is more powerful. In contrast, if the true mean μ is sparse in the sense that there are only few nonzero components with large magnitude in μ, then the maximum-type test is more powerful. In practice, since the true alternative hypothesis is unknown, it is unclear how to choose a powerful test. Furthermore, there are denser and intermediate situations in which neither type of test is powerful [21].

Li et al. [15] studied the projection test and derived the optimal projection direction which leads to the best power under alternative hypothesis. However, the estimation of the optimal projection direction has not been systematically studied yet. This paper aims to fill this gap by studying how to construct a sparse optimal projection test to achieve better power. We propose an estimation procedure of the sparse optimal projection direction by regularized quadratic programming with nonconvex penalty and linear constraint. We further examine the finite sample performance of the proposed procedure and illustrate it by an empirical analysis of a real data set.

The rest of this paper is organized as follows. In Sect. 19.2, we propose a new projection test with the optimal projection being estimated by the regularized quadratic programming. In Sect. 19.3, simulation studies are conducted to examine the finite sample performance of different tests in terms of type I error and power. In Sect. 19.4, we apply various tests to a real data example, which shows that the proposed projection test is more powerful than existing tests.

19.2 Projection Test with Sparse Optimal Direction

Li et al. [15] proposed an exact projection test using the optimal projection direction. They showed that the optimal choice of k in P_k is 1 and the optimal projection is $\Sigma^{-1}\mu$ in the sense that the power is maximized. Let $\theta = \Sigma^{-1}\mu$ and $y_i = \theta^\top x_i, i = 1, \ldots, n$. The projection Hotelling's T^2 test is

$$T_\theta^2 = n\bar{x}^\top\theta(\theta^\top S\theta)^{-1}\theta^\top\bar{x},$$

which follows the $F_{1,n-1}$ distribution under H_0 and normality assumption. It is equivalent to the one-sample t-test based on $y_1, \ldots, y_n$. In order to control the type I error, [15] also proposed a data-splitting strategy to estimate the optimal projection direction and obtained an exact t-test. The entire sample is randomly partitioned into two separate sets $\mathscr{S}_1 = \{x_1, \ldots, x_{n1}\}$ and $\mathscr{S}_2 = \{x_{n1+1}, \ldots, x_n\}$. Set $\mathscr{S}_1$ is used to estimate the projection direction θ and set $\mathscr{S}_2$ is used to construct the test statistic T_θ^2. To estimate θ, a ridge-type estimator is constructed $\widehat{\theta} = (S_1 + \lambda D_1)^{-1}\bar{x}_1$, where $\bar{x}_1$ and S_1 are the sample mean and the sample covariance matrix computed from $\mathscr{S}_1$ and $D_1 = \text{diag}(S_1)$, the diagonal matrix of S_1. Therefore, the estimator $\widehat{\theta}$ is independent of set $\mathscr{S}_2$. Then the data points from $\mathscr{S}_2$ are projected onto a 1-dimensional space by left-multiplying $\widehat{\theta}$. The one-sample t-test is performed based on the new data points $\widehat{\theta}^\top x_{n_1+1}, \ldots, \widehat{\theta}^\top x_n$. In order to have high power, [15] recommended to use $n_1 = \lfloor \kappa n \rfloor$ with $\kappa \in [0.4, 0.6]$ and $\lambda = n_1^{-1/2}$ in practice based on their empirical study. If κ is small, only a small portion of sample is used to estimate the optimal projection and the estimator is not accurate. If κ is large, only a small portion of sample is used to perform the test. As a result, a too small or too large κ leads to significant loss in the power of the test. The advantage of the data-splitting procedure is that we can obtain an exact t-test, meanwhile we may lose power since the sample in $\mathscr{S}_1$ is discarded when performing the test.

We propose a new estimation of the optimal projection under the assumption that the optimal projection $\Sigma^{-1}\mu$ is sparse. The assumption that the optimal projection direction is sparse is relatively mild and can be satisfied in different scenarios. For example, if Σ has the autocorrelation structure and μ is sparse and then the optimal projection $\Sigma^{-1}\mu$ is sparse. Another example is that if Σ has the compound symmetry structure and μ is sparse and then $\Sigma^{-1}\mu$ is approximately sparse in the sense that the first few entries in $\Sigma^{-1}\mu$ dominate the rest entries. Note that it is the direction rather than the magnitude of the projection $\Sigma^{-1}\mu$ that matters. In other words, $\Sigma^{-1}\mu$ and $a\Sigma^{-1}\mu$ have exactly the same performance for the one-sample problem (19.2), where a is some positive number. We observe that $\beta^* = \Sigma^{-1}\mu/\mu^\top\Sigma^{-1}\mu$, which is proportional to the optimal projection, is the solution to the following problem

$$\min_\beta \frac{1}{2}\beta^\top\Sigma\beta \text{ subject to } \mu^\top\beta = 1.$$

Based on the above observation, we propose the following estimation based on a regularized quadratic programming with nonconvex penalty and linear constraint,

$$\begin{aligned} &\min_{\beta} \frac{1}{2}\beta^\top S_1 \beta + \sum_{j=1}^{p} p_\lambda(|\beta_j|) \\ &\text{subject to } \bar{x}_1^\top \beta = 1, \end{aligned} \tag{19.5}$$

where $\bar{x}_1$ and S_1 are computed from set $\mathscr{S}_1$, $\beta = (\beta_1, \ldots, \beta_p)^\top \in \mathbb{R}^p$ and $p_\lambda(\cdot)$ is taken to be the smoothly clipped absolute deviation (SCAD) penalty [9]. Its first derivative is defined to be

$$p'_\lambda(|t|) = \lambda \left\{ I(|t| \le \lambda) + \frac{(a\lambda - |t|)_+}{(a-1)\lambda} I(|t| > \lambda) \right\},$$

where $a = 3.7$, $I(\cdot)$ is the indicator function and b_+ stands for the positive part of b. To solve the high-dimensional nonconvex optimization problem (19.5), we apply the local linear approximation (LLA) algorithm proposed in [22]. The idea is to approximate the nonconvex penalty by its first order expansion. Given the current solution $\beta^{(k)}$, (19.5) can be approximated by

$$\begin{aligned} &\min_{\beta} \frac{1}{2}\beta^\top S_1 \beta + \sum_{j=1}^{p} p'_\lambda(|\beta_j^{(k)}|)|\beta_j|, \\ &\text{subject to } \bar{x}_1^\top \beta = 1. \end{aligned}$$

Let

$$Q(\beta|\beta^{(k)}, \lambda) = \frac{1}{2}\beta^\top S_1 \beta + \sum_{j=1}^{p} p'_\lambda(|\beta_j^{(k)}|)|\beta_j|.$$

Wang et al. [20] and Fan et al. [10] studied how to implement the LLA under high-dimensional regression settings to obtain a sparse solution with oracle property. Here we apply their strategy for the above problem. Starting with initial value 0, we propose a two-step LLA estimator, which consists of the following two steps:

$$\begin{aligned} &\textbf{Step 1}: \widehat{\beta}^{(1)} = \operatorname*{argmin}_{\{\beta: \bar{x}_1^\top \beta = 1\}} Q(\beta|0, \tau\lambda)\ ; \\ &\textbf{Step 2}: \widehat{\beta} = \operatorname*{argmin}_{\{\beta: \bar{x}_1^\top \beta = 1\}} Q(\beta|\widehat{\beta}^{(1)}, \lambda). \end{aligned}$$

The solution $\widehat{\beta}$ in step 2 is our final estimator. Typically, we choose τ to be some small number such as $\tau = 1/\log n_1$ or $\tau = \lambda$. Instead of using the ridge-type estimator $\widehat{\theta} = (S_1 + \lambda D_1)^{-1}\bar{x}_1$, we use our two-step LLA estimator $\widehat{\beta}$ to carry out the projection test with data splitting. It can be shown that the resulting LLA estimator is consistent

under relatively mild conditions and thus the asymptotic power is valid for our new test with the data-splitting procedure. We call this new test LLA projection test.

19.3 Simulation Studies

In this section, we conduct numerical studies to examine the finite sample performance of different tests including the proposed LLA projection test for the one-sample problem. The LLA projection test is the same as that in [15] except that we use the LLA estimator as the projection direction. More specifically, we compare the LLA projection test with the ones proposed by [1, 6, 8, 14, 15]. We denote them by Li2015, D1958, BS1996, CQ2010 and L1996, respectively. We also compare the new test with the tests proposed in [19]. The authors considered two versions of their test, one with modification and one without modification, denoted by SD2008w and SD2008wo, respectively. Lopes et al. [16] proposed a single random projection test, labeled as LWJ2011.

We generate a random sample of size n from $N(\mu, \Sigma)$ with $\mu = c \cdot (1_{s_0}^\top, 0_{p-s_0}^\top)^\top$ and $s_0 = 10$. We set $c = 0, 0.5$ and 1 to examine the type I error rate and the power of the tests. For $\rho \in (0, 1)$, we consider the following three covariance structures:

(1) Compound symmetry with $\Sigma_1 = (1-\rho)I + \rho 11^\top$;
(2) Autocorrelation with $\Sigma_2 = (\rho^{|i-j|})_{i,j}$;
(3) Composite structure with $\Sigma_3 = 0.5\Sigma_1 + 0.5\Sigma_2$.

We consider $\rho = 0.25, 0.5, 0.75$ and 0.95 to examine the influence of correlation on the power of the test. We set sample size $n = 40, 160$ and dimension $p = 400, 1600$. We split the data set by setting $n_1 = \lfloor n\kappa \rfloor$ with $\kappa = 0.4$, where $\lfloor \cdot \rfloor$ is the rounding operator. To this end, we replace sample covariance matrix S_1 by $S_\phi = S_1 + \phi I$ with a small positive number $\phi = \sqrt{\log p/n_1}$. Such a perturbation does not noticeably affect the computational accuracy of the final solution and all the theoretical properties hold as well when $\phi \le \sqrt{\log p/n_1}$. All simulation results are based on 10,000 independent replicates. These results are summarized in Tables 19.1, 19.2 and 19.3.

Tables 19.1, 19.2 and 19.3 clearly indicate that the LLA projection test and the tests in [14–16] keep the type I error very well. This is not surprising since all these tests are exact tests. All other tests do not keep the type I error rate well because their critical values are determined from the asymptotic distributions. Next we compare the power of the LLA projection test with other existing methods. It can be seen from Tables 19.1, 19.2 and 19.3, the power of the tests strongly relies on the covariance structure as well as the values of ρ and c.

Table 19.1 reports the results for the compound symmetry covariance structure Σ_1. We first compare the LLA projection test and the Li2015 test since these two tests are of the same flavor but using different methods to estimate the projection direction. When we have relatively large sample size $n = 160$, both of the tests have high power and the LLA projection test slightly improves the performance of Li2015

Table 19.1 Power comparison for $N(\mu, \Sigma_1)$ (values in table are in percentage)

	$c = 0$				$c = 0.5$				$c = 1$			
ρ	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95
$n = 40$, $p = 400$												
LLA	4.98	4.50	4.94	5.19	71.53	89.92	99.00	99.96	99.97	99.88	99.99	100.0
Li2015	5.16	4.47	4.88	4.90	50.22	70.74	94.04	100.0	98.61	99.53	100.0	100.0
D1958	6.77	6.22	5.71	5.49	12.63	8.13	6.98	6.46	80.44	22.71	13.06	10.23
BS1996	7.73	7.80	7.79	7.80	14.64	10.55	9.39	9.11	88.21	30.28	18.51	15.33
CQ2010	7.72	7.82	7.79	7.77	14.64	10.50	9.41	9.11	88.18	30.22	18.50	15.32
SD2008w	4.20	1.71	0.52	0.15	7.97	2.29	0.63	0.22	54.21	6.41	1.29	0.36
SD2008wo	8.48	8.21	7.87	7.71	16.34	11.15	9.53	8.96	90.25	32.69	18.93	15.06
L1996	5.18	5.18	5.17	5.15	5.66	5.21	5.17	5.11	6.25	5.59	5.31	5.22
LJW2011	5.01	4.99	4.86	5.03	13.80	20.65	40.58	98.34	54.05	74.46	95.94	100.0
$n = 40$, $p = 1600$												
LLA	5.22	4.99	5.21	5.08	50.43	79.97	98.51	99.98	99.92	99.94	99.99	100.0
Li2015	5.01	4.71	5.06	4.94	14.62	23.71	54.68	98.14	71.49	81.98	95.74	100.0
D1958	6.93	6.19	5.73	5.48	7.81	6.71	5.96	5.73	12.45	8.12	6.96	6.45
BS1996	7.74	7.79	7.78	7.79	8.85	8.30	8.12	8.06	14.47	10.42	9.35	8.92
CQ2010	7.76	7.80	7.77	7.77	8.88	8.32	8.14	8.05	14.49	10.41	9.34	8.92
SD2008w	2.76	0.69	0.14	0.00	3.22	0.73	0.17	0.00	5.26	0.97	0.20	0.01
SD2008wo	8.37	8.12	7.86	7.71	9.67	8.70	8.24	7.94	15.79	11.14	9.52	8.78
L1996	5.15	5.15	5.15	5.15	5.30	5.22	5.19	5.18	5.50	5.29	5.23	5.18
LJW2011	4.65	5.08	4.99	4.95	6.77	7.68	11.52	52.16	14.49	20.55	42.17	98.29
$n = 160$, $p = 400$												
LLA	4.77	5.10	4.96	4.83	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Li2015	4.97	4.89	4.80	4.99	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
D1958	5.74	5.26	4.89	4.77	87.20	19.91	12.09	9.73	100.0	100.0	99.95	89.50
BS1996	6.66	6.71	6.69	6.71	94.00	26.42	16.69	13.78	100.0	100.0	100.0	99.41
CQ2010	6.66	6.71	6.69	6.71	94.02	26.45	16.69	13.76	100.0	100.0	100.0	99.39
SD2008w	3.11	0.99	0.34	0.07	50.59	3.63	0.72	0.17	100.0	92.93	7.69	1.27
SD2008wo	6.87	6.83	6.71	6.65	94.39	26.76	16.83	13.72	100.0	100.0	100.0	99.36
L1996	4.76	4.74	4.74	4.73	5.98	5.23	4.95	4.87	7.08	5.88	5.32	5.11
LJW2011	4.81	5.15	4.99	4.84	98.07	99.92	100.0	100.0	100.0	100.0	100.0	100.0
$n = 160$, $p = 1600$												
LLA	4.63	4.99	4.79	4.96	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0
Li2015	4.91	5.43	5.40	4.74	98.84	99.92	100.0	100.0	100.0	100.0	100.0	100.0
D1958	5.76	5.22	4.87	4.77	11.15	7.32	6.11	5.68	93.07	19.49	11.60	9.41
BS1996	6.71	6.69	6.69	6.69	13.09	9.46	8.37	8.11	97.90	26.09	16.49	13.60
CQ2010	6.71	6.69	6.70	6.70	13.10	9.47	8.37	8.11	97.91	26.11	16.48	13.61
SD2008w	2.10	0.40	0.05	0.02	3.82	0.53	0.05	0.02	29.18	1.19	0.15	0.03
SD2008wo	6.90	6.82	6.71	6.66	13.48	9.46	8.43	8.09	98.05	26.51	16.39	13.53
L1996	4.72	4.73	4.73	4.74	4.76	4.71	4.69	4.71	4.93	4.73	4.73	4.71
LJW2011	5.23	4.83	4.80	4.70	34.28	55.48	91.85	100.0	98.27	99.95	100.0	100.0

Table 19.2 Power comparison for $N(\mu, \Sigma_2)$ (values in table are in percentage)

	$c = 0$				$c = 0.5$				$c = 1$			
ρ	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95
$n = 40, p = 400$												
LLA	5.18	5.19	5.26	4.78	61.15	50.17	36.46	27.19	100.0	99.99	99.67	97.72
Li2015	5.29	4.46	5.16	4.81	46.27	35.27	21.13	13.86	99.98	99.53	91.08	68.03
D1958	5.06	4.97	4.75	5.30	89.47	77.24	51.45	17.29	100.0	100.0	99.96	84.57
BS1996	5.57	5.57	5.46	6.86	90.19	78.40	53.88	20.81	100.0	100.0	99.99	88.16
CQ2010	5.59	5.57	5.44	6.85	90.16	78.39	53.83	20.81	100.0	100.0	99.99	88.15
SD2008w	3.75	3.68	3.30	2.72	84.86	70.93	44.71	9.94	100.0	100.0	99.85	68.93
SD2008wo	7.25	7.28	7.61	8.52	90.57	80.54	57.97	23.86	100.0	100.0	99.96	87.61
L1996	4.69	4.67	4.93	4.96	36.78	28.64	16.49	6.63	92.38	77.29	40.81	9.26
LJW2011	5.52	5.11	5.00	4.97	12.71	12.15	11.51	15.28	44.17	43.04	42.40	60.42
$n = 40, p = 1600$												
LLA	5.25	5.19	5.09	5.12	38.03	30.96	22.88	16.49	100.0	99.94	98.81	91.04
Li2015	4.61	4.95	5.30	4.92	17.85	14.57	9.55	6.10	94.90	84.59	58.09	22.43
D1958	4.91	5.14	4.88	4.74	48.45	37.63	23.47	9.96	99.99	99.91	94.58	42.28
BS1996	5.05	5.46	5.36	5.49	49.13	38.40	24.63	11.40	99.99	99.91	94.96	45.81
CQ2010	5.08	5.48	5.29	5.50	49.26	38.35	24.60	11.44	99.99	99.91	94.94	45.74
SD2008w	1.77	1.91	2.04	1.81	30.97	22.82	12.73	3.66	99.92	99.03	86.28	23.53
SD2008wo	7.04	7.13	7.19	7.55	53.38	43.79	29.11	14.45	99.98	99.79	95.07	50.80
L1996	4.92	5.11	5.08	4.99	15.69	13.57	9.77	6.17	45.99	34.47	19.55	7.99
LJW2011	4.61	4.99	4.87	4.89	6.04	6.47	6.17	6.68	11.71	12.12	11.46	13.14
$n = 160, p = 400$												
LLA	5.10	4.94	4.82	5.03	100.0	99.97	99.37	99.98	100.0	100.0	100.0	100.0
Li2015	5.33	4.68	5.03	5.16	99.99	99.43	89.97	96.04	100.0	100.0	100.0	100.0
D1958	4.61	4.97	5.12	5.34	100.0	100.0	100.0	85.83	100.0	100.0	100.0	100.0
BS1996	5.03	5.50	5.83	6.61	100.0	100.0	100.0	89.10	100.0	100.0	100.0	100.0
CQ2010	5.03	5.49	5.83	6.62	100.0	100.0	100.0	89.10	100.0	100.0	100.0	100.0
SD2008w	4.20	4.42	4.17	2.73	100.0	100.0	100.0	72.60	100.0	100.0	100.0	100.0
SD2008wo	5.41	5.78	6.19	6.93	100.0	100.0	100.0	88.85	100.0	100.0	100.0	100.0
L1996	4.87	4.71	4.70	5.00	89.99	71.70	34.04	7.34	100.0	100.0	71.60	10.28
LJW2011	4.65	4.95	4.75	5.27	89.44	85.36	80.43	98.54	100.0	100.0	100.0	100.0
$n = 160, p = 1600$												
LLA	4.85	5.04	4.92	4.79	100.0	99.94	97.61	93.27	100.0	100.0	100.0	100.0
Li2015	5.24	4.83	4.97	5.01	97.18	88.69	61.66	35.37	100.0	100.0	100.0	99.60
D1958	4.73	4.72	4.99	5.11	99.99	99.89	95.03	42.55	100.0	100.0	100.0	100.0
BS1996	4.86	5.00	5.30	5.98	100.0	99.90	95.35	45.67	100.0	100.0	100.0	100.0
CQ2010	4.86	4.98	5.29	5.99	100.0	99.90	95.35	45.66	100.0	100.0	100.0	100.0
SD2008w	3.47	3.46	3.57	2.70	100.0	99.83	93.02	29.91	100.0	100.0	100.0	99.99
SD2008wo	5.40	5.48	5.65	6.33	100.0	99.87	95.47	46.65	100.0	100.0	100.0	100.0
L1996	5.27	5.08	4.84	4.78	42.34	31.61	16.88	6.42	97.49	83.24	39.61	9.04
LJW2011	4.86	4.67	4.56	5.45	25.35	24.48	23.41	37.24	92.06	90.94	90.47	98.77

Table 19.3 Power comparison for $N(\mu, \Sigma_3)$ (values in table are in percentage)

	$c = 0$				$c = 0.5$				$c = 1$			
ρ	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95
$n = 40$, $p = 400$												
LLA	5.13	4.61	5.68	4.93	60.97	60.90	57.77	55.26	99.99	99.86	99.81	99.73
Li2015	5.08	4.76	4.76	4.73	43.34	42.40	35.76	28.98	98.94	98.15	97.05	94.31
D1958	6.82	6.68	6.38	6.02	26.75	12.64	9.38	8.05	99.93	78.81	38.79	21.82
BS1996	7.37	7.73	7.75	7.86	29.27	14.57	11.73	10.39	99.95	86.13	48.62	29.47
CQ2010	7.40	7.71	7.71	7.89	29.30	14.54	11.68	10.35	99.96	86.09	48.67	29.42
SD2008w	5.39	4.27	2.71	1.32	20.68	8.06	4.18	1.88	99.61	52.80	15.70	5.47
SD2008wo	8.15	8.36	8.34	8.27	33.33	15.93	12.72	10.85	99.96	88.48	52.41	32.24
L1996	5.07	5.17	5.13	5.19	6.12	5.57	5.33	5.22	8.10	6.30	5.80	5.57
LJW2011	4.86	5.20	4.95	5.06	12.83	14.60	15.63	26.31	48.15	53.40	60.76	86.04
$n = 40$, $p = 1600$												
LLA	4.96	5.13	5.06	4.62	40.72	42.46	42.69	36.21	99.95	99.82	99.79	99.21
Li2015	4.60	5.13	5.16	5.05	12.67	13.26	11.95	8.45	70.81	69.11	66.18	46.23
D1958	7.17	6.90	6.48	6.19	9.58	7.78	7.18	6.81	27.30	12.45	9.31	8.28
BS1996	7.74	7.80	7.80	7.78	10.34	9.01	8.46	8.30	30.00	14.44	11.59	10.42
CQ2010	7.76	7.81	7.78	7.72	10.28	8.95	8.46	8.29	30.05	14.45	11.54	10.40
SD2008w	4.19	2.70	1.44	0.73	5.79	3.18	1.61	0.77	15.53	5.26	2.19	1.04
SD2008wo	8.48	8.43	8.32	8.19	11.70	9.76	9.05	8.82	34.82	15.94	12.59	11.17
L1996	5.10	5.14	5.19	5.20	5.39	5.27	5.21	5.21	5.71	5.44	5.30	5.31
LJW2011	5.00	4.84	4.79	5.23	6.68	6.81	7.05	8.80	13.07	14.15	16.65	23.86
$n = 160$, $p = 400$												
LLA	5.35	5.05	4.75	5.42	100.0	99.99	100.0	100.0	100.0	100.0	100.0	100.0
Li2015	5.01	5.03	4.86	5.36	100.0	100.0	99.91	99.99	100.0	100.0	100.0	100.0
D1958	5.98	5.73	5.44	5.04	100.0	83.26	33.73	18.91	100.0	100.0	100.0	100.0
BS1996	6.45	6.67	6.73	6.72	100.0	90.99	43.97	25.76	100.0	100.0	100.0	100.0
CQ2010	6.47	6.67	6.72	6.72	100.0	91.00	43.98	25.74	100.0	100.0	100.0	100.0
SD2008w	4.92	3.04	1.81	0.82	99.99	48.74	10.03	2.98	100.0	100.0	99.99	78.09
SD2008wo	6.70	6.80	6.89	6.85	100.0	91.47	45.28	26.38	100.0	100.0	100.0	100.0
L1996	4.70	4.75	4.75	4.74	7.09	6.01	5.56	5.20	10.27	7.09	6.21	5.83
LJW2011	5.36	5.28	4.97	4.91	94.31	96.07	97.07	100.0	100.0	100.0	100.0	100.0
$n = 160$, $p = 1600$												
LLA	4.99	4.85	5.10	4.63	100.00	100.0	99.99	99.96	100.0	100.0	100.0	100.0
Li2015	5.22	5.02	5.16	4.38	97.38	97.63	93.89	80.36	100.0	100.0	100.0	100.0
D1958	6.12	5.77	5.45	5.25	23.85	11.17	8.51	7.42	100.0	91.94	33.55	20.31
BS1996	6.66	6.68	6.70	6.77	26.24	13.24	10.44	9.46	100.0	97.40	44.83	27.33
CQ2010	6.67	6.68	6.70	6.77	26.26	13.22	10.41	9.44	100.0	97.42	44.81	27.32
SD2008w	4.13	2.06	0.84	0.39	15.66	3.70	1.26	0.58	100.0	29.28	4.52	1.28
SD2008wo	6.83	6.80	6.85	6.86	27.30	13.49	10.70	9.61	100.0	97.42	46.28	28.33
L1996	4.69	4.71	4.75	4.74	4.93	4.78	4.71	4.70	5.44	4.91	4.79	4.74
LJW2011	4.97	5.30	5.41	5.20	28.38	33.95	40.01	68.67	94.88	97.86	99.36	99.99

test. When we have a relatively small sample size $n = 40$, LLA projection test can dramatically improve the performance of Li2015 test especially when the signal is not strong ($c = 0.5$). A weaker correlation ρ results in a more significant improvement. This is because a small correlation makes the optimal direction $\Sigma^{-1}\mu$ closer to a sparse direction. When $c = 0.5$, the power of both tests increases significantly as ρ increases. As the value of c increases from 0.5 to 1, the power of the two tests increases dramatically. As the dimension p increases, there is a downward trend for the two tests. Even in the most challenging case $(n, p, c) = (40, 1600, 0.5)$, our LLA projection test has high power as well and is much powerful than Li2015 test. These two tests outperform all other tests. Some of the tests, such as D1958, BS1996, CQ2010 and SD2008w, tend to become less powerful when ρ increases. This is because these methods ignore the correlation among the variables and therefore their overall performance is not satisfactory.

Table 19.2 reports the results for the autocorrelation covariance structure Σ_2. Under this setting, the LLA projection test improves the performance of Li2015 test in all the combinations of c and ρ. In particular, the LLA projection test improves the performance of Li2015 dramatically when the sample size is relatively small and the correlation is large. For example, when $(n, p, c, \rho) = (40, 1600, 1, 0.95)$, the LLA test improves the power from 22.42% to 91.04%. The D1958, BS1996, CQ2010 and SD2008wo have more satisfactory performance than LLA test when $(n, c) = (40, 0.5)$ and ρ is not 0.95. Notice that the D1958, BS1996, CQ2010 and SD2008wo tests ignore the correlation among variables and replace Σ^{-1} by diagonal matrix. When Σ has the autocorrelation structure, its inverse is a 3-banded matrix – only its diagonal and first off-diagonal elements are nonzero. As a result, replacing Σ^{-1} by identity matrix does not lose much information. This explains why tests of D1958, BS1996, CQ2010 and SD2008wo have more satisfactory performance when Σ has autocorrelation structure and ρ is low. It is also observed that the power of these four tests decreases significantly as the correlation increases and become less powerful than the LLA test when $\rho = 0.95$. This is not surprising since all the four tests ignore the correlations among the variables. In general, the proposed test is preferred if Σ^{-1} is far away from identity matrix.

Table 19.3 reports the results for Σ_3. The LLA test is more powerful than Li2015 test in all the combinations of n, p, c, ρ and improves the power dramatically when ρ is large. The LLA test outperforms all other tests. The patterns for D1958, BS1996, CQ2010, SD2008w and SD2008wo are similar to the first scenario where $\Sigma = \Sigma_1$.

We also investigate the finite sample performance of the LLA projection test without the normality assumption. To this end, we generate random samples from the multivariate t distribution with degrees of freedom 6. To examine the robustness of the LLA test, we use the same critical values as those used in settings with normality assumption. Simulation results for Σ_2 are summarized in Table 19.4, from which it can be seen that the LLA test and Li2015 test can still retain the type I error rate very well. This implies that these two projection tests are not very sensitive to the normality assumption. All other alternative tests except for CQ2010 test fail to retain the type I error. In terms of power, LLA projection test is more powerful than Li2015 test in all combinations of n, p, c, ρ. For this autocorrelation covariance (i.e., Σ_2)

case, the LLA test and the CQ2010 test have similar performance and these two tests outperform all other tests. The overall patterns for Σ_1 and Σ_3 are similar to those in Tables 19.1 and 19.3. Results are not presented in this paper to save space.

19.4 Real Data Example

In this section, we apply the LLA projection test to a real dataset of high resolution micro-computed tomography. This dataset contains the bone density of 58 mice's skull of three different genotypes ("T0A0", "T0A1", and "T1A1") measured at different bone density levels in a genetic mutation study. For each mouse, bone density is measured for 16 different areas of its skull. For each area, bone volume is measured at density levels from 130 to 249. This dataset was collected at Center for Quantitative X-Ray Imaging at the Pennsylvania State University. See [18] for a detailed description of protocols. In this empirical analysis, we are interested in comparing the bone density patterns of two different areas in mice's skull. We compare the performance of the proposed LLA projection test with several existing methods. To emphasize the high-dimensionality nature of this dataset, we only use half sample of the dataset. We select the mice of the genotype "T0A1" and there are 29 samples available in the dataset, i.e., sample size $n = 29$. The two areas of the skull "Mandible" and "Nasal" are selected. We use all density levels from 130 to 249 for our analysis, hence dimension $p = 120$. We first take the difference of the bone density of the two selected areas at the corresponding density level for each subject since the two areas come from the same mouse. Then we normalize the bone density in the sense that $\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$ for all $1 \le j \le 120$.

We apply LLA projection test and several other existing methods to this dataset. Due to the relatively small sample size ($n = 29$), we opt to use slightly more data points to estimate the projection direction such that the estimator is reasonably well. As a result, we set $\kappa = 0.6$. The p-values are reported in the first row in Table 19.5. The p-values of all methods are 0, implying that the bone volume is significantly different. To see which test is more powerful, we also compute the p-values of these tests when we decrease the signals. Let $\bar{x}$ be the sample mean and $r_i = x_i - \bar{x}$ is the residual for the ith subject. Then a new observation $z_i = \delta\bar{x} + r_i$ is constructed for the ith subject. By the construction, a smaller δ results in a weaker signal and would make the test more challenging. Table 19.5 reports the p-values of all these tests for the new data z_i with $\delta = 1, 0.8, \ldots, 0.2$. As expected, the p-values of all tests increase as δ decreases. When $\delta = 0.8$ or 0.6, all these tests perform well and reject the null hypothesis at level 0.05. When $\delta = 0.4$, the Lauter's test fails to reject the null hypothesis. When $\delta = 0.2$, all the tests except for our method fail to reject the null hypothesis, which suggests that our method would perform well even though the signal is weak. Among those tests that fail to reject H_0 when $\delta = 0.2$, Li2015 projection test has the smallest p-value.

Table 19.4 Power comparison for $t_6(\mu, \Sigma_2)$ (Values in table are in percentage)

	$c = 0$				$c = 0.5$				$c = 1$			
ρ	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95	0.25	0.5	0.75	0.95
$n = 40, p = 400$												
LLA	4.74	4.44	4.81	5.19	46.94	38.23	27.23	19.02	99.95	99.52	96.50	91.21
Li2015	4.77	5.00	4.53	5.52	36.00	26.34	16.44	11.40	99.27	95.97	79.71	55.13
D1958	0.05	0.17	0.80	3.27	10.04	11.00	11.39	8.56	92.77	91.15	83.76	48.47
BS1996	0.08	0.25	1.02	4.65	12.42	13.43	14.05	11.29	94.45	92.97	86.68	55.35
CQ2010	5.49	5.71	5.83	6.77	68.47	55.52	36.29	15.44	100.0	99.97	97.97	64.15
SD2008w	0.04	0.05	0.37	1.34	7.13	8.03	8.00	3.84	91.59	89.34	79.00	33.06
SD2008wo	0.16	0.48	1.57	5.99	20.35	20.70	19.99	14.67	97.78	96.67	91.34	61.40
L1996	0.46	0.76	1.66	4.00	1.53	2.31	3.49	4.98	5.90	6.81	7.87	6.73
LJW2011	3.85	4.40	4.34	4.12	10.16	10.22	10.12	13.08	37.12	36.14	35.35	51.57
$n = 40, p = 1600$												
LLA	4.89	4.61	4.85	4.77	28.90	24.90	17.90	12.35	99.89	99.19	94.29	78.77
Li2015	5.24	4.58	5.08	5.37	13.82	11.03	8.69	5.40	83.00	69.65	44.60	17.05
D1958	0.00	0.00	0.02	1.13	0.00	0.00	0.22	2.09	7.62	7.79	9.38	9.44
BS1996	0.00	0.00	0.06	1.58	0.00	0.00	0.30	2.72	9.59	9.90	11.96	11.86
CQ2010	5.07	5.16	5.23	5.93	30.83	24.57	16.62	9.58	98.44	94.23	75.89	29.10
SD2008w	0.00	0.00	0.00	0.11	0.00	0.00	0.02	0.30	1.74	2.05	2.91	2.41
SD2008wo	0.00	0.00	0.13	2.51	0.00	0.05	0.57	4.55	18.33	18.81	19.84	17.45
L1996	0.05	0.12	0.40	2.08	0.10	0.18	0.52	2.30	0.21	0.40	0.80	2.87
LJW2011	4.26	4.24	4.22	4.18	5.60	5.31	5.49	6.01	10.32	9.83	9.97	11.53
$n = 160, p = 400$												
LLA	4.68	4.94	4.29	4.69	100.00	99.58	94.11	99.66	100.0	100.0	99.98	100.0
Li2015	4.77	4.98	4.82	4.95	99.58	96.61	78.27	88.18	100.0	100.0	99.97	100.0
D1958	0.41	1.03	2.41	4.33	99.52	99.02	95.86	53.53	99.99	99.99	99.99	99.96
BS1996	0.52	1.30	2.97	5.57	99.61	99.27	96.64	59.70	99.99	100.0	99.99	99.98
CQ2010	5.32	5.68	5.75	6.38	99.99	99.92	98.91	62.70	100.0	100.0	100.0	100.0
SD2008w	0.31	0.81	1.70	2.14	99.59	99.00	94.47	37.44	99.99	100.0	99.99	99.93
SD2008wo	0.77	1.48	3.28	6.09	99.80	99.46	96.90	61.76	100.0	100.0	100.0	99.99
L1996	0.92	1.60	2.87	4.35	6.34	8.40	9.53	5.81	24.25	25.56	22.83	8.63
LJW2011	4.55	4.42	4.36	4.38	82.42	76.72	70.40	95.73	100.0	100.0	100.0	100.0
$n = 160, p = 1600$												
LLA	5.19	5.08	5.35	4.85	99.95	98.95	89.73	79.71	100.0	100.0	100.0	100.0
Li2015	5.19	4.68	4.39	4.81	88.97	75.32	45.68	27.00	100.0	100.0	99.90	97.58
D1958	0.00	0.03	0.40	3.04	43.74	40.56	33.32	17.30	99.72	99.67	99.64	95.47
BS1996	0.00	0.07	0.47	3.80	47.26	43.91	36.50	19.82	99.78	99.74	99.75	96.40
CQ2010	4.98	4.93	5.16	6.08	98.83	95.34	75.76	27.93	100.0	100.0	100.0	98.82
SD2008w	0.00	0.00	0.16	1.22	33.20	30.69	24.45	9.06	99.59	99.57	99.37	88.00
SD2008wo	0.00	0.05	0.57	4.16	53.10	49.61	40.86	21.79	99.92	99.86	99.90	96.36
L1996	0.15	0.26	0.65	3.31	0.29	0.66	1.22	4.07	0.95	1.56	2.51	5.19
LJW2011	4.48	3.98	4.32	4.05	19.83	19.19	18.86	30.07	84.29	84.13	82.51	96.48

Table 19.5 Bone density dataset: p-value of one-sample test

δ	LLA	Li2015	D1958	BS1996	CQ2010	SD2008w	SD2008wo	L1996	LJW2011
1	0	0	0	0	0	0	0	0	0
0.8	0	0	0	0	0	0	0	0	0
0.6	0	0	0	0	0	0	0	0.0005	0
0.4	0	4×10^{-5}	0.0015	0	0	0.0088	0	0.6775	2×10^{-5}
0.2	0.0390	0.0906	0.2145	0.2710	0.2714	0.4136	0.2999	0.8870	0.3073

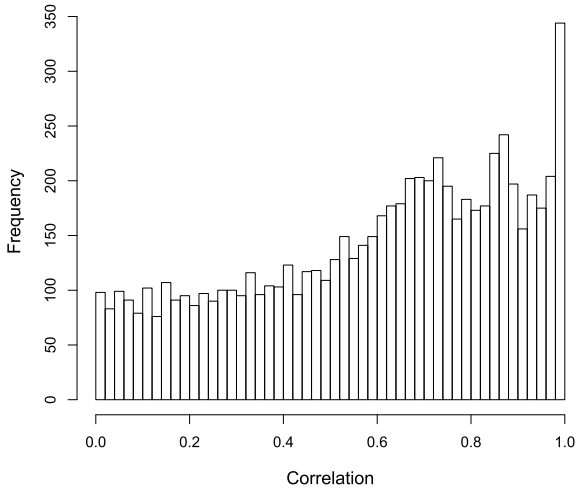


Fig. 19.1 Histogram of absolute values of paired sample correlations among bone densities at all different bone density levels

We plot the histogram of absolute values of paired sample correlations among all bone density levels in Fig. 19.1. It indicates that some bone density levels are highly correlated. This may explain why our method is more powerful than Dempster test, BS test and SD test since these methods do not take the dependence among variables into account.

Acknowledgments This work was supported by a NSF grant DMS 1820702 and a NIDA, NIH grant P50 DA039838. The content is solely the responsibility of the authors and does not necessarily represent the official views of NSF, NIH or NIDA.

References

1. Bai, Z., Saranadasa, H.: Effect of high dimension: by an example of a two sample problem. Statistica Sinica **6**, 311–329 (1996)
2. Cai, T., Liu, W.: Adaptive thresholding for sparse covariance matrix estimation. J. Am. Stat. Assoc. **106**(494), 672–684 (2011)
3. Cai, T., Liu, W., Luo, X.: A constrained ℓ_1 minimization approach to sparse precision matrix estimation. J. Am. Stat. Assoc. **106**(494), 594–607 (2011)

4. Cai, T., Liu, W., Xia, Y.: Two-sample test of high dimensional means under dependence. J. R. Stat. Soc. Series B (Statistical Methodology) **76**(2), 349–372 (2014)
5. Chen, S.X., Li, J., Zhong, P.-S.: Two-sample and ANOVA tests for high dimensional means. Ann. Stat. **47**(3), 1443–1474 (2019)
6. Chen, S.X., Qin, Y.-L.: A two-sample test for high-dimensional data with applications to gene-set testing. Ann. Stat. **38**(2), 808–835 (2010)
7. Consortium, W. T. C. C: Genome-wide association study of 14,000 cases of seven common diseases and 3,000 shared controls. Nature **447**(7145), 661 (2007)
8. Dempster, A.P.: A high dimensional two sample significance test. Ann. Math. Stat. **29**(4), 995–1010 (1958)
9. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**(456), 1348–1360 (2001)
10. Fan, J., Xue, L., Zou, H.: Strong oracle optimality of folded concave penalized estimation. Ann. Stat. **42**(3), 819 (2014)
11. Fang, K.-T., Kotz, S., Ng, K.: Symmetric Multivariate and Related Distributions. Chapman and Hall (1990)
12. Fang, K.-T., Zhang, Y.-T.: Generalized Multivariate Analysis. Science Press, Springer-Verlag, Beijing (1990)
13. Hotelling, H.: The generalization of student's ratio. Ann. Math. Stat. **2**(3), 360–378 (1931)
14. Lauter, J.: Exact t and F tests for analyzing studies with multiple endpoints. Biometrics **52**(3), 964–970 (1996)
15. Li, R., Huang, Y., Wang, L., Xu, C.: Projection Test for High-dimensional Mean Vectors with Optimal Direction (2015)
16. Lopes, M., Jacob, L., Wainwright, M.J.: A more powerful two-sample test in high dimensions using random projection. In: Advances in Neural Information Processing Systems, pp. 1206–1214 (2011)
17. Pan, G., Zhou, W.: Central limit theorem for hotellings T^2 statistic under large dimension. Ann. Appl. Probab. **21**(5), 1860–1910 (2011)
18. Percival, C.J., Huang, Y., Jabs, E.W., Li, R., Richtsmeier, J.T.: Embryonic craniofacial bone volume and bone mineral density in fgfr2+/p253r and nonmutant mice. Dev. Dyn. **243**(4), 541–551 (2014)
19. Srivastava, M.S., Du, M.: A test for the mean vector with fewer observations than the dimension. J. Multivar. Anal. **99**(3), 386–402 (2008)
20. Wang, L., Kim, Y., Li, R.: Calibrating nonconvex penalized regression in ultra-high dimension. Ann. Stat. **41**(5), 2505 (2013)
21. Xu, G., Lin, L., Wei, P., Pan, W.: An adaptive two-sample test for high-dimensional means. Biometrika **103**(3), 609–624 (2016)
22. Zou, H., Li, R.: One-step sparse estimates in nonconcave penalized likelihood models. Ann. Stat. **36**(4), 1509 (2008)

Chapter 20
Goodness-of-fit Tests for Correlated Bilateral Data from Multiple Groups

Xiaobin Liu and Chang-Xing Ma

Abstract Correlated bilateral data often arise in ophthalmological and otolaryngological studies, where responses of paired body parts of each subject are measured. A number of statistical methods have been proposed to tackle this intra-class correlation problem, and in practice it is important to choose the most suitable one which fits the observed data well. Tang et al. (Stat Methods Med Res 21(4):331–345, 2012, [16]) compared different goodness-of-fit statistics for correlated data including only two groups. In this article, we investigate the general situation for $g \geq 2$ groups. Our simulation results show that the performance of the goodness-of-fit test methods, as measured by the power and the type I error rate, is model depending. The observed performance difference is more significant in scenario with small sample size and/or highly dependent data structure. Examples from ophthalmologic studies are used to illustrate the application of these goodness-of-fit test methods.

20.1 Introduction

In clinical studies, paired data arise naturally if investigators collect information from paired body parts, say, legs, arms, eyes, etc. For example, the ophthalmologist records results from both left and right eyes in a routine eye examination. The outcome can be bilateral responses, unilateral response or no response. The conditional probability of having a response at one eye given a response at the other is usually different from the unconditional probability.

This intra-class correlation problem has attracted a lot of attentions and a number of statistical models have been proposed. Rosner [12] showed that treating each eye as an independent random variable was invalid in the presence of intra-class correlation. He proposed a constant R model that the conditional probability of having a response

X. Liu · C.-X. Ma (✉)
Department of Biostatistics, University at Buffalo, Buffalo, NY 14214, USA
e-mail: cxma@buffalo.edu

X. Liu
e-mail: xiaobinl@buffalo.edu

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_20

Table 20.1 Binary correlated data structure

Number of responses	Group					
	1	2	3	...	g	Total
0	m_{01}	m_{02}	m_{03}	...	m_{0g}	m_{0+}
1	m_{11}	m_{12}	m_{13}	...	m_{1g}	m_{1+}
2	m_{21}	m_{22}	m_{23}	...	m_{2g}	m_{2+}
Total	m_{+1}	m_{+2}	m_{+3}	...	m_{+g}	m_{++}

at one eye given a response at the other was R times the unconditional probability. The constant R is a measure of dependence between two eyes of the same person. Specifically, in an eye examination, patients are randomly assigned to g groups. Let $Z_{ijk} = 1$ if the kth eye of the jth person in the ith group had a response, $i = 1, \ldots, g$, $j = 1, \ldots, m_i, k = 1, 2$. If $\Pr(Z_{ijk} = 1) = \lambda_i$, Rosner's model assumed $\Pr(Z_{ijk} = 1 | Z_{ij(3-k)} = 1) = R\lambda_i$ for the same constant R in each of the g groups. Dallal [2] pointed out that Rosner's model would give a poor fit if binary responses were almost certain to happen with widely varying group-specific prevalence. He then offered a model that assumed the conditional probability was a fixed constant that $\Pr(Z_{ijk} = 1 | Z_{ij(3-k)} = 1) = \gamma$. Dallal's model also had its own limitation. Donner [3] proposed a constant ρ model where the common correlation coefficient in each of the g groups was a fixed constant ρ, which is $\mathrm{Corr}(Z_{ijk}, Z_{ij(3-k)}) = \rho$. In addition, independent model and saturated model for this type of data are also available. Different methods of data analysis under these models, such as testing the homogeneity of proportions and constructing confidence intervals, have been proposed [1, 4–10, 13–15, 17].

With the aforementioned medels in hand, an important factor for consideration in practice is to choose the most suitable one which fits well the observed data. Tang et al. [16] compared different goodness-of-fit test methods for correlated binary data including only two groups. However, the general situation where the correlated bilateral data have $g \geq 2$ groups has not been investigated, which is our focus in this article. More specifically, we choose likelihood ratio statistic (G^2), Pearson Chi-square statistic (X^2), adjusted Pearson Chi-square statistic (X^2_{adj}) and three bootstrap processes as candidate methods, and examine their performance in testing goodness-of-fit of each model for correlated binary data with multiple groups.

The rest part of the article is organized as follows. In Sect. 20.2, we present details of the five models for correlated binary data with multiple groups, including the formulas for computing log-likelihood of a specific dataset and maximum likelihood estimates (MLE) of model parameters. Three statistics and three bootstrap processes for testing goodness-of-fit of different models and their asymptotic distributions are presented in Sect. 20.3. Simulation studies are conducted in Sect. 20.4 with respect to type I error and power. Two examples are used to illustrate the different goodness-of-fit test methods in Sect. 20.5. Section 20.6 contains a brief summary and discussion.

20.2 Models for Correlated Bilateral Data

Suppose there are m_{li} subjects having l responses in the ith group, $l = 0, 1, 2$, $i = 1, \ldots, g$. The data structure is shown in Table 20.1. Hereafter we record a specific table as a $1 \times 3g$ vector

$$\mathbf{m} = (m_{01}, m_{11}, m_{21}, \ldots, m_{0g}, m_{1g}, m_{2g}).$$

20.2.1 Independence Model

This model assumes $R = 1$ or $\rho = 0$, such that the paired body parts are independent with each other. The parameter space for this model is $\Omega_I = \{(\lambda_1, \ldots, \lambda_g) : 0 \leq \lambda_i \leq 1, i = 1, \ldots, g\}$ and the log-likelihood function for a specific dataset is

$$l_0(\lambda_1, \ldots, \lambda_g; \mathbf{m}) = \sum_{i=1}^{g} [(2m_{0i} + m_{1i}) \log(1 - \lambda_i) + (m_{1i} + 2m_{2i}) \log \lambda_i].$$

It can be calculated that the MLEs of λ_i's are

$$\widehat{\lambda}_i = \frac{m_{1i} + 2m_{2i}}{2m_{+i}}.$$

The independent model is easy to use but often criticized for its limitation in practice.

20.2.2 Rosner's Model

The Rosner's model uses a constant R to describe the correlation between the paired body parts. The probability of getting a response at one part conditioning on having a response at the other is R times the unconditional probability. The parameter space of Rosner's model is $\Omega_R = \{(\lambda_1, \ldots, \lambda_g) : 0 < R \leq 1/a, \texttt{if } a \leq 1/2; (2 - 1/a)/a \leq R \leq 1/a, \texttt{if } a > 1/2, \texttt{where } a = \max\{\lambda_i, i = 1, \ldots, g\}\}$. When $R = 1$, this model reduces to independent model. It can be shown that the log-likelihood of a specific dataset is

$$\begin{aligned} l_1(\lambda_1, \ldots, \lambda_g, R; \mathbf{m}) = \sum_{i=1}^{g} [&m_{0i} \log(R\lambda_i^2 - 2\lambda_i + 1) \\ &+ m_{1i} \log(2\lambda_i(1 - R\lambda_i)) + m_{2i} \log(R\lambda_i^2)]. \end{aligned}$$

To get MLEs of the parameters, we differentiate the above function with respect to λ_i's and R and set them equal to 0.

$$\frac{\partial l}{\partial R} = \frac{m_{0+}\lambda_i^2}{R\lambda_i^2 - 2\lambda_i + 1} - \frac{m_{1+}\lambda_i}{1 - R\lambda_i} + \frac{m_{2+}}{R} = 0$$
$$\frac{\partial l}{\partial \lambda_i} = \frac{2m_{0i}(R\lambda_i - 1)}{R\lambda_i^2 - 2\lambda_i + 1} - \frac{m_{1i}R}{1 - R\lambda_i} + \frac{2m_{2i} + m_{1i}}{\lambda_i} = 0, i = 1, \ldots, g$$

As there is no close-form solution to these equations, we use the Fisher-Score iterative method to get the MLEs by repeating the following steps derived from [8].

The equation sets (1) can be simplified as a 3rd order polynomial

$$\lambda_i^3 - \frac{4m_{0i} + 5m_{1i} + 6m_{2i}}{2Rm_{+i}}\lambda_i^2 + \frac{m_{0i} + (1+R)m_{1i} + (2+R)m_{2i}}{R^2 m_{+i}}\lambda_i - \frac{m_{1i} + 2m_{2i}}{2R^2 m_{+i}} = 0.$$

The $(t+1)$th update for λ_i is the real root of the 3rd order polynomial after replacing R with $R^{(t+1)}$; and the $(t+1)$th update for R can be calculated with the following formula

$$R^{(t+1)} = R^{(t)} - \left(\frac{\partial^2 l_1}{\partial R^2}\left(\lambda_1^{(t)}, \ldots, \lambda_g^{(t)}; R^{(t)}\right)\right)^{-1} \frac{\partial l_1}{\partial R}\left(\lambda_1^{(t)}, \ldots, \lambda_g^{(t)}; R^{(t)}\right),$$

where

$$\frac{\partial l_1}{\partial R} = \sum_{i=1}^{g}\left(\frac{m_{2i}}{R} + \frac{\lambda_i^2 m_{0i}}{R\lambda_i^2 - 2\lambda_i + 1} + \frac{\lambda_i m_{1i}}{R\lambda_i - 1}\right),$$
$$\frac{\partial^2 l_1}{\partial R^2} = -\frac{m_{2+}}{R^2} - \sum_{i=1}^{g}\frac{\lambda_i^2 m_{1i}}{(R\lambda_i - 1)^2} - \sum_{i=1}^{g}\frac{\lambda_i^4 m_{0i}}{(R\lambda_i^2 - 2\lambda_i + 1)^2}.$$

20.2.3 Equal Correlation Coefficients Model

This model arises when assuming the correlation coefficients of the paired body parts in all g groups are equal. That is, $\texttt{Corr}(Z_{ijk}, Z_{ij(3-k)}) = \rho$ for all j. The parameter space is $\Omega_E = \{(\lambda_1, \ldots, \lambda_g, \rho) : 0 \le \lambda_i \le 1, 0 \le \rho \le 1, i = 1, \ldots, g\}$. The log-likelihood function for a specific dataset is

$$l_2(\lambda_1, \ldots, \lambda_g, \rho; \mathbf{m}) = \sum_{i=1}^{g}[m_{0i}\log((1-\lambda_i)(\rho\lambda_i - \lambda_i + 1)) + m_{1i}\log(2\lambda_i(1-\rho)(1-\lambda_i)) + m_{2i}\log(\lambda_i^2 + \rho\lambda_i(1-\lambda_i))].$$

Now we calculate MLEs of the parameters. By partial differentiating l_2 with respect to $\lambda_i's$ and ρ we have

$$\frac{\partial l_2}{\partial \rho} = \sum_{i=1}^{g}\left[\frac{m_{0i}\lambda_i}{\rho\lambda_i - \lambda_i + 1} - \frac{m_{1i}}{1-\rho} + \frac{m_{2i}(1-\lambda_i)}{\lambda_i + \rho(1-\lambda_i)}\right]$$

$$\frac{\partial l_2}{\partial \lambda_i} = \frac{m_{0i}(2(1-\rho)\lambda_i + \rho - 2)}{(1-\lambda_i)(\rho\lambda_i - \lambda_i + 1)} + \frac{m_{1i}(1-2\lambda_i)}{\lambda_i(1-\lambda_i)} + \frac{m_{2i}(2\lambda_i + \rho - 2\rho\lambda_i)}{\lambda_i^2 + \rho\lambda_i(1-\lambda_i)},$$
$$i = 1, \ldots, g$$

Setting the above partial differentiations equal to zero, the MLEs of $\lambda_i's$ and ρ are the solution to these equations. Similar to Rosner's model, no explicit solution exists. With Fisher-Score iteration, MLEs can be calculated by repeating the following steps derived by [7] until convergence.
The equation sets (2) can be simplified as a 3rd order polynomial

$$2\rho(2-\rho)m_{+i}\lambda_i^3 + [3\rho^2 m_{+i} - \rho(5m_{0i} + 6m_{1i} + 7m_{2i}) + 2m_{0i} + 3m_{1i} + 4m_{2i}]\lambda_i^2$$
$$+ [(4\rho - \rho^2)m_{+i} - 2\rho m_{0i} - m_{1i} - 2m_{2i}]\lambda_i - \rho(m_{1i} + m_{2i}) = 0.$$

The $(t+1)$th update for λ_i can be obtained by directly solving the polynomial and choosing the real root after replacing ρ with $\rho^{(t+1)}$; ρ can be updated by the following iteration

$$\rho^{(t+1)} = \rho^{(t)} - \left(\frac{\partial^2 l_2}{\partial \rho^2}\left(\lambda_1^{(t)}, \ldots, \lambda_g^{(t)}, \rho^{(t)}\right)\right)^{-1} \frac{\partial l}{\partial \rho}\left(\lambda_1^{(t)}, \ldots, \lambda_g^{(t)}, \rho^{(t)}\right),$$

where

$$\frac{\partial l_2}{\partial \rho} = \sum_{i=1}^{g}\left[\frac{m_{1i}}{\rho - 1} - \frac{(\lambda_i - 1)m_{2i}}{\rho + \lambda_i - \rho\lambda_i} + \frac{\lambda_i m_{0i}}{\rho\lambda_i - \lambda_i + 1}\right],$$

$$\frac{\partial^2 l_2}{\partial \rho^2} = -\sum_{i=1}^{g}\left[\frac{m_{1i}}{(\rho-1)^2} + \frac{\lambda_i^2 m_{0i}}{(\rho\lambda_i - \lambda_i + 1)^2} + \frac{(\lambda_i - 1)^2 m_{2i}}{(\rho + \lambda_i - \rho\lambda_i)^2}\right].$$

20.2.4 Dallal's Model

This model assumes that the conditional probability is a constant, which is, $Pr(Z_{ijk} = 1 | Z_{ij(3-k)} = 1) = \gamma$. The parameter space is $\Omega_D = \{(\lambda_1, \ldots, \lambda_g, \gamma) : 0 \le \gamma \le 1$ if $a \le 1/2; 2 - 1/a \le \gamma \le 1$ if $a > 1/2$, where $a = \max\{\lambda_i, i = 1, \ldots, g\}\}$. The log-likelihood function for a specific dataset is

$$l_3(\lambda_1, \ldots, \lambda_g, \gamma; \mathbf{m}) = \sum_{i=1}^{g} \big[m_{0i} \log((\gamma - 2)\lambda_i + 1) + m_{1i} \log(2(1-\gamma)\lambda_i) + m_{2i} \log(\gamma \lambda_i)\big].$$

The MLEs of parameters can be calculated by partial differentiating l_3 with respect to $\lambda_i' s$ and γ and setting them equal to zero. Here

$$\frac{\partial l_3}{\partial \lambda_i} = \frac{m_{0i}(\gamma - 2)}{(\gamma - 2)\lambda_i + 1} + \frac{m_{1i} + m_{2i}}{\lambda_i} = 0, i = 1, \ldots, g,$$
$$\frac{\partial l}{\partial \gamma} = \sum_{i=1}^{g} [\frac{m_{0i}\lambda_i}{(\gamma - 2)\lambda_i + 1} - \frac{m_{1i}}{1-\gamma} + \frac{m_{2i}}{\gamma}] = 0.$$

Solving these equations, we have

$$\widehat{\gamma} = \frac{2m_{2+}}{2m_{2+} + m_{1+}},$$
$$\widehat{\lambda_i} = \frac{(m_{1i} + m_{2i})(2m_{2+} + m_{1+})}{2m_{+i}(m_{2+} + m_{1+})}, i = 1, \ldots, g.$$

20.2.5 Saturated Model

This is the model with full parameters, and is often used as a reference model in the goodness-of-fit test. The parameter space is $\Omega_S = \{(p_{01}, p_{11}, \ldots, p_{0g}, p_{1g}) : 0 \leq p_{li} \leq 1, l = 0, 1, i = 1, \ldots, g\}$. For a given configuration, the log-likelihood is

$$l(p_{01}, p_{11}, \ldots, p_{0g}, p_{1g}; \mathbf{m}) = \sum_{l=0}^{2} \sum_{i=1}^{g} m_{li} \log(p_{li}),$$

where $p_{2i} = 1 - p_{0i} - p_{1i}$ for $i = 1, \ldots, g$. MLEs of the parameters are directly calculated by $\widehat{p}_{li} = \frac{m_{li}}{m_{+i}}$.

20.3 Methods for Goodness-of-Fit Test

Several statistics have been developed for goodness-of-fit test. Here we choose three most frequently used ones: likelihood ratio (G^2), Pearson chi-square (X^2) and the adjusted chi-square ($X^2_{\texttt{adj}}$) statistics. Their formulas are given by

$$G^2 = 2\sum(\texttt{observed})\log\frac{\texttt{observed}}{\texttt{expected}} = 2\sum_{l=0}^{2}\sum_{i=1}^{g} m_{li}\log\frac{m_{li}}{\widehat{m}_{li}},$$

$$X^2 = \sum\frac{(\texttt{observed}-\texttt{expected})^2}{\texttt{expected}} = \sum_{l=0}^{2}\sum_{i=1}^{g}\frac{(m_{li}-\widehat{m}_{li})^2}{\widehat{m}_{li}},$$

$$X^2_{\texttt{adj}} = \sum\frac{(|\texttt{observed}-\texttt{expected}|-1/2)^2}{\texttt{expected}} = \sum_{l=0}^{2}\sum_{i=1}^{g}\frac{(|m_{li}-\widehat{m}_{li}|-1/2)^2}{\widehat{m}_{li}}.$$

When the sample size is large enough, all of the three statistics asymptotically follow chi-square distribution with the degree of freedom being the difference of the number of parameters between the two models in the null and alternative hypotheses. In our cases, the saturated model has $2g$ parameters; Independent model has g parameters; Rosner's model, equal correlation coefficients model and Dallal's model all have $g+1$ parameters. So for the last three models of interest, the asymptotic chi-square distribution has $g-1$ degree of freedom. For the independent model, the degree of freedom is g.

In addition, we introduce three bootstrap methods here. They differ from each other in what statistic (value) is used for ordering the samples. The bootstrap procedure can be summarized as

1. Randomly generate N samples based on the estimated parameters under the null model;
2. Calculate the probability (or G^2 or X^2) of each sample under the null model. The probability of a given table is calculated as

$$Pr(\mathbf{m}|H_0) = \prod_{i=1}^{g}\frac{m_{+i}!}{m_{0i}!m_{1i}!m_{2i}!}\tilde{p}_{0i}^{m_{0i}}\tilde{p}_{1i}^{m_{1i}}\tilde{p}_{2i}^{m_{2i}},$$

   where $\tilde{p}_{li}$ is the corresponding MLE of p_{li} under the null model;
3. Count the number of probabilities (or G^2 or X^2) which are less (larger) than the probability (or G^2 or X^2) of the current observed data. If the number is $<5\%N$, then reject the null hypothesis.

We denote the three bootstrap procedures based on G^2, X^2 and probability as B_1, B_2 and B_3, respectively. In the next simulation section, we compare the performance of the three test statistics and three bootstrap methods with respect to empirical type I error and power.

20.4 Simulation Study

In this section, we examine the performance of the six proposed methods for testing goodness-of-fit with Rosner's, equal correlation coefficients and Dallal's model. The independent model is nested under these three model by setting $R = 1$ or $\rho = 0$, and the saturated model is used as a reference model.

We first investigate the empirical type I error. We considered scenarios when $g = 2, 4, 8$ and $m_{+1} = \cdots = m_{+g} = m = 25, 50, 100$, and chose different sets of parameters under each m, g combination. Under each of these configurations, we generated 10,000 random samples based on the null model. For bootstrap methods, 2000 bootstrap samples were generated for each of the 10,000 samples. Our empirical type I error was computed by `number of rejection`/10,000.

Tables 20.2, 20.3, and 20.4 report the empirical type I error under all scenarios based on Rosner's, equal correlation coefficients and Dallal's model, respectively. For Rosner's model, we let $R = 1.2$, 1.5, and 1.8; for equal correlation coefficients model, let $\rho = 0.5$, 0.7, and 0.9; for Dallal's model, let $\gamma = 0.3$, 0.5, and 0.7. In these three tables, different λ setups are used to be more careful about its influence on the results. These setups are

Case I: $\lambda = (0.3, 0.5)$;
Case II: $\lambda = (0.5, 0.5)$;
Case III: $\lambda = (0.1, 0.2, 0.3, 0.4)$;
Case IV: $\lambda = (0.2, 0.2, 0.4, 0.4)$;
Case V: $\lambda = (0.1, 0.2, 0.3, 0.4, 0.1, 0.2, 0.3, 0.4)$;
Case VI: $\lambda = (0.2, 0.2, 0.4, 0.4, 0.2, 0.2, 0.4, 0.4)$.

From the three tables, the results of $g = 2$ show similar pattern as those in [16]. Combining them with results of $g = 4$ and $g = 8$, it shows that for all the three models, the adjusted Pearson χ^2 is over conservative with very small empirical type I error. Therefore, its usage is not recommended here. G^2 works very well for all the three models. X^2 produces good empirical type I error for Rosner's model. For the other two models, its empirical type I error is greater than 0.05 when the sample size is relatively small, and it gets better as the sample size increases. For the three bootstrap methods, their empirical type I errors are nearly as good as G^2. However, they work poorer when R is large in Rosner's model and ρ is large in equal correlation model with small sample size. Generally B_1 and B_2 have smaller empirical type I error than B_3. Their performance also improve as the sample size gets larger.

Then we considered power under different parameter configurations. The results are shown in Table 20.5. The parameter setups are

Case I : $\lambda = (0.2, 0.2)$, $R = (1.2, 1.5)$ (or $\rho = (0.5, 0.7)$ or $\gamma = (0.5, 0.7)$);
Case II: $\lambda = (0.2, 0.4)$, $R = (1.2, 1.5)$ (or $\rho = (0.5, 0.7)$ or $\gamma = (0.5, 0.7)$);
Case III: $\lambda = (0.1, 0.2, 0.3, 0.4)$, $R = (1.2, 1.2, 1.5, 1.5)$ (or $\rho = (0.5, 0.5, 0.7, 0.7)$ or $\gamma = (0.5, 0.5, 0.7, 0.7)$);
Case IV: $\lambda = (0.2, 0.2, 0.4, 0.4)$, $R = (1.2, 1.2, 1.5, 1.5)$ (or $\rho = (0.5, 0.5, 0.7, 0.7)$ or $\gamma = (0.5, 0.5, 0.7, 0.7)$);
Case V: $\lambda = (0.1, 0.2, 0.3, 0.4, 0.1, 0.2, 0.3, 0.4)$, $R = (1.2, 1.2, 1.2, 1.2, 1.5, 1.5, 1.5, 1.5)$ (or $\rho = (0.5, 0.5, 0.5, 0.5, 0.7, 0.7, 0.7, 0.7)$ or $\gamma = (0.5, 0.5, 0.5, 0.5, 0.7, 0.7, 0.7, 0.7)$);
Case VI: $\lambda = (0.2, 0.2, 0.4, 0.4, 0.2, 0.2, 0.4, 0.4)$, $R = (1.2, 1.2, 1.2, 1.2, 1.5, 1.5, 1.5, 1.5)$ (or $\rho = (0.5, 0.5, 0.5, 0.5, 0.7, 0.7, 0.7, 0.7)$ or $\gamma = (0.5, 0.5, 0.5, 0.5, 0.7, 0.7, 0.7, 0.7)$).

We found that for all the three models, X^2 has the highest power among the six methods. Power of B_3 is also high especially when g is large. However, the difference among the six methods is not significant.

20.5 Real World Examples

We present two real examples to examine the performance of the six methods. The first one was 218 outpatients aged from 20 to 29 with retinitis pigmentosa (RP). They were seen at Massachusetts Eye and Ear Infirmary from 1970 to 1979 [12]. These outpatients were assigned into four groups based on genetic types of autosomal dominant RP (DOM), recessive RP (AR), sex-linked RP (SL) and isolate RP (ISO). Snellen visual acuity (VA) was recorded for each person. If VA was 20/50 or worse, the eye was considered affected while an eye was normal if VA was 20/40 or better. The number of effected eyes for persons in each genetic type group is shown in Table 20.6. We use all the four models to fit the data. P-values for the six methods are calculated as well as AICs. The results are shown in Table 20.7.

From Table 20.7, we can find that p-values of all six methods for Rosner's model, equal correlation coefficients model and Dallal's model are larger than 0.05. However, p-values for equal correlation coefficients model are much larger than the other models. This indicates that equal correlation coefficients model fits our data the best. Also, equal correlation coefficients model has the smallest AIC, which also makes it the best model to fit our data.

The second example was from a study of extend and causes of blindness and visual impairment (VI) in the Varamin district of Iran [11]. Visual acuity (VA) of 2819 persons were examined. The prevalence of VI, defined as VA $\geq$ 6/60 and $<$ 6/18 in the better eye with available correction, were shown in Table 20.8 among different age groups. We fit all the four models and calculate the p-values of the six methods as those in the first example. As shown in Table 20.9, equal correlation coefficient model has the smallest AIC, and all p-values are larger than 0.05. Therefore, similar to the first example, equal correlation coefficients model fits our data the best.

Table 20.2 The empirical type I error for Rosner's model

m	g	Case	R	X^2	G^2	X^2_{adj}	B_1	B_2	B_3
25	2	I	1.2	0.065	0.051	0.015	0.062	0.060	0.063
			1.5	0.061	0.054	0.019	0.058	0.059	0.068
			1.8	0.052	0.049	0.021	0.065	0.067	0.070
		II	1.2	0.053	0.053	0.017	0.049	0.048	0.058
			1.5	0.046	0.042	0.014	0.050	0.048	0.057
			1.8	0.050	0.047	0.013	0.0845	0.0845	0.0805
	4	III	1.2	0.047	0.051	0.049	0.051	0.047	0.050
			1.5	0.049	0.046	0.041	0.043	0.039	0.043
			1.8	0.046	0.042	0.041	0.036	0.038	0.041
		IV	1.2	0.057	0.044	0.008	0.042	0.037	0.046
			1.5	0.055	0.040	0.008	0.056	0.046	0.062
			1.8	0.059	0.043	0.008	0.053	0.045	0.059
	8	V	1.2	0.048	0.050	0.064	0.039	0.031	0.040
			1.5	0.047	0.043	0.062	0.039	0.037	0.043
			1.8	0.053	0.047	0.061	0.030	0.027	0.045
		VI	1.2	0.068	0.048	0.003	0.055	0.051	0.069
			1.5	0.066	0.047	0.005	0.048	0.046	0.059
			1.8	0.058	0.042	0.006	0.042	0.038	0.058
50	2	I	1.2	0.055	0.051	0.025	0.052	0.053	0.058
			1.5	0.053	0.052	0.025	0.044	0.045	0.048
			1.8	0.053	0.053	0.026	0.088	0.089	0.091
		II	1.2	0.047	0.046	0.024	0.044	0.045	0.048
			1.5	0.046	0.045	0.021	0.045	0.045	0.048
			1.8	0.051	0.050	0.023	0.062	0.062	0.063
	4	III	1.2	0.044	0.043	0.010	0.045	0.041	0.039
			1.5	0.054	0.047	0.012	0.047	0.049	0.051
			1.8	0.054	0.045	0.013	0.050	0.045	0.048
		IV	1.2	0.064	0.044	0.011	0.047	0.043	0.044
			1.5	0.065	0.049	0.014	0.061	0.061	0.062
			1.8	0.053	0.045	0.014	0.048	0.049	0.059
	8	V	1.2	0.055	0.048	0.008	0.054	0.048	0.058
			1.5	0.058	0.048	0.008	0.047	0.044	0.052
			1.8	0.053	0.046	0.007	0.053	0.050	0.060
		VI	1.2	0.067	0.047	0.007	0.055	0.051	0.060
			1.5	0.067	0.049	0.009	0.059	0.055	0.066
			1.8	0.060	0.047	0.009	0.053	0.050	0.059

(continued)

Table 20.2 (continued)

m	g	Case	R	X^2	G^2	X^2_{adj}	B_1	B_2	B_3
100	2	I	1.2	0.053	0.051	0.029	0.051	0.055	0.055
			1.5	0.048	0.048	0.029	0.049	0.048	0.051
			1.8	0.051	0.051	0.033	0.056	0.055	0.057
		II	1.2	0.053	0.052	0.033	0.054	0.055	0.056
			1.5	0.051	0.050	0.033	0.043	0.043	0.043
			1.8	0.048	0.048	0.028	0.044	0.044	0.047
	4	III	1.2	0.054	0.049	0.017	0.052	0.051	0.056
			1.5	0.056	0.046	0.016	0.049	0.048	0.051
			1.8	0.056	0.046	0.018	0.053	0.050	0.056
		IV	1.2	0.060	0.051	0.020	0.051	0.051	0.052
			1.5	0.058	0.052	0.023	0.049	0.051	0.052
			1.8	0.052	0.049	0.021	0.041	0.043	0.048
	8	V	1.2	0.059	0.045	0.010	0.052	0.048	0.052
			1.5	0.060	0.049	0.012	0.053	0.052	0.056
			1.8	0.055	0.044	0.011	0.050	0.052	0.058
		VI	1.2	0.064	0.052	0.016	0.052	0.053	0.056
			1.5	0.055	0.048	0.013	0.051	0.050	0.051
			1.8	0.053	0.049	0.017	0.043	0.043	0.051

Table 20.3 The empirical type I error for equal correlation coefficients model

m	g	Case	ρ	X^2	G^2	X^2_{adj}	B_1	B_2	B_3
25	2	I	0.5	0.060	0.054	0.016	0.053	0.056	0.063
			0.7	0.065	0.052	0.012	0.064	0.067	0.069
			0.9	0.067	0.022	0.001	0.053	0.054	0.041
		II	0.5	0.064	0.059	0.019	0.058	0.059	0.064
			0.7	0.065	0.054	0.012	0.057	0.061	0.064
			0.9	0.079	0.021	0.001	0.049	0.044	0.037
	4	III	0.5	0.072	0.044	0.005	0.055	0.051	0.059
			0.7	0.083	0.043	0.003	0.063	0.055	0.057
			0.9	0.043	0.043	0.002	0.069	0.069	0.039
		IV	0.5	0.068	0.047	0.009	0.055	0.055	0.065
			0.7	0.079	0.045	0.007	0.055	0.057	0.059
			0.9	0.045	0.039	0.002	0.064	0.064	0.035
	8	V	0.5	0.079	0.039	0.001	0.033	0.037	0.037
			0.7	0.091	0.038	0.002	0.052	0.040	0.052
			0.9	0.046	0.046	0.004	0.053	0.047	0.021
		VI	0.5	0.079	0.045	0.003	0.039	0.043	0.055
			0.7	0.087	0.043	0.003	0.041	0.039	0.049
			0.9	0.053	0.046	0.003	0.047	0.045	0.019

(continued)

Table 20.3 (continued)

m	g	Case	ρ	X^2	G^2	X^2_{adj}	B_1	B_2	B_3
50	2	I	0.5	0.051	0.048	0.024	0.046	0.046	0.048
			0.7	0.055	0.050	0.020	0.055	0.055	0.063
			0.9	0.079	0.042	0.008	0.077	0.076	0.061
		II	0.5	0.056	0.055	0.028	0.056	0.057	0.057
			0.7	0.059	0.055	0.025	0.050	0.050	0.053
			0.9	0.088	0.051	0.008	0.067	0.066	0.064
	4	III	0.5	0.065	0.050	0.014	0.058	0.059	0.063
			0.7	0.068	0.047	0.008	0.050	0.048	0.053
			0.9	0.069	0.043	0.006	0.057	0.049	0.040
		IV	0.5	0.059	0.052	0.016	0.051	0.054	0.057
			0.7	0.061	0.049	0.014	0.048	0.051	0.054
			0.9	0.070	0.042	0.004	0.053	0.047	0.039
	8	V	0.5	0.061	0.044	0.006	0.041	0.040	0.047
			0.7	0.067	0.045	0.006	0.050	0.055	0.059
			0.9	0.075	0.048	0.003	0.054	0.054	0.038
		VI	0.5	0.064	0.053	0.010	0.064	0.061	0.069
			0.7	0.064	0.047	0.007	0.039	0.045	0.047
			0.9	0.082	0.048	0.005	0.059	0.049	0.044
			0.7	0.054	0.053	0.030	0.046	0.047	0.049
			0.9	0.058	0.048	0.016	0.047	0.052	0.052
		II	0.5	0.049	0.048	0.028	0.053	0.053	0.055
			0.7	0.053	0.052	0.029	0.049	0.049	0.050
			0.9	0.065	0.046	0.019	0.048	0.054	0.057
	4	III	0.5	0.055	0.050	0.020	0.055	0.054	0.056
			0.7	0.056	0.048	0.017	0.053	0.052	0.054
			0.9	0.069	0.043	0.009	0.052	0.055	0.053
		IV	0.5	0.055	0.050	0.021	0.049	0.047	0.050
			0.7	0.056	0.052	0.022	0.057	0.056	0.058
			0.9	0.067	0.039	0.010	0.051	0.042	0.045
	8	V	0.5	0.058	0.049	0.014	0.044	0.048	0.050
			0.7	0.056	0.047	0.011	0.043	0.044	0.045
			0.9	0.077	0.048	0.007	0.050	0.049	0.048
		VI	0.5	0.053	0.048	0.016	0.051	0.050	0.056
			0.7	0.057	0.049	0.015	0.049	0.051	0.054
			0.9	0.073	0.046	0.008	0.049	0.050	0.051

Table 20.4 The empirical type I error for Dallal's model

m	g	Case	γ	X^2	G^2	X^2_{adj}	B_1	B_2	B_3
25	2	I	0.3	0.076	0.050	0.029	0.069	0.067	0.068
			0.5	0.061	0.052	0.021	0.041	0.043	0.050
			0.7	0.065	0.060	0.019	0.047	0.049	0.056
		II	0.3	0.064	0.046	0.045	0.055	0.057	0.056
			0.5	0.056	0.053	0.026	0.050	0.051	0.057
			0.7	0.058	0.055	0.023	0.053	0.054	0.059
	4	III	0.3	0.068	0.050	0.006	0.062	0.051	0.051
			0.5	0.076	0.045	0.006	0.056	0.052	0.057
			0.7	0.079	0.045	0.004	0.060	0.063	0.065
		IV	0.3	0.073	0.045	0.007	0.053	0.047	0.051
			0.5	0.077	0.051	0.009	0.052	0.056	0.059
			0.7	0.078	0.050	0.008	0.053	0.058	0.066
	8	V	0.3	0.073	0.049	0.003	0.055	0.053	0.051
			0.5	0.086	0.044	0.003	0.065	0.062	0.067
			0.7	0.085	0.038	0.001	0.056	0.063	0.068
		VI	0.3	0.082	0.048	0.005	0.052	0.051	0.055
			0.5	0.079	0.045	0.002	0.053	0.048	0.062
			0.7	0.085	0.049	0.004	0.059	0.059	0.075
50	2	I	0.3	0.056	0.049	0.022	0.046	0.049	0.052
			0.5	0.054	0.052	0.025	0.051	0.052	0.054
			0.7	0.053	0.051	0.026	0.052	0.052	0.058
		II	0.3	0.052	0.047	0.025	0.046	0.047	0.051
			0.5	0.051	0.049	0.027	0.049	0.049	0.051
			0.7	0.049	0.048	0.025	0.044	0.044	0.045
	4	III	0.3	0.069	0.050	0.011	0.059	0.055	0.063
			0.5	0.063	0.047	0.013	0.050	0.047	0.050
			0.7	0.068	0.049	0.012	0.055	0.052	0.056
		IV	0.3	0.070	0.045	0.012	0.053	0.055	0.055
			0.5	0.059	0.050	0.015	0.048	0.048	0.052
			0.7	0.061	0.052	0.018	0.049	0.049	0.053
	8	V	0.3	0.073	0.050	0.008	0.056	0.058	0.055
			0.5	0.065	0.042	0.006	0.049	0.048	0.056
			0.7	0.068	0.045	0.005	0.053	0.052	0.061
		VI	0.3	0.073	0.050	0.007	0.049	0.047	0.051
			0.5	0.058	0.044	0.007	0.054	0.053	0.059
			0.7	0.063	0.048	0.009	0.055	0.055	0.064

(continued)

Table 20.4 (continued)

m	g	Case	γ	X^2	G^2	X^2_{adj}	B_1	B_2	B_3
100	2	I	0.3	0.047	0.046	0.025	0.048	0.049	0.051
			0.5	0.051	0.050	0.032	0.048	0.051	0.051
			0.7	0.053	0.051	0.032	0.048	0.048	0.049
		II	0.3	0.051	0.049	0.032	0.046	0.047	0.048
			0.5	0.050	0.049	0.032	0.042	0.043	0.043
			0.7	0.051	0.051	0.033	0.048	0.049	0.050
	4	III	0.3	0.065	0.050	0.019	0.052	0.045	0.048
			0.5	0.055	0.048	0.018	0.051	0.051	0.052
			0.7	0.060	0.052	0.020	0.056	0.058	0.062
		IV	0.3	0.055	0.049	0.020	0.051	0.055	0.050
			0.5	0.051	0.047	0.022	0.042	0.040	0.042
			0.7	0.056	0.052	0.025	0.051	0.051	0.053
	8	V	0.3	0.063	0.049	0.013	0.042	0.043	0.045
			0.5	0.061	0.049	0.015	0.055	0.055	0.059
			0.7	0.061	0.051	0.014	0.065	0.065	0.071
		VI	0.3	0.054	0.044	0.013	0.041	0.042	0.045
			0.5	0.051	0.046	0.015	0.049	0.051	0.055
			0.7	0.053	0.047	0.015	0.051	0.051	0.055

Table 20.5 Estimated power (in %) when group sizes are $m = 150$

g	Case	X^2	G^2	X^2_{adj}	B_1	B_2	B_3
Rosner's model							
2	I	9.7	9.4	5.4	8.9	8.9	9.0
	II	13.4	12.4	8.4	11.8	11.3	10.6
4	III	11.3	7.7	3.5	10.8	8.7	8.6
	IV	14.3	12.7	7.7	12.9	11.9	11.4
8	V	30.1	27.1	14.2	28.1	28.2	28.9
	VI	40.1	39.3	26.2	39.0	39.8	40.5
Equal correlation coefficients model							
2	I	40.6	40.3	32.0	39.5	39.6	40.0
	II	48.4	48.2	40.1	49.4	50.0	49.6
4	III	53.3	52.5	40.2	51.4	52.0	52.1
	IV	61.3	60.7	49.6	60.8	61.1	60.9
8	V	75.9	74.8	59.8	76.0	76.1	76.8
	VI	83.6	82.7	70.9	82.8	82.8	83.3
Dallal's model							
2	I	47.7	47.3	38.3	46.0	46.3	46.7
	II	59.6	59.2	51.6	59.6	59.2	59.8
4	III	64.8	63.4	52.1	63.7	63.6	64.0
	IV	75.7	74.9	65.8	74.6	74.6	74.7
8	V	89.5	89.3	80.4	88.5	88.8	89.1
	VI	94.7	94.5	89.4	93.9	94.3	94.5

Table 20.6 Number of effected eyes for persons in each genetic type group

Number of affected eyes	Genetic type			
	DOM	AR	SL	ISO
0	15	7	3	67
1	6	5	2	24
2	7	9	14	57

Table 20.7 p-values of different methods and AIC for the four models

Model	p-values						
	G^2	X^2	X^2_{adj}	B_1	B_2	B_3	AIC
Independent	0	0	0	0.2845	0.0015	0.2645	413.8090
Rosner's	0.0595	0.0798	0.2033	0.0764	0.0911	0.0575	80.2841
Equal correlation	0.7355	0.7205	0.9030	0.7509	0.7314	0.5978	67.9795
Dallal's	0.2162	0.2424	0.4418	0.2170	0.2310	0.1935	74.3464

Table 20.8 Prevalence of VI by age groups in the sample population

Number of affected eyes	Age groups (yrs)						
	50–54	54–59	60–64	65–69	70–74	75–79	80+
0	885	478	372	198	135	63	28
1	39	30	48	33	44	21	23
2	21	23	31	17	40	30	32

Table 20.9 p-values of different methods and AIC for the four models in VI study

Model	p-values						
	G^2	X^2	X^2_{adj}	B_1	B_2	B_3	AIC
Independent	0	0	0	0	0	0	9060
Rosner's	0	0	0	0	0	0	458.1489
Equal correlation	0.7525	0.7557	0.8797	0.7745	0.7750	0.7525	161.8042
Dallal's	0.0287	0.0294	0.0662	0.0350	0.0340	0.0350	183.1064

20.6 Conclusions

In the analysis of correlated bilateral data, an important factor for consideration is to choose the most suitable statistical model which fits well the observed data. This topic has not been systematically investigated for correlated binary data with $g \geq 2$ groups. In this article, we describe five popular models for correlated paired data with multiple groups, and provide the formulas for computing log-likelihood of a specific dataset as well as MLE for model parameters. We then investigate different methods for goodness-of-fit test of these models. Simulation study is performed to calculate empirical type I error rates and power. G^2 statistic keeps the empirical type I errors for all the three models. Based on simulation, X^2 performs well for Rosner's model, and the three bootstrap methods have problems of inflated type I error when R and ρ are relatively large in Rosner's and equal correlation coefficient models. In general the performance difference between different goodness-of-fit statistics is more significant when the sample size is small and/or the data are highly dependent. Two real examples of ophthalmologic studies are used to illustrate the different goodness-of-fit test methods.

So far our tests are all done based on asymptotic distribution of the statistics. The exact methods for small samples are left as interesting future work.

References

1. Ahn, C., Jung, S.-H., Donner, A.: Application of an adjusted χ^2 statistics to site-specific data in observational dental studies. J. Clin. Periodontol. **29**(1), 79–82 (2002). January
2. Dallal, G.E.: Paired bernoulli trials. Biometrics **44**, 253–257 (1988). March
3. Donner, A.: Statistical methods in opthalmology: an adjusted chi-square approach. Biometrics **45**(2), 605–611 (1989)
4. Donner, A.: Cluster randomization trials in epidemiology: theory and application. J. Stat. Plann. Inference **42**(1–2), 37–56 (1994)
5. Donner, A., Eliasziw, M.: A goodness-of-fit approach to inference procedures for kappa statistics: confidence interval construction, significance-testing and sample size estimation. Stat. Med. **11**, 1511–1519 (1992)
6. Liu, X. Shan, G., Tian, L., Ma, C.-X.: Exact methods for testing homogeneity of proportions for correlated multiple clustered binary data. Commun. Stat. Simul. Comput. **46**(8), 6074–6082 (2017)
7. Ma, C.-X., Liu, S.: Testing equality of proportions for correlated binary data in opthalmologic studies. J. Biopharm. Stat. **27**(4), 611–619 (2017). https://doi.org/10.1080/10543406.2016.1167072
8. Ma, C.-X., Shan, G., Liu, S.: Homogeneity test for binary correlated data. PLoS One **10**(4), e0124337 (2015)
9. Pei, Y.-B., Tang, M.-L., Guo, J.-H.: Testing the equality of two proportions for combined unilateral and bilateral data. Commun. Stat. Simul. Comput. **37**, 1515–1529 (2008)
10. Pei, Y.-B., Tang, M.-L., Wong, W.-K., Guo, J.-H.: Confidence intervals for correlated proportion differences from paired data in a two-arm randomized clinical trial. Stat. Methods Med. Res. **21**(2), 167–187 (2012). April
11. Rajavi, Z., Katibeh, M., Ziaei, H., Fardesmaeilpour, N., Sehat, M., Ahmadieh, H., Javadi, M.A.: Rapid assessment of avoidable blindness in Iran. Ophthalmology **118**(9), 1812–1818 (2011)

12. Rosner, B.: Statistical methods in ophthalmology: an adjustment for the intraclass correlation between eyes. Biometrics **38**, 105–114 (1982). March
13. Shan, G., Ma, C.-X.: Exact methods for testing the equality of proportions for clustered data. Stat. biopharma. Res. **6**(1), 115–122 (2014)
14. Tang, M.-L., Tang, N.-S., Rosner, B.: Statistical inference for correlated data in ophthalmologic studies. Stat. Med. **25**, 2771–2783 (2006)
15. Tang, N.-S., Tang, M.-L., Qiu, S.-F.: Testing the equality of proportions for correlated otolaryngologic data. Comput. Stat. Data Anal. **52**(7), 2719–3729 (2008)
16. Tang, M.-L., Pei, Y.-B., Wong, W.-K., Li, J.-L.: Goodness-of-fit tests for correlated paired binary data. Stat. Methods Med. Res. **21**(4), 331–345 (2012)
17. Tang, N.-S., Tang, M.-L., Qiu, S.-F., Pei, Y.-B.: Asymptotic confidence interval construction for proportion difference in medical studies with bilateral data. Stat. Methods Med. Res. **20**(3), 233–259 (2011). June

Chapter 21
A Bilinear Reduced Rank Model

Chengcheng Hao, Feng Li, and Dietrich von Rosen

Abstract This article considers a bilinear model that includes two different latent effects. The first effect has a direct influence on the response variable, whereas the second latent effect is assumed to first influence other latent variables, which in turn affect the response variable. In this article, latent variables are modelled via rank restrictions on unknown mean parameters and the models which are used are often referred to as reduced rank regression models. This article presents a likelihood-based approach that results in explicit estimators. In our model, the latent variables act as covariates that we know exist, but their direct influence is unknown and will therefore not be considered in detail. One example is if we observe hundreds of weather variables, but we cannot say which or how these variables affect plant growth.

21.1 Introduction

In the early age of statistics, variations in data were studied through applications of linear models. Analysis of variance, regression analysis and analysis of covariance were developed simultaneously and later were put under the same umbrella with

Dedicated to Professor Kai-Tai Fang and his wife Ting Mei.

C. Hao
Shanghai University of International Business and Economics, No. 1900 Wenxiang Road, Songjiang District, 201620 Shanghai, People's Republic of China
e-mail: chengcheng.hao@outlook.com

F. Li
Central University of Finance and Economics, 39 South College Road, Haidian District, Beijing 100081, People's Republic of China
e-mail: feng.li@cufe.edu.cn

D. von Rosen (✉)
Swedish University of Agricultural Sciences, Box 7032, 750 07 Uppsala, Sweden
e-mail: dietrich.von.rosen@slu.se

Linköping University, SE-581 83, Linköping, Sweden

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_21

matrix theory. In the beginning, only one response variable was considered but soon, due to the knowledge of how to simultaneously handle sample variances and sample covariances, multivariate analysis was developed including multivariate analysis of variance (MANOVA), principal component analysis and canonical correlation analysis. An interesting article that reviews multivariate analysis up to the 1940s was written by Rao [6]. Shortly after Rao's article, Andersson [1] wrote a seminal paper on multivariate analysis in which, among other methods, reduced rank problems were considered; i.e., the cases in which the matrix of regression parameters was not of full rank, which is a case that obviously does not exist in univariate linear regression models, where the rank always equals one.

Bilinear regression models were indirectly considered in the above-mentioned article by Anderson [1], in which bilinear restrictions in a MANOVA model were handled. However, the article by Potthoff and Roy [5] is usually considered the first article on the analysis of bilinear models, which introduced the so called growth curve model, but there had been several earlier contributions in the analysis of growth curves, which were all bilinear. For references on bilinear regression models (growth curve models/GMANOVA) see, for example, von Rosen [9]. Moreover, our bilinear models are balanced multivariate models with a linearly structured mean in contrast to a MANOVA model for which an arbitrary mean structure is assumed to hold.

This article mainly considers reduced rank analysis applied to the analysis of growth curve models via knowledge about the analysis of extended bilinear regression models. The book by Reinsel and Velu [8] includes many references to reduced rank regression analysis presents many examples in which the models are used. In von Rosen and von Rosen [10], some of Reinsel and Velu's [8] work on growth curves with rank restrictions was extended.

Reduced rank models imaginarily can be connected to latent variables. For example, if observing many weather variables, such as hourly temperature and precipitation data collected over a month, the effect on plant growth will occur via some unobserved latent processes. In this article, we will introduce the case in which latent variables influence a response variable directly, but the latent variables also influence other latent variables that then influence the response variable. In our example, temperature and precipitation also influence many soil characteristics, and these variables, through some unobserved latent processes, affect plant growth. Thus, temperature and precipitation also form a basis for latent variables to act indirectly. Some details connected to this example will be provided in Sect. 21.3. This way of thinking has been implemented in statistical model building which is connected to graphical models, but in this case, the focus is on modelling covariance matrices. Moreover, in factor analysis (i.e. structural equation modelling), one models latent variables via covariance structures.

Suppose that there is a parameter matrix $\boldsymbol{\Theta}$ of size $p \times q$ and rank r; i.e., $r(\boldsymbol{\Theta}) = r$. This supposition means that there exist linear combinations $\mathbf{L}'\boldsymbol{\Theta} = \mathbf{0}$, but where $\mathbf{L}$: $(p - r) \times p$ is unknown. Solving $\mathbf{L}'\boldsymbol{\Theta} = \mathbf{0}$ as a function of $\boldsymbol{\Theta}$ implies the factorization $\boldsymbol{\Theta} = \boldsymbol{\Theta}_1\boldsymbol{\Theta}_2$, $\boldsymbol{\Theta}_1$ $p \times r$ and $\boldsymbol{\Theta}_2$ $r \times q$, where both matrices are of rank equal to r. This implication follows from the fact that $\mathcal{C}(\boldsymbol{\Theta}_1) = \mathcal{C}(\mathbf{L})^{\perp}$, where $\mathcal{C}(\bullet)$ and $\mathcal{C}(\bullet)^{\perp}$ denote the column space and its orthogonal complement, respec-

tively, and $\boldsymbol{\Theta}_2$ is an arbitrary matrix that generats all solutions to $\mathbf{L}'\boldsymbol{\Theta} = \mathbf{0}$. However, since $\mathbf{L}$ is arbitrary, $\boldsymbol{\Theta}_1$ is also arbitrary. Moreover, the rank restriction also implies restrictions among the columns of $\boldsymbol{\Theta}$; i.e., $\mathbf{H}\boldsymbol{\Theta}' = \mathbf{0}$. Thus, $\boldsymbol{\Theta} = \boldsymbol{\Theta}_1\boldsymbol{\Theta}_2$, where $\boldsymbol{\Theta}_2$ is of full rank, satisfying $\mathscr{C}(\boldsymbol{\Theta}_2') = \mathscr{C}(\mathbf{H}')^{\perp}$, and of course, $\boldsymbol{\Theta}_1$ is unknown. Therefore, with rank restrictions, we have difficulties interpreting $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ because we do not know if we have row- or column-restrictions. It also follows that without further conditions, $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ are not estimable; therefore, in this article, the focus will be on estimating $\boldsymbol{\Theta}$ and not $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$. We can conclude that by putting rank restrictions on a matrix is not very informative. However, there exists another type of modelling where one starts with $\boldsymbol{\Theta}_1$ and $\boldsymbol{\Theta}_2$ and multiplies these matrices together; and this case, one usually has a clear interpretation of the matrices. This type of model has been studies in so-called cointegration analysis in econometrics (see Johansen [3]).

In Sect. 21.2, the proposed model is described in detail. To the best of our knowledge, this method for modelling the indirect effects of latent variables has not been considered before. Moreover, the chapter provides one example in which the model can be used. In Sect. 21.3, a likelihood-based approach to estimate the parameters is proposed.

21.2 Model

Denote $\mathbf{X} : p \times n$ as the random matrix corresponding to p repeated measurements of n independent observations. The model that will be studied is given by

$$\mathbf{X} = \mathbf{ABC}_1 + \boldsymbol{\Theta}\mathbf{C}_2 + \boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3 + \mathbf{E}, \tag{21.1}$$

where $\mathbf{A} : p \times q$ is a known within-individuals design matrix, the matrices $\mathbf{C}_1 : k \times n, \mathbf{C}_2 : k_1 \times n$ and $\mathbf{C}_3 : k_1 \times n$ are known between-individuals design matrices with column spaces satisfying $\mathscr{C}(\mathbf{C}_3') \subset \mathscr{C}(\mathbf{C}_2') \subset \mathscr{C}(\mathbf{C}_1')$, and $\mathbf{E} : p \times n$ is a random error matrix following the matrix normal distribution $N_{p,n}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I_n})$, where the covariance matrix $\boldsymbol{\Sigma} : p \times p$ is an unknown positive definite matrix. Moreover, the matrices $\mathbf{B} : q \times k$, $\boldsymbol{\Theta} : p \times k_1$ and $\boldsymbol{\Psi} : p \times p$ are unknown mean parameters, where $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$ have rank restrictions $r(\boldsymbol{\Theta}) = r_1 < \min(p, k_1)$ and $r(\boldsymbol{\Psi}) = r_2 < p$. Models with a product such as $\boldsymbol{\Psi}\boldsymbol{\Theta}$ in (21.1), with rank restrictions on both included matrices, have, to the best of our knowledge, not been considered before. In this model, the main purpose is to estimate $\mathbf{B}$ while adjusting for the latent effects that are introduced in the model via rank restrictions.

The condition $\mathscr{C}(\mathbf{C}_3') \subset \mathscr{C}(\mathbf{C}_2') \subset \mathscr{C}(\mathbf{C}_1')$ is a technical condition and is motivated by knowledge about extended bilinear regression models (see von Rosen [9]). In the condition, strict subspace inequalities, which are pure estimability conditions since $\boldsymbol{\Theta}$ is included in two effect terms, are necessary. It will also be assumed that

$$\mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\boldsymbol{\Theta}) = \{\mathbf{0}\},$$

which is a very natural condition and will, in principle, not put any restrictions on the use of the model in (21.1).

If the terms $\boldsymbol{\Theta}\mathbf{C}_2$ and $\boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3$ are not included in (21.1), we would have the traditional growth curve model (see Potthoff and Roy [5], von Rosen [9]), or if $\boldsymbol{\Theta}$ does not have rank restrictions, and the term $\boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3$ is not included, then the model is a GMANOVA+MANOVA model (see Chinchilli and Elsewick [2]). Moreover, in (21.1), the term $\boldsymbol{\Theta}\mathbf{C}_2$ represents the direct latent effects, whereas the term $\boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3$ mimics indirect latent effects or the interaction between two latent variables within a nested regime. Throughout the article, it will be assumed that n is so large that the parameters can be estimated.

We finally stress that the main purpose is to estimate the growth curves that are adjusted for latent effects, and we do not discuss the latent effects directly because it is difficult to interpret the estimators of the rank restricted parameter matrices.

21.2.1 Example

The purpose of this example is to motivate the model given by (21.1). We have chosen to use plant growth and weather characteristics to illustrate the model, but currently, when we have the ability to measure many complex systems, there are also many other examples where the model can be applied, for example, within neurosciences or when studying financial markets.

Plant Growth

Suppose that the aim of a study is to compare two treatments (blocks). The study comprises 10 types of plants. Two types of plants have long roots and are not affected by the characteristics of the upper layer of soil (e.g., chemical variables, soil textures, organic materials); two types of plants are very robust against weather conditions (different types of summary measures of temperature and precipitation) and soil characteristics, whereas the remaining six types of plants are affected by both weather and soil variables. Let the study comprise $n = 200$ observations, and suppose that for all 10 plant types, there are two blocks per plant type, all of which are of equal size. A between-individuals design matrix $\mathbf{C}_1$ for specifying the "growth curve" $\mathbf{ABC}_1$ can have the following form:

$$\mathbf{C}_1 = \mathbf{I}_{10} \otimes \begin{pmatrix} \mathbf{1}'_{10} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}'_{10} \end{pmatrix},$$

where $\mathbf{1}_{10}$ is a vector of ones of size 10 and $\otimes$ denotes the Kronecker product of two matrices. The within-individuals design matrix $\mathbf{A}$ in the model given in (21.1) equals, if $p = 10$ and there is a linear growth (t_i represents the ith time point),

$$\mathbf{A}' = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \mathrm{t}_1 & \mathrm{t}_2 & \mathrm{t}_3 & \mathrm{t}_4 & \mathrm{t}_5 & \mathrm{t}_6 & \mathrm{t}_7 & \mathrm{t}_8 & \mathrm{t}_9 & \mathrm{t}_{10} \end{pmatrix}.$$

Concerning the weather and soil variables, suppose that we have 10 weather variables and 10 soil variables. We will specify $\mathbf{C}_2$ and $\mathbf{C}_3$ in the model presented in (21.1). Let $\mathbf{g}_i$ denote a vector of size 10 in which the 10 weather variables for plant type i are stored, and let $\mathbf{s}_i$ be a vector in which the soil observations for plant type i are stored. It will be supposed that weather and soil variables are constant for each plant type, meaning that all plants from a specific plant type grow in places with the same soil and weather characteristics. Therefore, for each plant type, we have specific background matrices. To handle weather observations in the model, we define $\mathbf{V}_i$, $i \in \{1, \ldots, 10\}$, such that

$$\mathbf{V}_i = \begin{cases} \mathbf{1}'_{20} \otimes \mathbf{g}_i, & 10 \times 20, \quad \text{if } i \in \{1, \ldots, 8\}, \\ \mathbf{1}'_{20} \otimes \mathbf{0}, & 10 \times 20, \quad \text{if } i \in \{9, 10\}. \end{cases}$$

Then $\mathbf{C}_2$ in (21.1) is defined as

$$\mathbf{C}_2 = \text{Block}(\mathbf{V}_1, \ldots, \mathbf{V}_{10}), \qquad 100 \times 200,$$

where Block denotes the block diagonal operator. It follows that $\mathscr{C}(\mathbf{C}'_2) \subset \mathscr{C}(\mathbf{C}'_1)$. To see this relationship, note that

$$\mathbf{V}'_i = \begin{pmatrix} \mathbf{1}_{10} & \mathbf{0} \\ \mathbf{0} & \mathbf{1}_{10} \end{pmatrix} \begin{pmatrix} \mathbf{g}'_i \\ \mathbf{g}'_i \end{pmatrix}, \quad i \in \{1, \ldots, 8\}.$$

Strict inclusion between the subspaces holds because for $i \in \{9, 10\}$, $\mathbf{V}_i$ equals $\mathbf{0}$. It can be noted that the model states that there are eight plant types that are directly affected by weather conditions via the term $\boldsymbol{\Theta}\mathbf{C}_2$.

Concerning the soil variable, let

$$\mathbf{M}_i = \begin{cases} \mathbf{1}'_{20} \otimes \mathbf{s}_i, & 10 \times 20, \quad \text{if } i \in \{1, \ldots, 6\}, \\ \mathbf{1}'_{20} \otimes \mathbf{0}, & 10 \times 20, \quad \text{if } i \in \{7, \ldots, 10\}. \end{cases}$$

Then, $\mathbf{C}_3$ in (21.1) is given by

$$\mathbf{C}_3 = \text{Block}(\mathbf{M}_1, \ldots, \mathbf{M}_{10}), \qquad 100 \times 200.$$

Since, $\mathscr{C}(\mathbf{M}'_i) = \mathscr{C}(\mathbf{V}'_i)$, $i \in \{1, \ldots, 6\}$, it follows that $\mathscr{C}(\mathbf{C}'_3) \subset \mathscr{C}(\mathbf{C}'_2)$.

Here, $\mathbf{C}_3$ is constructed so that only plant types on which there is an influence by weather and soil characteristics are included. We can think of the latent processes of weather variables as effecting the soil characteristics, which in turn have an influence on plant growth via some new latent variables.

21.3 Estimation

In this section, likelihood-inspired estimators are established. For notational convenience, $(\mathbf{Q})(\mathbf{Q})'$ is written as $(\mathbf{Q})()'$, where $\mathbf{Q}$ can be any matrix expression. The likelihood function for model (21.1) equals

$$L(\mathbf{B}, \boldsymbol{\Theta}, \boldsymbol{\Psi}, \boldsymbol{\Sigma}) = (2\pi)^{-\frac{1}{2}pn} e^{-\frac{1}{2}\mathrm{tr}\{\boldsymbol{\Sigma}^{-1}(\mathbf{X}-\mathbf{ABC}_1-\boldsymbol{\Theta}\mathbf{C}_2-\boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3)()'\}} |\boldsymbol{\Sigma}|^{-\frac{1}{2}n}.$$

Using a well-known inequality (see Srivastava and Khatri [11], Theorem 1.10.4)

$$L(\mathbf{B}, \boldsymbol{\Theta}, \boldsymbol{\Psi}, \boldsymbol{\Sigma}) \leq |n^{-1}(\mathbf{X} - \mathbf{ABC}_1 - \boldsymbol{\Theta}\mathbf{C}_2 - \boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3)()'|^{-\frac{1}{2}n}(2\pi)^{-\frac{1}{2}pn} e^{-\frac{1}{2}pn}, \tag{21.2}$$

with equality if and only if

$$n\boldsymbol{\Sigma} = (\mathbf{X} - \mathbf{ABC}_1 - \boldsymbol{\Theta}\mathbf{C}_2 - \boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3)()'.$$

Thus, $\boldsymbol{\Sigma}$ will be estimated if $\mathbf{B}$, $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$ can be estimated.

Let $\mathbf{S}_1 = \mathbf{X}(\mathbf{I} - \mathbf{P}_{\mathbf{C}_1'})\mathbf{X}'$, $\mathbf{P}_{\mathbf{C}_1'} = \mathbf{C}_1'(\mathbf{C}_1\mathbf{C}_1')^-\mathbf{C}_1$, and $\mathbf{V}_1 = \mathbf{XP}_{\mathbf{C}_1'} - \mathbf{ABC}_1 - \boldsymbol{\Theta}\mathbf{C}_2 - \boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3$, where for an arbitrary $\mathbf{Q}$ the notation $(\mathbf{Q})^-$ denotes any g-inverse of $\mathbf{Q}$ satisfying the well-known relation $\mathbf{QQ}^-\mathbf{Q} = \mathbf{Q}$.

Moreover, in the subsequent calculations, the determinant relation $|\mathbf{I} + \mathbf{QR}| = |\mathbf{I} + \mathbf{RQ}|$ for arbitrary $\mathbf{Q}$ and $\mathbf{R}$ will be used many times. Minimizing the determinant in (21.2) will be the main objective, and we start by performing a number of calculations leading to the following chain of equalities:

$$\begin{aligned}
&|(\mathbf{X} - \mathbf{ABC}_1 - \boldsymbol{\Theta}\mathbf{C}_2 - \boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3)()'| \\
&\quad = |\mathbf{S}_1 + (\mathbf{XP}_{\mathbf{C}_1'} - \mathbf{ABC}_1 - \boldsymbol{\Theta}\mathbf{C}_2 - \boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3)()'| = |\mathbf{S}_1||\mathbf{I} + \mathbf{V}_1'\mathbf{S}_1^{-1}\mathbf{V}_1| \\
&\quad = |\mathbf{S}_1||\mathbf{I} + \mathbf{V}_1'\mathbf{S}_1^{-1}\mathbf{A}(\mathbf{A}'\mathbf{S}_1^{-1}\mathbf{A})^-\mathbf{A}'\mathbf{S}_1^{-1}\mathbf{V}_1 + \mathbf{V}_1'\mathbf{A}^o(\mathbf{A}^{o'}\mathbf{S}_1\mathbf{A}^o)^-\mathbf{A}^{o'}\mathbf{V}_1|, \quad (21.3)
\end{aligned}$$

where $\mathbf{S}^{-1} = \mathbf{S}^{-1}\mathbf{P}_{\mathbf{A},\mathbf{S}} + \mathbf{P}_{\mathbf{A}^o,\mathbf{S}^{-1}}\mathbf{S}^{-1}$, with $\mathbf{P}_{\mathbf{A},\mathbf{S}} = \mathbf{A}(\mathbf{A}'\mathbf{S}^{-1}\mathbf{A})^-\mathbf{A}'\mathbf{S}^{-1}$, and $\mathbf{A}^o$ is any matrix generating $\mathscr{C}(\mathbf{A})^\perp$ (see Kollo and von Rosen [4], Theorem 1.2.25). Since $\mathbf{A}^{o'}\mathbf{ABC}_1 = \mathbf{0}$, it follows that

$$\begin{aligned}
&\text{the r.h.s. of (21.3)} \\
&\geq |\mathbf{S}_1||\mathbf{I} + (\mathbf{XP}_{\mathbf{C}_1'} - \boldsymbol{\Theta}\mathbf{C}_2 - \boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3)'\mathbf{A}^o(\mathbf{A}^{o'}\mathbf{S}_1\mathbf{A}^o)^-(\mathbf{XP}_{\mathbf{C}_1'} - \boldsymbol{\Theta}\mathbf{C}_2 - \boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3)|, \quad (21.4)
\end{aligned}$$

(r.h.s. is an abbreviation for "right-hand side") with equality if and only if $\mathbf{A}'\mathbf{S}_1^{-1}\mathbf{V}_1 = \mathbf{0}$; that is,

$$\mathbf{ABC}_1 = \mathbf{A}(\mathbf{A}'\mathbf{S}_1^{-1}\mathbf{A})^-\mathbf{A}'\mathbf{S}_1^{-1}(\mathbf{XP}_{\mathbf{C}_1'} - \boldsymbol{\Theta}\mathbf{C}_2 - \boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3).$$

Thus, $\mathbf{B}$ is estimated if $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$ can be estimated because as a function of $\boldsymbol{\Theta}$ and $\boldsymbol{\Psi}$ we have a consistent system of linear equations. The above block of calculations will be repeated two times before the estimators are obtained.

Let us continue with (21.4). We need a few more definitions (compare with $\mathbf{S}_1$ and $\mathbf{V}_1$). Let

$$\begin{aligned}
\mathbf{T}_1 &= \mathbf{S}_1\mathbf{A}^o(\mathbf{A}^{o'}\mathbf{S}_1\mathbf{A}^o)^-\mathbf{A}^{o'} = \mathbf{P}'_{\mathbf{A}^{o'},\mathbf{S}_1^{-1}},\\
\mathbf{S}_2 &= \mathbf{S}_1 + \mathbf{T}_1\mathbf{X}(\mathbf{P}_{\mathbf{C}'_1} - \mathbf{P}_{\mathbf{C}'_2})\mathbf{X}'\mathbf{T}'_1,\\
\mathbf{V}_2 &= \mathbf{X}\mathbf{P}_{\mathbf{C}_2} - \boldsymbol{\Theta}\mathbf{C}_2 - \boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3.
\end{aligned}$$

Then,

$$\begin{aligned}
&\text{the r.h.s. of (21.4)}\\
&= |\mathbf{S}_1 + \mathbf{T}_1(\mathbf{X}\mathbf{P}_{\mathbf{C}'_1} - \boldsymbol{\Theta}\mathbf{C}_2 - \boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3)()'\mathbf{T}'_1|\\
&= |\mathbf{S}_2 + \mathbf{T}_1\mathbf{V}_2\mathbf{V}'_2\mathbf{T}'_1| = |\mathbf{S}_2||\mathbf{I} + \mathbf{V}'_2\mathbf{T}'_1\mathbf{S}_2^{-1}\mathbf{T}_1\mathbf{V}_2|\\
&= |\mathbf{S}_2||\mathbf{I} + \mathbf{V}'_2\mathbf{T}'_1\mathbf{S}_2^{-1}\mathbf{P}_{\mathbf{T}_1\boldsymbol{\Theta}_1,\mathbf{S}_2}\mathbf{T}_1\mathbf{V}_2 + \mathbf{V}'_2\mathbf{T}'_1\mathbf{P}_{(\mathbf{T}_1\boldsymbol{\Theta}_1)^o,\mathbf{S}_2^{-1}}\mathbf{S}_2^{-1}\mathbf{T}_1\mathbf{V}_2|, \qquad (21.5)
\end{aligned}$$

because of the rank restrictions $r(\boldsymbol{\Theta}) = r_1$ and $\boldsymbol{\Theta} = \boldsymbol{\Theta}_1\boldsymbol{\Theta}_2$, for some $\boldsymbol{\Theta}_1 : p \times r_1$, $\boldsymbol{\Theta}_2 : r_1 \times k_1$ which both are of rank r_1 and unknown. Moreover,

$$\text{the r.h.s. of (21.5)} \geq |\mathbf{S}_2||\mathbf{I} + \mathbf{V}'_2\mathbf{T}'_1\mathbf{P}_{(\mathbf{T}_1\boldsymbol{\Theta}_1)^o,\mathbf{S}_2}\mathbf{S}_2^{-1}\mathbf{T}_1\mathbf{V}_2| \qquad (21.6)$$

with equality if and only if $\boldsymbol{\Theta}'_1\mathbf{T}'_1\mathbf{S}_2^{-1}\mathbf{T}_1\mathbf{V}_2 = \mathbf{0}$, which in turn implies that

$$\boldsymbol{\Theta}_1\boldsymbol{\Theta}_2\mathbf{C}_2 = \boldsymbol{\Theta}_1(\boldsymbol{\Theta}'_1\mathbf{T}'_1\mathbf{S}_2^{-1}\mathbf{T}_1\boldsymbol{\Theta}_1)^{-1}\boldsymbol{\Theta}'_1\mathbf{T}'_1\mathbf{S}_2^{-1}(\mathbf{X}\mathbf{P}_{\mathbf{C}'_2} - \boldsymbol{\Psi}\boldsymbol{\Theta}_1\boldsymbol{\Theta}_2\mathbf{C}_3), \qquad (21.7)$$

where the inverse exists because $\boldsymbol{\Theta}_1$ is of full column rank. If we can find an estimator for $\boldsymbol{\Theta}_1$ and consider $\boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3$ to be known, we have a system of consistent linear equations in $\boldsymbol{\Theta}_2$.

We proceed with (21.6). Let

$$\begin{aligned}
\mathbf{T}_2 &= \mathbf{S}_2(\mathbf{T}_1\boldsymbol{\Theta}_1)^o((\mathbf{T}_1\boldsymbol{\Theta}_1)^{o'}\mathbf{S}_2(\mathbf{T}_1\boldsymbol{\Theta}_1)^o)^-(\mathbf{T}_1\boldsymbol{\Theta}_1)^{o'} = \mathbf{P}_{(\mathbf{T}_1\boldsymbol{\Theta}_1)^o,\mathbf{S}_2^{-1}},\\
\mathbf{S}_3 &= \mathbf{S}_2 + \mathbf{T}_2\mathbf{T}_1\mathbf{X}(\mathbf{P}_{\mathbf{C}'_2} - \mathbf{P}_{\mathbf{C}'_3})\mathbf{X}'\mathbf{T}'_1\mathbf{T}'_2,\\
\mathbf{V}_3 &= \mathbf{X}\mathbf{P}_{\mathbf{C}'_3} - \boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3.
\end{aligned}$$

Then,

$$\begin{aligned}
\text{the r.h.s. of (21.6)} &= |\mathbf{S}_2 + \mathbf{T}_2\mathbf{T}_1(\mathbf{X}\mathbf{P}_{\mathbf{C}'_2} - \boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3)()'\mathbf{T}'_1\mathbf{T}'_2|\\
&= |\mathbf{S}_3 + \mathbf{T}_2\mathbf{T}_1(\mathbf{X}\mathbf{P}_{\mathbf{C}'_3} - \boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3)()'\mathbf{T}'_1\mathbf{T}'_2| = |\mathbf{S}_3 + \mathbf{T}_2\mathbf{T}_1\mathbf{V}_3\mathbf{V}'_3\mathbf{T}'_1\mathbf{T}'_2|\\
&= |\mathbf{S}_3||\mathbf{I} + \mathbf{V}'_3\mathbf{T}'_1\mathbf{T}'_2\mathbf{S}_3^{-1}\mathbf{T}_2\mathbf{T}_1\mathbf{V}_3|. \qquad (21.8)
\end{aligned}$$

The determinant $|\mathbf{S}_3|$ is a function of $\boldsymbol{\Theta}_1$ since $\mathbf{T}_2$ is a function of $\boldsymbol{\Theta}_1$. Now, we will minimize $|\mathbf{S}_3|$ with respect to $\boldsymbol{\Theta}_1$, which implies that we are not aiming to find maximum likelihood estimates because $\boldsymbol{\Theta}_1$ is also included in the other determinant of (21.8). However, by focusing only on $|\mathbf{S}_3|$, it will be shown that explicit estimators can be obtained. Let

$$\begin{aligned}\mathbf{P}_{\mathbf{C}_2'\backslash\mathbf{C}_3'} &= \mathbf{P}_{\mathbf{C}_2'} - \mathbf{P}_{\mathbf{C}_3'},\\ \mathbf{R}_1 &= \mathbf{I} + \mathbf{P}_{\mathbf{C}_2'\backslash\mathbf{C}_3'}\mathbf{X}'\mathbf{H}_1\mathbf{H}_1'\mathbf{X}\mathbf{P}_{\mathbf{C}_2'\backslash\mathbf{C}_3'},\end{aligned}$$

where for some $\mathbf{H}_1 : p \times (p - r(\mathbf{A}))$

$$\mathbf{T}_1'\mathbf{S}_2^{-1}\mathbf{T}_1 = \mathbf{H}_1\mathbf{H}_1'.$$

It follows that

$$\begin{aligned}|\mathbf{S}_3| &= |\mathbf{S}_2|\mathbf{I} + \mathbf{P}_{\mathbf{C}_2'\backslash\mathbf{C}_3'}\mathbf{X}'\mathbf{T}_1'\mathbf{T}_2'\mathbf{S}_2^{-1}\mathbf{T}_2\mathbf{T}_1\mathbf{X}\mathbf{P}_{\mathbf{C}_2'\backslash\mathbf{C}_3'} \qquad (21.9)\\ &= |\mathbf{S}_2||\mathbf{R}_1 - \mathbf{P}_{\mathbf{C}_2'\backslash\mathbf{C}_3'}\mathbf{X}'\mathbf{H}_1\mathbf{H}_1'\boldsymbol{\Theta}_1(\boldsymbol{\Theta}_1'\mathbf{H}_1\mathbf{H}_1'\boldsymbol{\Theta}_1)^{-1}\boldsymbol{\Theta}_1'\mathbf{H}_1\mathbf{H}_1'\mathbf{X}\mathbf{P}_{\mathbf{C}_2'\backslash\mathbf{C}_3'}\\ &= |\mathbf{S}_2||\mathbf{R}_1||\mathbf{I} - \mathbf{F}_1'\mathbf{H}_1'\mathbf{X}\mathbf{P}_{\mathbf{C}_2'\backslash\mathbf{C}_3'}\mathbf{R}_1^{-1}\mathbf{P}_{\mathbf{C}_2'\backslash\mathbf{C}_3'}\mathbf{X}'\mathbf{H}_1\mathbf{F}_1|,\end{aligned}$$

where

$$\mathbf{F}_1 = \mathbf{H}_1'\boldsymbol{\Theta}_1(\boldsymbol{\Theta}_1'\mathbf{H}_1\mathbf{H}_1'\boldsymbol{\Theta}_1)^{-1/2}, \quad \mathbf{F}_1'\mathbf{F}_1 = \mathbf{I}_{r_1}, \quad \mathbf{F}_1 : (p - r(\mathbf{A})) \times r_1.$$

Let

$$\mathbf{U} = \mathbf{I} - \mathbf{H}_1'\mathbf{X}\mathbf{P}_{\mathbf{C}_2'\backslash\mathbf{C}_3'}\mathbf{R}_1^{-1}\mathbf{P}_{\mathbf{C}_2'\backslash\mathbf{C}_3'}\mathbf{X}'\mathbf{H}_1$$

which is positive definite since

$$\mathbf{U}^{-1} = \mathbf{I} + \mathbf{H}_1'\mathbf{X}\mathbf{P}_{\mathbf{C}_2'\backslash\mathbf{C}_3'}\mathbf{X}'\mathbf{H}_1$$

is positive definite. Thus,

$$\text{the r.h.s. of (21.9)} = |\mathbf{S}_2||\mathbf{R}_1||\mathbf{F}_1'\mathbf{U}\mathbf{F}_1| \geq |\mathbf{S}_2||\mathbf{R}_1|\prod_{i=1}^{r_1}\lambda_{p-r(\mathbf{A})-r_1+i}, \quad (21.10)$$

where $\lambda_1 \geq \cdots \geq \lambda_{p-r(\mathbf{A})}$ are the ordered eigen-values of $\mathbf{U}$, which are all independent of $\boldsymbol{\Theta}_1$ since $\mathbf{U}$ is not a function of $\boldsymbol{\Theta}_1$. The inequality follows from Rao [7], Theorem 2.1 (Poincaré separation theorem). Let $\{\mathbf{v}_i\}$ be the corresponding eigenvectors to $\{\lambda_{p-r(\mathbf{A})-r_1+i}\}$. Then, the minimum in (21.10) is obtained if $\mathbf{F}_1$ is chosen to equal

$$\widetilde{\mathbf{F}}\mathbf{F}_1 = (\mathbf{v}_1, \ldots, \mathbf{v}_{r_1})$$

and it remains to find a $\boldsymbol{\Theta}_1$ such that

$$\widetilde{\mathbf{F}}\mathbf{F}_1 = \mathbf{H}_1'\boldsymbol{\Theta}_1(\boldsymbol{\Theta}_1'\mathbf{H}_1\mathbf{H}_1'\boldsymbol{\Theta}_1)^{-1/2}.$$

Since $\widetilde{\mathbf{F}}\mathbf{F}_1'\widetilde{\mathbf{F}}\mathbf{F}_1 = \mathbf{I}_{r_1}$, one solution is given by

$$\widehat{\boldsymbol{\Theta}}_1 = \mathbf{H}_1(\mathbf{H}_1'\mathbf{H}_1)^{-1}\widetilde{\mathbf{F}}\mathbf{F}_1. \tag{21.11}$$

We will later return to the estimation of $\boldsymbol{\Theta}$ because according to (21.7), the estimator of $\boldsymbol{\Theta}$ will be a function of the estimator of $\boldsymbol{\Psi}$, and therefore, $\boldsymbol{\Psi} = \boldsymbol{\Psi}_1\boldsymbol{\Psi}_2$, where $\boldsymbol{\Psi}_1 : p \times r_2$, $\boldsymbol{\Psi}_2 : r_2 \times p$ has to be discussed. Let us start with (21.8) and

the r.h.s. of (21.8)

$$\begin{aligned}
&= |\mathbf{S}_3||\mathbf{I} + \mathbf{V}_3'\mathbf{T}_1'\mathbf{T}_2'\mathbf{S}_3^{-1}\mathbf{P}_{\mathbf{T}_2\mathbf{T}_1\boldsymbol{\Psi}_1,\mathbf{S}_3}\mathbf{T}_2\mathbf{T}_1\mathbf{V}_3 + \mathbf{V}_3'\mathbf{T}_1'\mathbf{T}_2'\mathbf{P}_{(\mathbf{T}_2\mathbf{T}_1\boldsymbol{\Psi}_1)^o,\mathbf{S}_3^{-1}}\mathbf{S}_3^{-1}\mathbf{T}_2\mathbf{T}_1\mathbf{V}_3|\\
&\geq |\mathbf{S}_3||\mathbf{I} + \mathbf{V}_3'\mathbf{T}_1'\mathbf{T}_2'\mathbf{P}_{(\mathbf{T}_2\mathbf{T}_1\boldsymbol{\Psi}_1)^o,\mathbf{S}_3^{-1}}\mathbf{S}_3^{-1}\mathbf{T}_2\mathbf{T}_1\mathbf{V}_3|. \qquad (21.12)
\end{aligned}$$

Equality holds if and only if

$$\boldsymbol{\Psi}_1'\mathbf{T}_1'\mathbf{T}_2'\mathbf{S}_3^{-1}\mathbf{T}_2\mathbf{T}_1\mathbf{V}_3 = \mathbf{0},$$

where $\mathbf{S}_3$ and $\mathbf{T}_2$ are functions of $\boldsymbol{\Theta}_1$ for which an estimate was presented in (21.11). Hence,

$$\boldsymbol{\Psi}_1\boldsymbol{\Psi}_2\boldsymbol{\Theta}\mathbf{C}_3 = \boldsymbol{\Psi}_1(\boldsymbol{\Psi}_1'\mathbf{T}_1'\mathbf{T}_2'\mathbf{S}_3^{-1}\mathbf{T}_2\mathbf{T}_1\boldsymbol{\Psi}_1)^{-1}\boldsymbol{\Psi}_1'\mathbf{T}_1'\mathbf{T}_2'\mathbf{S}_3^{-1}\mathbf{X}\mathbf{P}_{\mathbf{C}_3'} \tag{21.13}$$

and $\boldsymbol{\Psi}_2$ can be estimated as a function of $\boldsymbol{\Psi}_1$ and $\boldsymbol{\Theta}$. Note that (21.13) implies that $\boldsymbol{\Psi}\boldsymbol{\Theta}\mathbf{C}_3$ is determined if $\boldsymbol{\Psi}_1$ is replaced by an estimate since $\boldsymbol{\Theta}_1$ in $\mathbf{S}_3$ and $\mathbf{T}_2$ has been estimated. Moreover, the expression in (21.13) can be inserted into (21.7) and given $\widehat{\boldsymbol{\Theta}}_1$, we can write

$$\begin{aligned}
\widehat{\boldsymbol{\Theta}}\mathbf{C}_2 &= \widehat{\boldsymbol{\Theta}}_1(\widehat{\boldsymbol{\Theta}}_1'\mathbf{T}_1'\mathbf{S}_2^{-1}\mathbf{T}_1\widehat{\boldsymbol{\Theta}}_1)^{-1}\widehat{\boldsymbol{\Theta}}_1'\mathbf{T}_1'\mathbf{S}_2^{-1}(\mathbf{X}\mathbf{P}_{\mathbf{C}_2'} - \widehat{\boldsymbol{\Psi}\boldsymbol{\Theta}}\mathbf{C}_3) \qquad (21.14)\\
&= \widehat{\boldsymbol{\Theta}}_1\widehat{\boldsymbol{\Theta}}_1'\mathbf{T}_1'\mathbf{S}_2^{-1}(\mathbf{X}\mathbf{P}_{\mathbf{C}_2'} - \widehat{\boldsymbol{\Psi}\boldsymbol{\Theta}}\mathbf{C}_3),
\end{aligned}$$

where $\widehat{\boldsymbol{\Psi}\boldsymbol{\Theta}}$ indicates that (21.13) is used, assuming that $\boldsymbol{\Psi}_1$ can be estimated. Thus, from this expression, $\boldsymbol{\Theta}_2$ can be estimated; however, $\boldsymbol{\Theta}_2$ is not unique and not of any greater interest.

The parameter matrix $\boldsymbol{\Psi}_1$ remains to estimate, which is important because this will give explicit estimators for $\widehat{\boldsymbol{\Theta}}\mathbf{C}_2$ and $\widehat{\boldsymbol{\Psi}\boldsymbol{\Theta}}\mathbf{C}_3$. The estimation will be carried out in the same way as when $\boldsymbol{\Theta}_1$ was estimated. Let

$$\begin{aligned}\mathbf{R}_2 &= \mathbf{I} + \mathbf{P}_{\mathbf{C}_3'}\mathbf{X}'\mathbf{T}_1'\mathbf{T}_2'\mathbf{S}_3^{-1}\mathbf{T}_2\mathbf{T}_2\mathbf{X}\mathbf{P}_{\mathbf{C}_3'},\\ \mathbf{T}_1'\mathbf{T}_2'\mathbf{S}_3^{-1}\mathbf{T}_2\mathbf{T}_2 &= \mathbf{H}_2\mathbf{H}_2', \quad \mathbf{H}_2 : \ p \times (p - r(\mathbf{A} : \boldsymbol{\Theta})),\\ \mathbf{F}_2 &= \mathbf{H}_2'\boldsymbol{\Psi}_1(\boldsymbol{\Psi}_1'\mathbf{H}_2\mathbf{H}_2'\boldsymbol{\Psi}_1)^{-1/2},\end{aligned}$$

where $\mathbf{R}_2$ and $\mathbf{H}_2$ depend on only one unknown quantity, $\boldsymbol{\Theta}_1$, which has been estimated in (21.11). From (21.12), it follows that

$$\text{the r.h.s. of (21.12)} = |\mathbf{S}_3||\mathbf{R}_2||\mathbf{F}_2'(\mathbf{I} - \mathbf{H}_2'\mathbf{X}\mathbf{P}_{\mathbf{C}_3'}\mathbf{R}_2^{-1}\mathbf{P}_{\mathbf{C}_3'}\mathbf{X}'\mathbf{H}_2)\mathbf{F}_2|. \quad (21.15)$$

Furthermore, define

$$\mathbf{U}_2 = \mathbf{I} - \mathbf{H}_2'\mathbf{X}\mathbf{P}_{\mathbf{C}_3'}\mathbf{R}_2^{-1}\mathbf{P}_{\mathbf{C}_3'}\mathbf{X}'\mathbf{H}_2, \qquad (p - r(\mathbf{A} : \boldsymbol{\Theta})) \times (p - r(\mathbf{A} : \boldsymbol{\Theta})),$$

with $\mathbf{U}_2^{-1} = \mathbf{I} + \mathbf{H}_2'\mathbf{X}\mathbf{P}_{\mathbf{C}_3'}\mathbf{X}'\mathbf{H}_2$. Thus, $\mathbf{U}_2$ is positive definite. The assumption $\mathscr{C}(\mathbf{A}) \cap \mathscr{C}(\boldsymbol{\Theta}) = \{\mathbf{0}\}$ used in (21.1) implies that $p - r(\mathbf{A} : \boldsymbol{\Theta}) = p - r(\mathbf{A}) - r_1$. Then,

$$\text{the r.h.s. of (21.15)} = |\mathbf{S}_3||\mathbf{R}_2||\mathbf{F}_2'\mathbf{U}_2\mathbf{F}_2| \geq |\mathbf{S}_3||\mathbf{R}_2| \prod_{i=1}^{r_2} \lambda_{p-r(\mathbf{A})-r_1-r_2+i},$$

where $\{\lambda_{p-r(\mathbf{A})-r_1-r_2+i}\}$ are eigen-values of $\mathbf{U}_2$. Moreover, let

$$\widetilde{\mathbf{F}}\mathbf{F}_2 = (\mathbf{w}_1, \ldots, \mathbf{w}_{r_2})$$

be the matrix of eigen-vectors of $\mathbf{U}_2$ corresponding to $\{\lambda_{p-r(\mathbf{A})-r_1-r_2+i}\}$, $i \in \{1, \ldots, r_2\}$. Thus, an estimated $\boldsymbol{\Psi}_1$ must satisfy

$$\mathbf{H}_2'\boldsymbol{\Psi}_1(\boldsymbol{\Psi}_1'\mathbf{H}_2\mathbf{H}_2'\boldsymbol{\Psi}_1)^{-1/2} = \widetilde{\mathbf{F}}\mathbf{F}_2$$

Since $\widetilde{\mathbf{F}}\mathbf{F}_2'\widetilde{\mathbf{F}}\mathbf{F}_2 = \mathbf{I}$

$$\widehat{\boldsymbol{\Psi}}_1 = \mathbf{H}_2(\mathbf{H}_2'\mathbf{H}_2)^{-1}\widetilde{\mathbf{F}}\mathbf{F}_2 \quad (21.16)$$

is an estimator.

This result means that we have estimated both $\widehat{\boldsymbol{\Theta}}\mathbf{C}_2$ and $\widehat{\boldsymbol{\Psi}\boldsymbol{\Theta}}\mathbf{C}_3$, and therefore, an explicit estimator of $\mathbf{B}$ can be presented, which was the main purpose of this article. We do not present estimators of $\boldsymbol{\Theta}_2$ and $\boldsymbol{\Psi}_2$, since they are of no real interest, although they can be obtained from (21.15) and (21.13), respectively.

Proposition 1 *For the model presented in Sect. 21.2 in (21.1), let* $\widehat{\boldsymbol{\Theta}}\mathbf{C}_2$ *and* $\widehat{\boldsymbol{\Psi}\boldsymbol{\Theta}}\mathbf{C}_3$ *be given by (21.15) and (21.13), respectively, where for the last relation* $\widehat{\boldsymbol{\Psi}}_1$*, presented in (21.16), has been inserted. The following estimators are proposed:*

(*i*) $\widehat{\boldsymbol{\Theta}}_1$ *is given in (21.11);*
(*ii*) $\widehat{\boldsymbol{\Psi}}_1$ *is given in (21.16);*

(*iii*) *If* $r(\mathbf{A}) = q$ *and* $r(\mathbf{C}_1) = k$ *then*

$$\begin{aligned}\widehat{\mathbf{B}} &= (\mathbf{A}'\mathbf{S}_1^{-1}\mathbf{A})^{-1}\mathbf{A}'\mathbf{S}_1^{-1} \\ &\quad \times (\mathbf{X}\mathbf{C}_1'(\mathbf{C}_1\mathbf{C}_1')^{-1} - \widehat{\boldsymbol{\Theta}}\mathbf{C}_2\mathbf{C}_1'(\mathbf{C}_1\mathbf{C}_1')^{-1} - \widehat{\boldsymbol{\Psi}\boldsymbol{\Theta}}\mathbf{C}_3\mathbf{C}_1'(\mathbf{C}_1\mathbf{C}_1')^{-1}),\end{aligned}$$

where $\mathbf{S}_1 = \mathbf{X}(\mathbf{I} - \mathbf{P}_{\mathbf{C}_1'})\mathbf{X}'$;

(*iv*) $\mathbf{A}\widehat{\mathbf{B}}\mathbf{C}_1 = \mathbf{A}(\mathbf{A}'\mathbf{S}_1^{-1}\mathbf{A})^{-}\mathbf{A}'\mathbf{S}_1^{-1}(\mathbf{X}\mathbf{C}_1'(\mathbf{C}_1\mathbf{C}_1')^{-}\mathbf{C}_1 - \widehat{\boldsymbol{\Theta}}\mathbf{C}_2 - \widehat{\boldsymbol{\Psi}\boldsymbol{\Theta}}\mathbf{C}_3)$;

(*v*) $n\widehat{\boldsymbol{\Sigma}} = (\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}}\mathbf{C}_1 - \widehat{\boldsymbol{\Theta}}\mathbf{C}_2 - \widehat{\boldsymbol{\Psi}\boldsymbol{\Theta}}\mathbf{C}_3)(\mathbf{X} - \mathbf{A}\widehat{\mathbf{B}}\mathbf{C}_1 - \widehat{\boldsymbol{\Theta}}\mathbf{C}_2 - \widehat{\boldsymbol{\Psi}\boldsymbol{\Theta}}\mathbf{C}_3)'$.

21.4 Discussion

The model is overparameterized, which implies estimability problems (parameter identiliability problems). In fact, estimability in complex statistical models has become an important topic in the era of analysing large data-sets. Regularization in loss functions is one tool that nowadays is often applied. A different approach that constitutes the main idea of this work is to introduce rank restrictions on mean parameters to model the effects of latent variables, which are thought to govern a large set of measurable variables. Moreover, we link the latent mean variables with an extended Bilinear regression model, which yields a new class of models. An explicit estimator of the latent variable effect is derived. In the future, based on this estimate, the aim is to study statistical properties, including the interpretation, of estimators and estimates and to study different types of model validation procedures.

Acknowledgements Chengcheng Hao and Feng Li are supported by the National Natural Science Foundation of China (no. 11601319 and no. 11501587, respectively). Feng Li is also supported by Beijing Universities Advanced Disciplines Initiative (no. 6JJ2019163). Dietrich von Rosen is supported by the Swedish Research Council (2017-03003).

References

1. Anderson, T.W.: Estimating linear restrictions on regression coefficients for multivariate normal distributions. Ann. Math. Statist **22**, 327–351 (1951)
2. Chinchilli, V.M., Elswick, R.K.: A mixture of the MANOVA and GMANOVA models. Comm. Statist. Theory Methods **14**, 3075–3089 (1985)
3. Johansen, S.: Estimation and hypothesis testing of cointegration vectors in Gaussian vector autoregressive models. Econometrica **59**, 1551–1580 (1991)
4. Kollo, T., von Rosen, D.: Advanced Multivariate Statistics with Matrices. Springer, New York (2005)
5. Potthoff, R.F., Roy, S.N.: A generalized multivariate analysis of variance model useful especially for growth curve problems. Biometrika **51**, 313–326 (1964)
6. Rao, C.R.: Tests of significance in multivariate analysis. Biometrika **35**, 58–79 (1948)
7. Rao, C.R.: Separation theorems for singular values of matrices and their applications in multivariate analysis. J. Multivar. Anal. **9**, 362–377 (1979)

8. Reinsel, G.C., Velu, R.P.: Multivariate Reduced-Rank Regression. Springer, New York (1998)
9. von Rosen, D.: Bilinear Regression Analysis: An Introduction. Springer, New York (2018)
10. von Rosen, T., von Rosen, D.: On estimation in some reduced rank extended growth curve models. Math. Methods Statist. **26**, 299–310 (2017)
11. Srivastava, M.S., Khatri, C.G.: An Introduction to Multivariate Statistics. North-Holland, New York (1979)

Chapter 22
Simultaneous Multiple Change Points Estimation in Generalized Linear Models

Xiaoying Sun and Yuehua Wu

Abstract In this paper, the problem of multiple change points estimation is considered for generalized linear models, in which both the number of change-points and their locations are unknown. The proposed method is to first partition the data sequence into segments to construct a new design matrix, secondly convert the multiple change points estimation problem into a variable selection problem, and then apply a regularized model selection technique and obtain the regression coefficient estimation. The consistency of the estimator is established regardless if there is a change point in which the number of coefficients can diverge as the sample size goes to infinity. An algorithm is provided to estimate the multiple change points. Simulation studies are conducted for the logistic and log-linear models. A real data application is also presented.

22.1 Introduction

Change point analysis is the process of detecting distributional changes within time-ordered observations [13]. Applications can be found in many research areas including climate studies, medical and health sciences, financial econometrics and risk management. For instance, change point analysis is used to examine the North Atlantic tropical cyclone record for statistical discontinuities (change points) [16], confirm the effect of the seat belt legislation on the monthly deaths and serious injuries, detect speech signals [3], and estimate changes in the 1982 Urakawa-Oki earthquake records [10].

Page [14, 15] first introduced the undocumented change point problem. Since then, change point problems have been intensively studied in the literature. The change point problems often considered in the literature can roughly be categorized into two

X. Sun · Y. Wu (✉)
York University, Toronto, Canada
e-mail: wuyh@mathstat.yorku.ca

X. Sun
e-mail: sunying@mathstat.yorku.ca

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_22

classes. One class is the change point detection in the distributions of a time-ordered sequence of independent observations. Examples include a nonparametric approach for detecting multiple change points in the distributions of a multivariate sequence of independent observations [13]; construction of a nonparametric test statistic and a nonparametric estimator respectively to detect and estimate a change point in the distributions of an independent univariate sequence [8, 17]; use of a test statistic to detect a single change point in a categorical data sequence [16]. The other class of change point problems is to detect or estimate all the locations in a data sequence such that before and after each of them the data sequence follows different models. The single change point detection in the linear regression models can be found in Csörgö and Horváth [2] among others. For linear regression models, a fast algorithm was proposed in Jin et al. [11] to simultaneously estimate multiple change points in linear regression models. For nonstationary time series models, Davis et al. [3] and Jin et al. [10] studied the multiple structural break estimation and variable selection problem. For cumulative logit models, Lu and Wang [12] developed a likelihood ratio test for detecting a sudden change in parameters for a multinomial sequence. For generalized linear models (GLMs), Antoch et al. [1] proposed a statistic to test if there is a structural change. More examples can be found in literature. However, the literature on the multiple change points estimation in the GLM is relatively thin.

In this paper, we focus on the problem of multiple change points estimation in GLMs in which both number of change points and their locations are unknown. In light of Jin et al. [10], we propose a simultaneous multiple change points estimation method which first partitions the data sequence into several segments to construct a new design matrix, secondly convert the multiple change points estimation problem into a variable selection problem, and then estimate the regression coefficients by maximizing a penalized likelihood function. The consistency of the coefficient estimator is established in which the number of coefficients can diverge as the sample size goes to infinity. The nonzero coefficient estimates provide the information about which segments potentially contain a change point. An algorithm is provided to estimate the change point in each possible segment. Note that in this algorithm, the test statistic proposed in Antoch et al. [1] is used to test if there exists a change point in each possible segment.

The rest of this article is organized as follows. In Sect. 22.2, we present a GLM with multiple change points and describe our change point detection methodology, whose theoretic justification is also provided. In Sect. 22.3, an algorithm is built to implement the method given in Sect. 22.2. Simulation studies and a real data application are presented in Sects. 22.4 and 22.5 respectively. The paper is concluded in Sect. 22.6. The test proposed by Antoch et al. [1], and the proof of the theorem are respectively given in the appendix.

Throughout this paper, A^T denotes the transpose of a matrix A. v^T, v_j and $\|v\|$ denote the transpose, j^{th} component and the L_2 norm of a vector v, respectively. For $v = (v_1, v_2, \ldots, v_p)^T$ being a $p \times 1$ vector, $A = (a_{ij}) = (a_1, \ldots, a_p)$ being a $q \times p$ matrix with a_{ij}'s as its elements and a_j's as its column vectors, and $\mathscr{B} = \{i_1, i_2, \ldots, i_k\}$, $1 \le i_1 \le \ldots \le i_k \le p$, being an index set with its size denoted by $|\mathscr{B}|$, write $v_{[\mathscr{B}]} = (v_{i_1}, \ldots, v_{i_k})^T$ and $A_{[\mathscr{B}]} = (a_{i_1}, \ldots, a_{i_k})$. Let $I_S(t)$ be the indicator

function such that $I_S(t) = 1$ if $t \in S$ and $I_S(t) = 0$ otherwise. Define $a_+ = a$ if $a > 0$ and $a_+ = 0$ otherwise. Denote the inverse function of $f(x)$ as $f^{-1}(x)$. Let $f'(x)$ and $f''(x)$ denote the first and second order derivatives of a univariate function, $f(x)$, with respect to the scalar x. If g is a univariate function of a vector v, let $\partial g(v)/\partial v$ and $\partial^2 g(v)/(\partial v \partial v^T)$ denote the first and second order derivatives with respect to the vector v. Define $\lfloor x \rfloor$ and $\lceil x \rceil$ respectively as the largest integer smaller than or equal to x and the smallest integer larger than or equal to x.

22.2 Simultaneous Multiple Change Points Detection

22.2.1 The GLM with Multiple Change Points

Let $(y_{n1}, x_{n1}), (y_{n2}, x_{n2}), \cdots, (y_{nn}, x_{nn})$ be a double-indexed series of random samples where y_{nt} is a scalar response and $x_{nt} = (x_{nt1}, x_{nt2}, \cdots, x_{ntp})^T$ is a vector of covariates for all $t = 1, 2, \cdots, n$. Suppose that for every n and given x_{nt}, Y_{nt} has a distribution in the exponential family, taking the form

$$f_{nt}(y_{nt}|x_{nt}) = \exp\left\{\frac{y_{nt}\theta(x_{nt}) - b(\theta(x_{nt}))}{a(\phi)} + c(y_{nt}, \phi)\right\}$$

for some specific function $a(\cdot), b(\cdot)$ and $c(\cdot)$. Then the expectation and variance of Y_{nt} given x_{nt} are respectively $\mu_{nt} = E(Y_{nt}|x_{nt}) = b'(\theta(x_{nt}))$ and $\sigma_{nt}^2 = Var(Y_{nt}|x_{nt}) = a(\phi)b''(\theta(x_{nt}))$.

The GLM is formulated as

$$g(\mu_{nt}) = \sum_{j=1}^{p} \beta_j x_{ntj} = x_{nt}^T \beta$$

where $\beta = (\beta_1, \beta_2, \cdots, \beta_p)^T$ is the vector of parameters, and $g(\cdot)$ is a proper link function. In this paper, we consider the canonical link, i.e., $g(\mu_{nt}) = (db/d\theta)^{-1}(\mu_{nt})$, and hence $\theta(x_{nt}) = x_{nt}^T \beta$.

Denote all the change points as $\{l_{n,1}, l_{n,2}, \cdots, l_{n,s}\}$ satisfying that $0 = l_0 < l_{n,1} < l_{n,2} < \cdots < l_{n,s} < l_{n,s+1} = n$, where s is the total number of change points. Consider the following GLM with multiple change points formulated as

$$g(\mu_{nt}) = x_{nt}^T \beta_i, \quad l_{n,i-1} < t \le l_{n,i}, \quad i = 1, 2, \cdots, s+1, \quad t = 1, 2, \cdots, n, \quad (22.1)$$

where $\beta_i = (\beta_{i1}, \cdots, \beta_{ip})^T$ is the parameter vector associated with the i^{th} segment $\{l_{n,i-1}, \ldots, l_{n,i}\}$. The objective is to estimate the total number of change points, s, and their locations, $l_{n,1}, l_{n,2}, \cdots, l_{n,s}$.

The double subscripts in model (22.1) are to emphasize the dependence on the sample size n. Throughout this paper $\tau_i \in (0, 1)$ is defined such that $l_{n,i} = \lfloor \tau_i n \rfloor$ for

$i = 1, 2 \cdots, s$. Set $\tau_0 = 0$ and $\tau_{s+1} = 1$ for convenience. In the rest of this paper, the subscript n is suppressed if there is no confusion.

22.2.2 The Method

In order to detect all the change points in model (22.1), the proposed method is to convert the multiple change points detection problem into a model selection problem by partitioning the data sequence and rewriting model (22.1) into model (22.2), and then utilize the regularized model selection techniques to estimate the total number of change points, s and the change points l_i's simultaneously. The procedure is described as following.

1. Partition the data sequence into q_n segments, $\mathscr{Q}_1 = \{1, 2, \cdots, n - (q_n - 1)m\}$ as the first segment with length $n - (q_n - 1)m$ satisfying that $m \leq n - (q_n - 1)m \leq d_0 m$ for some $d_0 \geq 1$ and $\mathscr{Q}_k = \{n - (q_n - k + 1)m + 1, \cdots, n - (q_n - k)m\}$ as the k^{th} segment with length m for $k = 2, 3, \cdots, q_n$. Then there exist $n_1 < n_2 < \cdots < n_s$ such that $l_i \in \mathscr{Q}_{n_i}$ for $i = 1, 2, \cdots, s$.
2. Rewrite model (22.1) in order to incorporate the partition yields the following model

$$g(\mu_t) = x_t^T [\ \beta_1 + \sum_{k=2}^{q_n} \delta_k I_{\{n-(q_n-k+1)m+1, \ldots,\ n\}}(t)\ ] - v_t, \tag{22.2}$$

where

$$\delta_k = \begin{cases} \beta_{i+1} - \beta_i, & \text{for} k = n_i, \quad i = 1, 2, \ldots, s, \\ 0, & \text{otherwise}, \end{cases}$$

and

$$v_t = \begin{cases} x_t^T \delta_k, & \text{for} k = n_i, \quad t \in \{n - (q_n - k + 1)m + 1, \ldots, l_i\}, \\ 0, & \text{otherwise}, \end{cases}$$

$t = 1, 2, \cdots, n$. For the sake of convenience, denote $\varsigma_i = n - (q_n - n_i + 1)m + 1$.

3. Denote $g(\mu) = (g(\mu_1),\ g(\mu_2),\ \ldots,\ g(\mu_n))^T$. Let $\mathscr{A} = \cup_{i=0}^{s} \mathscr{B}_i$, where $\mathscr{B}_i = \{(n_i - 1)p + 1, \ldots, n_i p\}, \quad i = 1, \ldots, s, \mathscr{B}_0 = \{1, \ldots, p\}$ and $\mathscr{A}^c = \{1, \ldots, pq_n\} \backslash \mathscr{A}$. Put $\gamma = (\beta_1^T,\ \delta_2^T,\ \cdots,\ \delta_{q_n}^T)^T = (\gamma_1,\ \gamma_2,\ \cdots,\ \gamma_{pq_n})^T$, where $\gamma_{[\mathscr{A}^c]} = 0$. Now we write model (22.2) in the following matrix form

$$g = Z\gamma - W\gamma. \tag{22.3}$$

Here,

$$Z = [z_1,\ z_2,\ \cdots,\ z_n]^T\ =\ [\tilde{z}_1,\ \tilde{z}_2,\ \cdots,\ \tilde{z}_{pq_n}]$$
$$= \begin{pmatrix} Z^{(1)} & 0 & 0 & \cdots & 0 \\ Z^{(2)} & Z^{(2)} & 0 & \cdots & 0 \\ & & \cdots & & \\ Z^{(q_n)} & Z^{(q_n)} & Z^{(q_n)} & Z^{(q_n)} & Z^{(q_n)} \end{pmatrix}_{n\times(pq_n)},$$

$Z^{(1)} = (x_1,\ x_2,\ \cdots,\ x_{n-(q_n-1)m})^T$ of dimension $(n-(q_n-1)m)\ \times\ p$, $Z^{(2)} = (x_{n-(q_n-1)m+1},\ x_{n-(q_n-1)m+2},\ \cdots,\ x_{n-(q_n-2)m})^T$ of dimension $m\ \times\ p$, $\cdots$, $Z^{(q_n)} = (x_{n-m+1},\ x_{n-m+2},\ \cdots,\ x_n)^T$ of dimension $m\ \times\ p, z_t, t = 1,\ 2,\ \ldots,\ n$ are row vectors of Z, $\tilde{z}_j,\ j\ =\ 1,\ 2,\ \cdots,\ pq_n$ are column vectors of Z, and $W_{n\times(pq_n)} = (w_1, w_2, \ldots, w_n)^T$ with $w_t = 0$ for $t \notin \{n-(q_n-n_i+1)m+1, \ldots, l_i\}$, otherwise $(w_t)_{[\mathscr{B}_i]} = x_t$ and $(w_t)_{[\mathscr{B}_i^c]} = 0$, where $t\ =\ 1,\ 2,\ \ldots,\ n$ and $i = 1, 2, \ldots, s$.

Then the log-likelihood function for model (22.3) is

$$\mathscr{L}(\gamma) = \sum_{t=1}^{n}\left[\frac{y_t(z_t^T\gamma - w_t^T\gamma) - b(z_t^T\gamma - w_t^T\gamma)}{a(\phi)} + c(y_t,\phi)\right].$$

4. Denote $Q(\gamma) = \mathscr{L}_1(\gamma) - n\sum_{j=1}^{pq_n} p_{\lambda_n}(|\gamma_j|)$, where $\mathscr{L}_1(\gamma) = \sum_{t=1}^{n}(\frac{y_t(z_t^T\gamma)-b(z_t^T\gamma)}{a(\phi)} + c(y_t,\phi))$, obtained by letting $w_t \equiv \mathbf{0}$ in $\mathscr{L}(\gamma)$, for $t\ =\ 1,\ 2,\ \ldots,\ n$. We propose to estimate γ in model (22.3) by maximizing the following penalized log-likelihood function

$$\widehat{\gamma} = \arg\max_{\gamma} Q(\gamma) = \arg\max_{\gamma}\left\{\mathscr{L}_1(\gamma) - n\sum_{j=1}^{pq_n} p_{\lambda_n,d}(|\gamma_j|)\right\}, \qquad (22.4)$$

where $\lambda_n > 0, d > 0$, and the penalty function $p_{\lambda_n,d}(\theta)$ is symmetric about $\theta = 0$ and satisfies the following assumptions: $p_{\lambda_n,d}(0) = 0$, $p'_{\lambda_n,d}(\theta) = 0$ if $\theta > \lambda_n d$ and $p'_{\lambda_n,d}(0) = \lambda_n$. Here are two penalty functions among others that meet these assumptions. One is the SCAD penalty function defined in Fan and Li [4] satisfying that $p_{\lambda_n,d}(0) = 0$ and $p'_{\lambda_n,d}(\theta) = \lambda_n\{I_{(0,\lambda_n]}(\theta) + \frac{(d\lambda_n-\theta)_+}{(d-1)\lambda_n}I_{(\lambda_n,\infty)}(\theta)\}$. The other is the MCP penalty defined in Zhang [18] satisfying that $p_{\lambda_n,d}(\theta) = (\lambda_n\theta - \frac{\theta^2}{2d})I_{(0,d\lambda_n]}(\theta) + \frac{1}{2}d\lambda_n^2 I_{(d\lambda_n,\infty)}(\theta)$. In this paper, we use these two penalty functions for illustration purpose. Other penalty functions may also be used to derive the regression coefficient estimator $\widehat{\gamma}_n$.

22.2.3 The Consistency of the Proposed Estimator

To study the asymptotic properties of $\widehat{\gamma}_n$, we assume that there is an underlying true model with true change points $l_{n,i}^* = \lfloor n\tau_i^* \rfloor,\ i = 1, 2, \cdots, s$ and there exist true values of γ_n: $\gamma_n^0 = (\gamma_{n1}^0, \ldots, \gamma_{n,pq_n}^0)^T$ with $\gamma_{n[\mathscr{A}_n^c]}^0 = 0$. Note that the dimension of γ_n

goes to ∞ as $n \to \infty$. To prove the consistency of the estimator $\widehat{\gamma}_n$, we employ the techniques developed in Fan and Peng [6] which showed the asymptotic properties of the maximum nonconcave penalized likelihood estimator with a diverging number of parameters. The following assumptions are made for the technical proof. The first four assumptions are imposed on both likelihood and penalty terms. The last one is put on the term involving w.

Assumption 22.1 $\liminf_{n\to\infty} \liminf_{\gamma\to 0^+} p'_{\lambda_n}(\gamma)/\lambda_n > 0$.

Assumption 22.2 $\lambda_n \to 0$, $\sqrt{n/q_n}\lambda_n \to \infty$ as $n \to \infty$.

Assumption 22.3 $\min_{j\in\mathscr{A}} \{|\gamma_{nj}^0|/\lambda_n\} \to \infty$ as $n \to \infty$.

Assumption 22.4 For every n and i, $\{(Y_t, x_t), l_{i-1} < t \le l_i\}$ are independent and identically distributed with probability density $f_{n,i}(y_{l_i}, x_{l_i}, \beta_i)$, which has a common support, and the model is identifiable. Furthermore, they satisfy the following three regularity conditions.

(1) The first and second derivatives of the likelihood function satisfy the joint equations

$$E_{\beta_i}\left\{\frac{\partial \log f_{n,i}(y_{l_i}, x_{l_i}, \beta_i)}{\partial \beta_{ij}}\right\} = 0,$$

and

$$E_{\beta_i}\left\{\frac{\partial \log f_{n,i}(y_{l_i}, x_{l_i}, \beta_i)}{\partial \beta_{ij}} \frac{\partial \log f_{n,i}(y_{l_i}, x_{l_i}, \beta_i)}{\partial \beta_{ik}}\right\} = -E_{\beta_i}\left\{\frac{\partial^2 \log f_{n,i}(y_{l_i}, x_{l_i}, \beta_i)}{\partial \beta_{ij}\partial \beta_{ik}}\right\},$$

for $j, k = 1, 2, \ldots, p$.

(2) The Fisher information matrix

$$I(\beta_i) = E_{\beta_i}\left[\left\{\frac{\partial \log f_{n,i}(y_{l_i}, x_{l_i}, \beta_i)}{\partial \beta_i}\right\}\left\{\frac{\partial \log f_{n,i}(y_{l_i}, x_{l_i}, \beta_i)}{\partial \beta_i}\right\}^T\right]$$

satisfies conditions $0 < C_1 < e_{\min}\{I(\beta_i)\} \le e_{\max}\{I(\beta_i)\} < C_2 < \infty$ for all n with $e_{\min}\{I(\beta_i)\}$ and $e_{\max}\{I(\beta_i)\}$ denoting the minimum and maximum eigenvalues of $I(\beta_i)$ respectively. For $j, k = 1, 2, \ldots, p$,

$$E_{\beta_i}\left\{\frac{\partial \log f_{n,i}(y_{l_i}, x_{l_i}, \beta_i)}{\partial \beta_{ij}} \frac{\partial \log f_{n,i}(y_{l_i}, x_{l_i}, \beta_i)}{\partial \beta_{ik}}\right\}^2 < C_3 < \infty$$

and

$$E_{\beta_i}\left\{\frac{\partial^2 \log f_{n,i}(y_{l_i}, x_{l_i}, \beta_i)}{\partial\beta_{ij}\partial\beta_{ik}}\right\}^2 < C_4 < \infty.$$

(3) There is a large enough open subset ω_i of $\Omega \in R^p$ which contains the true parameter β_i, such that for almost all (Y_t, x_t), the density admits all third derivatives $\partial f_{n,i}(y_{l_i}, x_{l_i}, \beta_i)/\partial\beta_{ij}\partial\beta_{ik}\partial\beta_{il}$ for all $\beta_i \in \omega_i$. Furthermore, there are functions M_{njkl} such that

$$\left|\frac{\partial \log f_{n,i}(y_{l_i}, x_{l_i}, \beta_i)}{\partial\beta_{ij}\partial\beta_{ik}\partial\beta_{il}}\right| \le M_{njkl}(y_{l_i}, x_{l_i})$$

for all $\beta_i \in \omega_i$, and

$$E_{\beta_i}\left\{M^2_{njkl}(y_{l_i}, x_{l_i})\right\} < C_5 < \infty$$

for all p, n, j, k, l.

These regularity conditions correspond to Assumptions (E)–(G) in Fan and Peng [6].

Assumption 22.5 Assume that $\min\{\tau_i^* - \tau_{i-1}^*, \quad i = 1, 2, \cdots, s+1\} > \iota > 0$ where ι is a constant. Also assume that $q_n = O(n^{\frac{1}{6}})$ and $l_{n,i}^* - \varsigma_i = O(\sqrt{nq_n})$ where $\varsigma_i = n - (q_n - n_i + 1)m + 1$.

To this end, we state the theorem as follows and its proof is given in Appendix B.

Theorem 22.1 *If Assumptions 1–5 hold, there exists a local maximizer $\widehat{\gamma}_n$ to $Q(\gamma_n)$ and $\|\widehat{\gamma}_n - \gamma_n^0\| = O_p(\sqrt{q_n/n})$, where $\widehat{\gamma}_n$ is the SCAD estimator. Furthermore, we have* $\lim_{n\to\infty} P(\widehat{\gamma}_{n[\mathscr{A}_n^c]} = 0) = 1$.

Let $\widehat{\mathscr{A}} = \{j : \widehat{\gamma}_j \neq 0\}$. Then the total number of change points is estimated by the size of the set $\{\lceil j/p \rceil, j \in \widehat{\mathscr{A}}\}$ which is denoted as $\widehat{s}$. Theorem 22.1 implies the consistency of $\widehat{s}$ to s. It also provides the information that the $\widehat{k}_i^{th}$ segment contains a change for each $\widehat{k}_i \in \{\lceil j/p \rceil, j \in \widehat{\mathscr{A}}\}$, $j = 1, \ldots, \widehat{s}$.

22.3 An Algorithm

As showed in the previous section, $\widehat{\gamma}_n$ provides the information about which segments potentially contain a change point. Thus we are able to present an algorithm in this section to find out if a possible segment contains a change point and to locate it if it does exist. The algorithm consists of the following steps.

Step 1. First, we test if there exists a change point in the sequence by the test proposed in Antoch et al. [1]. The details are given in Appendix A.

- If there is no change point, set $\tilde{s} = 0$ and go to Step 5.
- Otherwise, estimate the change point by the estimate in Appendix A and denote it by $\widehat{l}$. Then set $\tilde{s} = 1$.

Step 2. Compute the estimate $\widehat{\gamma}$ defined in (22.4) by the R Package `SIS` [5] or `cvplogistic` [9].

Step 3. Let $\widehat{s}$ record the number of change point estimates, $\widehat{\mathbf{k}} = \{\widehat{k}_1, \widehat{k}_2, \ldots, \widehat{k}_{\widehat{s}}\}$ be a vector containing the change point estimates. Set $\widehat{s} = 0$.

- If $\widehat{\gamma}_j = 0$ for all $j > p$, go to Step 5.
- Otherwise, set $\tilde{\mathbf{k}} = \{\tilde{k}_1, \tilde{k}_2, \ldots, \tilde{k}_{s^*}\} = \{\lceil \frac{j}{p} \rceil$: for all $j > p$ such that $\widehat{\gamma}_j \neq 0\}$ with $\tilde{k}_1 < \tilde{k}_2 < \ldots < \tilde{k}_{s^*}$ which records the segment number that possibly contains a change point and s^* is the total number of possible change points. Set $l = 1$ where l is from 1 to s^*.

Step 4. Use the test proposed in Antoch et al. [1] to detect a change point in each segment which possibly contains a change point. The details are given in Appendix A. This step is to reduce the overestimation of the number of change points from Step 3 and also can improve the accuracy of change point estimates.

- If $l > s^*$, go to Step 5.
- Otherwise, test $H_0^{(l)}$ that there is no change point in $g(\mu_t) = x_t^T \beta$, $t = n - (q_n - \tilde{k}_l + 2)m + 1, \ldots, \leq n - (q_n - \tilde{k}_l)m$, at the significance level, 5% by Antoch et al. [1].
  - If the test is not significant, set $l = l + 1$, and repeat Step 4.
  - Otherwise, set $\widehat{s} = \widehat{s} + 1$, and $\widehat{k}_{\widehat{s}+1} = \tilde{k}_l$. Then we obtain a change point $\widehat{k}_{\widehat{s}}$ in this segment.

    If $\tilde{k}_{l+1} - \tilde{k}_l = 1$, set $l = l + 1$, and repeat Step 4.
    Otherwise, set $l = l + 2$, and repeat Step 4.

Step 5.

- If $\tilde{s} = 0$, there is no change point.
- If $\tilde{s} = 1$,
  - If $\widehat{s} \leq 1$, there exists one change point and the estimate of this change point, $\widehat{k}$ is given by the estimate, $\widehat{l}$ in Step 1.
  - If $\widehat{s} > 1$, the total number of change points is $\widehat{s}$ and the estimates of these change points are $\{\widehat{k}_1, \widehat{k}_2, \ldots, \widehat{k}_{\widehat{s}}\}$.

In the next two sections, data examples are presented to show the performance of the proposed methodology.

22.4 Simulation Studies

The false alarm rate (Type I error) and the accuracy of the change point estimates derived by the algorithm proposed in Sect. 22.3 are evaluated through Monte Carlo simulations in this section. More specifically, we will calculate the empirical probabilities that the proposed algorithm erroneously detects change points when they actually do not exist. Moreover, we show how frequently the algorithm detects the correct number of change points and how accurately it estimates the change points when they do exist. Two specific generalized linear models, the logistic and the log-linear models, are considered for demonstration purpose.

22.4.1 Two Specific Generalized Linear Models

For the binomial response, $y_t|x_t \sim \text{Binomial}(1, \pi(x_t))$. The density function is

$$f(y_t|x_t) = \pi(x_t)^{y_t}(1-\pi(x_t))^{1-y_t} = \exp\left\{y_t \log\frac{\pi(x_t)}{1-\pi(x_t)} + \log(1-\pi(x_t))\right\}.$$

Then $\theta(x_t) = \log\frac{\pi(x_t)}{1-\pi(x_t)}$, $b(\theta(x_t)) = \log\left(1+e^{\theta(x_t)}\right)$, $\mu_t = b'(\theta(x_t)) = \frac{e^{\theta(x_t)}}{1+e^{\theta(x_t)}}$, and $\sigma_t^2 = b''(\theta(x_t)) = \frac{e^{\theta(x_t)}}{(1+e^{\theta(x_t)})^2}$. It can be seen that the canonical link function is $g(\mu_t) = \log(\frac{\mu_t}{1-\mu_t})$.

For the Poisson response, $y_t|x_t \sim \text{Poisson}(\lambda(x_t))$. The density function is

$$f(y_t|x_t) = \frac{\lambda(x_t)^{y_t}e^{-\lambda(x_t)}}{y_t!} = \exp\{y_t\log\lambda(x_t) - \lambda(x_t) - \log(y_t!)\}.$$

Then $\theta(x_t) = \log\lambda(x_t), b(\theta(x_t)) = e^{\theta(x_t)}, \mu_t = b'(\theta(x_t)) = e^{\theta(x_t)}$, and $\sigma^2 = b''(\theta(x_t)) = e^{\theta(x_t)}$. It can be seen that the canonical link function is $g(\mu_t) = \log(\mu_t)$.

22.4.2 GLMs with No Change Point

To examine the false alarm rate of the proposed algorithm, we consider the following four models, two for the binomial response and the other two for the Poisson response:

$$\mathbb{B}1 : \log\frac{\mu_t}{1-\mu_t} = -0.7; \qquad \mathbb{B}2 : \log\frac{\mu_t}{1-\mu_t} = 12 - 3x_t;$$
$$\mathbb{P}1 : \log(\mu_t) = 2; \qquad \mathbb{P}2 : \log(\mu_t) = 2 - x_t,$$

where $t = 1, \ldots, n$.

All of these four models contain no change point. We first generate x_t from the uniform distribution U(0, 9) for $\mathbb{B}2$ and U(0, 1) for $\mathbb{P}2$. For each model, we generate 1,000 independent observations with length $n = 1,000$. The empirical

probabilities that the proposed algorithm erroneously detects change points in the generated sequences are 0.039 for $\mathbb{B}1$, 0.084 for $\mathbb{B}2$, 0.034 for $\mathbb{P}1$, and 0.044 for $\mathbb{P}2$. This demonstrates that our algorithm has low false alarm rates for all these four models.

22.4.3 GLMs with Multiple Change Points

The performance of the proposed methodology is also evaluated in this subsection through Monte Carlo simulations for GLMs with multiple change points. We will estimate how frequently the methodology detects the correct number of change points and how accurately it estimates the change points when they do exist. We consider the following four models. $\mathbb{B}3-\mathbb{B}4$ are for the binomial response and $\mathbb{P}3-\mathbb{P}4$ are for the Poisson response.

- $\mathbb{B}3 : \log \frac{\mu_t}{1-\mu_t} = -0.73 + 0.14x_t + (2.02 + 1.34x_t)I_{\{513,\ldots,769\}}(t) - (2.15 + 1.57x_t)I_{\{770,\ldots,1000\}}(t)$.
- $\mathbb{B}4 : \log \frac{\mu_t}{1-\mu_t} = 1.58 - 0.79x_t - (2.04 - 0.90x_t)I_{\{1428,\ldots,10000\}}(t) + (2.25 - 0.07x_t)I_{\{3085,\ldots,10000\}}(t) - 2.86I_{\{4503,\ldots,10000\}}(t)$ $+(1.66 - 0.02x_t)I_{\{5913,\ldots,10000\}}(t) - (0.59 + 0.79x_t)I_{\{7422,\ldots,10000\}}(t) + (0.67 + 1.27x_t)I_{\{8804,\ldots,10000\}}(t)$.
- $\mathbb{P}3 : \log(\mu_t) = 0.31 - 0.11x_t + 0.91I_{\{513,\ldots,769\}}(t) - (0.64 - 0.01x_t)$ $I_{\{770,\ldots,1000\}}(t)$.
- $\mathbb{P}4 : \log(\mu_t) = 1.58 - 0.79x_t - (2.04 - 0.90x_t)I_{\{1428,\ldots,10000\}}(t) + (0.95$ $- 0.18x_t)I_{\{3085,\ldots,10000\}}(t) - (1.06 + 0.12x_t)I_{\{4503,\ldots,10000\}}(t) + (0.95 + 0.41x_t)$ $I_{\{5913,\ldots,10000\}}(t) - (0.88 + 0.39x_t)I_{\{7422,\ldots,10000\}}(t) + (0.87 + 0.30x_t)$ $I_{\{8804,\ldots,10000\}}(t)$.

Both $\mathbb{B}3$ and $\mathbb{P}3$ contain two change points located at $t = 512, 769$ respectively. Both $\mathbb{B}4$ and $\mathbb{P}4$ contain 6 change points at $t = 1427, 3084, 4502, 5912, 7421, 8803$, respectively. First, we generate x_t from the uniform distribution U(0, 9) for $\mathbb{B}3-\mathbb{B}4$ and U(0, 1) for $\mathbb{P}3-\mathbb{P}4$, then we generate y_t according to each model for $t = 1, 2, \ldots, n$, with $n = 1,000$ for $\mathbb{B}3$ and $\mathbb{P}3$ and $n = 10,000$ for $\mathbb{B}4$ and $\mathbb{P}4$.

The accuracy of the change point estimates is calculated based on 1000 independent simulations. Let $\widehat{\mathcal{N}}_i^{(\mathbb{M})} = \{\widehat{t}_1^{(\mathbb{M})}, \ldots, \widehat{t}_{\widehat{s}}^{(\mathbb{M}j)}\}$ contain all change points estimated by the proposed methodology in the i^{th} simulation based on model $\mathbb{M}$ being $\mathbb{B}1$, $\mathbb{B}2$, $\mathbb{P}1$, or $\mathbb{P}2$ for $i = 1, 2, \ldots, 1,000$. Denote $\tilde{\epsilon}_{\mathbb{M}} = \{\widehat{\mathcal{N}}_i^{(\mathbb{M})} : |\widehat{\mathcal{N}}_i^{(\mathbb{M})}| = 1,\ i = 1, 2, \ldots, 1,000\}$ for $\mathbb{M}$ being $\mathbb{B}1$, $\mathbb{B}2$, $\mathbb{P}1$, or $\mathbb{P}2$. Thus, $|\tilde{\epsilon}_{\mathbb{M}}|$ denotes the number of simulations from model $\mathbb{M}$ out of 1,000 in which the number of change points has been correctly detected. Let $\text{Acc}(l, r) = |\{\widehat{k}_i : |\widehat{k}_i - l| \le r,\ i = 1, \ldots, 1000\}|$ with $r = 10$ or 15 denote the number of simulations out of 1,000 repetitions in which the change point estimate $\widehat{k}_i$ falls into the interval of length $2r$ centered at the true change point l, for $i = 1, \ldots, 1000$. The simulation results are reported in Table 22.1 for $\mathbb{B}3-\mathbb{B}4$ and Table 22.2 for $\mathbb{P}3-\mathbb{P}4$. From both tables, it can be seen that our

Table 22.1 Simulation results based on 1,000 simulations for $\mathbb{B}3$ and $\mathbb{B}4$

Model $\mathbb{M}$	$\lvert\tilde{\epsilon}_{\mathbb{M}}\rvert$							
	SCAD	MCP		SCAD	MCP		SCAD	MCP
$\mathbb{B}3$	927	927	Acc(512, 10)	916	971	Acc(512, 15)	931	988
			Acc(769, 10)	994	999	Acc(769, 15)	995	1000
$\mathbb{B}4$	824	723	Acc(1427, 10)	914	915	Acc(1427, 15)	955	956
			Acc(3084, 10)	882	884	Acc(3084, 15)	933	934
			Acc(4502, 10)	986	988	Acc(4502, 15)	992	994
			Acc(5913, 10)	856	850	Acc(5913, 15)	924	920
			Acc(7422, 10)	993	993	Acc(7422, 15)	998	998
			Acc(8804, 10)	957	972	Acc(8804, 15)	957	972

Table 22.2 Simulation results based on 1,000 simulations for $\mathbb{P}3$ and $\mathbb{P}4$

Model $\mathbb{M}$	$\lvert\tilde{\epsilon}_{\mathbb{M}}\rvert$		SCAD		SCAD
$\mathbb{P}3$	973	Acc(512, 10)	922	Acc(512, 15)	958
		Acc(769, 10)	885	Acc(769, 15)	942
$\mathbb{P}4$	873	Acc(1427, 10)	995	Acc(1427, 15)	998
		Acc(3084, 10)	965	Acc(3084, 15)	986
		Acc(4502, 10)	990	Acc(4502, 15)	998
		Acc(5913, 10)	997	Acc(5913, 15)	1000
		Acc(7422, 10)	982	Acc(7422, 15)	998
		Acc(8804, 10)	986	Acc(8804, 15)	986

methodology has a high power in detecting the correct number of multiple change points and a high accuracy in estimating them.

22.5 A Real Data Example

In this section, we apply our methodology on the bike sharing data set which contains the hourly counts of rental bikes in years 2011 and 2012 at Washington, D.C., USA. There are three reasons to justify this application. Firstly, the hourly count of rental bikes can be assumed to follow a Poisson distribution which describes such phenomenons. Secondly, the data set has been used in Fanaee-T and Gama [7] for event labeling which is a process of marking unusual data points as events. Their results show that there are many events marked in the hourly counts of rental bikes, which implies the existence of multiple change points in the mean hourly counts of rental bikes. Lastly, there are other variables such as hourly temperature and hourly humidity in the data set which may provide some justifications of the changes.

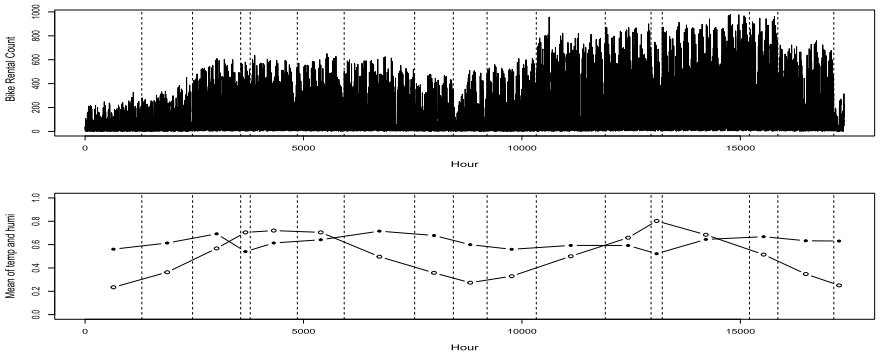


Fig. 22.1 The time series plot of the hourly rental bike counts together with the change points (upper panel) and the mean of hourly temperature and hourly humidity observations within each time interval separated by the change points (lower panel)

The time series of hourly counts including 17,379 h is plotted in Fig. 22.1 (upper panel). There are 16 change points in the series detected by our methodology which are displayed by the vertical lines in Fig. 22.1 (upper panel). Hence the whole time period is divided into 17 intervals by these 16 change points. The means of both hourly temperature and hourly humidity observations within each time interval separated by the change points are also plotted in Fig. 22.1 (lower panel). From this figure, it can be seen that for most of the time intervals, the changes in the means of the hourly counts for rental bikes conform with the changes in the means of the hourly temperatures within each time interval. However, for only two time intervals, the 4^{th} and 13^{th} intervals, the count of rental bikes drops while the mean of hourly temperatures increases. We suspect that, in these two time intervals, the increases of the mean of hourly temperatures and the drops of the mean of hourly humidities together would have caused the drops in the rental counts.

22.6 Discussion

To our best knowledge, this is the first paper of simultaneously multiple change points detection in GLM. In this paper, the regularized model selection technique has been used to solve the multiple change points estimation problem. The consistency of the estimates has been established when the number of parameters is diverging as the sample size goes to infinity.

Throughout the paper, we assume that the nuisance parameter ϕ is known. For instance, we consider the Binomial response and the Poisson response in the simulation studies. The nuisance parameters ϕ for these two distributions are equal to 1. For other distributions in which ϕ is unknown, one has to estimate ϕ first and then incorporate it into our methodology.

Acknowledgements The work was supported by the Natural Sciences and Engineering Research Council of Canada [RGPIN-2017-05720].

Appendix A: A Single Change Point Detection and Estimation in GLM

Consider the following model

$$g(\mu_t) = \begin{cases} x_t^T\beta, & t = 1, 2, \ldots, l, \\ x_t^T\beta^*, & t = l+1, l+2, \ldots, n. \end{cases}$$

Test $H_0 : \ l = n$ and $H_1 : \ l < n$.

The test statistic proposed in Antoch et al. [1] is summarized as follows. The maximum likelihood estimator $\widehat{\beta}$ of β is defined as the solution of the following system of equations: $\sum_{t=1}^{n}(y_t - g^{-1}(x_t^T\beta))x_{tj} = 0, \ j = 1, 2, \ldots, p$. Then $\widehat{\mu}_t = b'(x_t^T\widehat{\beta})$ and $\widehat{\sigma}^2 = a(\phi)b''(x_t^T\widehat{\beta})$, where ϕ is assumed to be known. Let $\widehat{S}(\tilde{l}) = \sum_{t=1}^{\tilde{l}}(y_t - \widehat{\mu}_t)^T x_t$, $\widehat{F}(\tilde{l}) = \sum_{t=1}^{\tilde{l}}\widehat{\sigma}_t^2 x_t x_t^T$, $\widehat{F}(n) = \sum_{t=1}^{n}\widehat{\sigma}_t^2 x_t x_t^T$, and $\widehat{D}(\tilde{l}) = \widehat{F}(\tilde{l}) - \widehat{F}(\tilde{l})\widehat{F}(n)^{-1}\widehat{F}(\tilde{l})^T$. Assume that there exists k_0 such that $\widehat{D}(\tilde{l})$ is positive definite for all $k_0 < \tilde{l} < n - k_0$. The test statistic proposed in Antoch et al. [1] is $T = \max_{k_0<\tilde{l}<n-k_0}\widehat{S}(\tilde{l})^T\widehat{D}(\tilde{l})^{-1}\widehat{S}(\tilde{l})$. They also showed that under H_0, the limiting distribution of the test statistic is

$$P(T \le 2\log\log n + (p+1)\log\log\log n + 2t - 2log\Gamma(\frac{p+1}{2})) \to \exp\{-2e^{-t}\}.$$

The asymptotic critical value for the test statistic at a given significance level can be obtained from this limiting distribution.

In the case that H_0 is rejected, the estimate of l is given by

$$\widehat{l} = \arg\max_{k_0<\tilde{l}<n-k_0}\widehat{S}(\tilde{l})^T\widehat{D}(\tilde{l})^{-1}\widehat{S}(\tilde{l}).$$

Appendix B: Proof of Theorem 22.1

Consider a ball $\|\gamma_n - \gamma_n^0\| \le M\sqrt{q_n/n}$ for some finite M.

$$\begin{aligned} Q(\gamma_n) &= \mathscr{L}_1(\gamma_n) - n\sum_{j=1}^{pq_n} p_{\lambda_n}(|\gamma_{nj}|) \\ &= \sum_{t=1}^{n}\left(\frac{y_{nt}(z_{nt}^T\gamma_n) - b(z_{nt}^T\gamma_n)}{a(\phi)} + c(y_{nt},\phi)\right) - n\sum_{j=1}^{pq_n} p_{\lambda_n}(|\gamma_{nj}|) \end{aligned}$$

$$= \mathscr{L}(\gamma_n) - n\sum_{j=1}^{pq_n} p_{\lambda_n}(|\gamma_{nj}|) + \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}} \frac{y_t(w_{nt}^T\gamma_n)}{a(\phi)}$$
$$- \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}} \frac{b(z_{nt}^T\gamma_n) - b(z_{nt}^T\gamma_n - w_{nt}^T\gamma_n)}{a(\phi)}$$

where $w_{nt} = 0$ for $t \notin \{n-(q_n-n_i+1)m+1, \ldots, l_{n,i}\}$.

First, we consider $\|\gamma_n - \gamma_n^0\| = M\sqrt{q_n/n}$.

$$\begin{aligned}
&Q(\gamma_n) - Q(\gamma_n^0)\\
&= (\mathscr{L}(\gamma_n) - \mathscr{L}(\gamma_n^0)) - n\sum_{j\in\mathscr{A}_n}(p_{\lambda_n}(|\gamma_{nj}|) - p_{\lambda_n}(|\gamma_{nj}^0|)) - n\sum_{j\in\mathscr{A}_n^c}(p_{\lambda_n}(|\gamma_{nj}|) - p_{\lambda_n}(|\gamma_{nj}^0|))\\
&\quad + \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}} \frac{y_{nt}(w_{nt}^T(\gamma_n - \gamma_n^0))}{a(\phi)} - \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}} \frac{b(z_{nt}^T\gamma_n) - b(z_{nt}^T\gamma_n^0)}{a(\phi)}\\
&\quad + \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}} \frac{b(z_{nt}^T\gamma_n - w_{nt}^T\gamma_n) - b(z_{nt}^T\gamma_n^0 - w_{nt}^T\gamma_n^0)}{a(\phi)}.
\end{aligned}$$

As $p_{\lambda_n}(0) = 0$ and $p_{\lambda_n}(|\gamma_{nj}|) \geq 0$, we have

$$\begin{aligned}
&Q(\gamma_n) - Q(\gamma_n^0)\\
&\leq [\mathscr{L}(\gamma_n) - \mathscr{L}(\gamma_n^0)] - n\sum_{j\in\mathscr{A}_n}[p'_{\lambda_n}(|\gamma_{nj}^0|)\mathrm{sign}(\gamma_{nj}^0)(\gamma_{nj} - \gamma_{nj}^0)\\
&\quad + p''_{\lambda_n}(|\gamma_{nj}^0|)(\gamma_{nj} - \gamma_{nj}^0)^2(1 + o_P(1))]\\
&\quad + \sum_{i=1}^{s}\sum_{t=\varsigma_i}^{l_{n,i}} a(\phi)^{-1}[y_{nt}(w_{nt}^T(\gamma_n - \gamma_n^0)) - \frac{\partial b(z_{nt}^T\gamma_n^*)}{\partial\gamma_n} z_{nt}^T(\gamma_n - \gamma_n^0)\\
&\quad + \frac{\partial b(z_{nt}^T\gamma_n^* - w_{nt}^T\gamma_n^*)}{\partial\gamma_n}(z_{nt}^T - w_{nt}^T)(\gamma_n - \gamma_n^0)]\\
&= A_1 + A_2 + A_3,
\end{aligned}$$

where $\|\gamma_n^* - \gamma_n^0\| \leq M\sqrt{q_n/n}$.

By the Taylor expansion and Assumption 4, $A_1 = \mathscr{L}(\gamma_n) - \mathscr{L}(\gamma_n^0) = -M^2 O_p(q_n)$. By Assumption 2, $p'_{\lambda_n}(|\gamma_{nj}^0|) = p''_{\lambda_n}(|\gamma_{nj}^0|) = 0$, for $j \in \mathscr{A}_n$ and large n. Then $|A_2| = o_p(\sqrt{q_n})$. By Assumption 5, $|A_3| = O_P(\sqrt{nq_n})(M\sqrt{q_n/n}) = O_p(q_n)$. By choosing a sufficiently large M, the first term dominates other terms. Since A_1 is negative, for $\varepsilon > 0$, there exists a large constant M such that

$$P\left\{\sup_{\|\gamma_n - \gamma_n^0\| = M\sqrt{q_n/n}} Q(\gamma_n) < Q(\gamma_n^0)\right\} \geq 1 - \varepsilon.$$

This implies that with probability at least $1-\varepsilon$ there exists a local maximum in the ball $\{\gamma_n : \|\gamma_n - \gamma_n^0\| \le M\sqrt{q_n/n}\}$. Hence, there exists a local maximizer such that $\|\widehat{\gamma}_n - \gamma_n^0\| = O_P(\sqrt{q_n/n})$.

Then we consider for $j \in \mathscr{A}_n^c$, by the standard Taylor expansion of the function $\partial\mathscr{L}(\gamma_n)/\partial\gamma_{nj}$ at γ_n^0, we obtain

$$\begin{aligned}
&\frac{\partial Q(\gamma_n)}{\partial\gamma_{nj}} \\
&= \frac{\partial\mathscr{L}(\gamma_n^0)}{\partial\gamma_{nj}} + \sum_{j'=1}^{pq_n}(\gamma_{nj'} - \gamma_{nj'}^0)\frac{\partial^2\mathscr{L}(\gamma_n^0)}{\partial\gamma_{nj}^2}(1+O_P(1)) - np'_{\lambda_n}(|\gamma_{nj}|)\mathrm{sign}(\gamma_{nj}) \\
&+ O_P(\sqrt{nq_n}) \\
&= O_P(\sqrt{nq_n}) + O_P(\sqrt{nq_n}) - np'_{\lambda_n}(|\gamma_{nj}|)\mathrm{sign}(\gamma_{nj}) + O_P(\sqrt{nq_n}) \\
&= n\lambda_n\left[O_P\left(\frac{\sqrt{q_n/n}}{\lambda_n}\right) - \lambda_n^{-1}p'_{\lambda_n}(|\gamma_{nj}|)\mathrm{sign}(\gamma_{nj})\right]
\end{aligned}$$

by Assumption 1. Since $\sqrt{q_n/n}/\lambda_n \to 0$ by Assumption 22.1, this entails that the sign of $\partial Q(\gamma_n)/\partial\gamma_{nj}$ is determined by the sign of γ_{nj} inside the neighborhood of γ_n^0 with radius $M\sqrt{q_n/n}$ by Assumption 3. That is, $\partial Q(\gamma_n)/\partial\gamma_{nj} > 0$ for $\gamma_{nj} < 0$ and $\partial Q(\gamma_n)/\partial\gamma_{nj} < 0$ for $\gamma_{nj} > 0$. Therefore, for any local maximizer $\widehat{\gamma}_n$ inside this ball, $\widehat{\gamma}_{n\mathscr{A}_n^c} = 0$ with probability tending to one. This completes the proof.

References

1. Antoch, J., Gregoire, G., Jaruvsková, D.: Detection of structural changes in generalized linear models. Stat. Probab. Lett. **69**, 315–332 (2004)
2. Csörgö, M., Horváth, L.: Limit Theorems in Change-point Analysis. Wiley, New York (1997)
3. Davis, R.A., Lee, T.C.M., Rodriguez-Yam, G.A.: Structural break estimation for nonstationary time series models. J. Am. Stat. Assoc. **101**, 223–239 (2006)
4. Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. **96**, 1348–1360 (2001)
5. Fan, J., Feng, Y., Saldana, D.F., Samworth, R., Wu, Y.: SIS: Sure Independence Screening (2010)
6. Fan, J., Peng, H.: Nonconcave penalized likelihood with a diverging number of parameters. Ann. Stat. **32**, 928–961 (2004)
7. Fanaee-T, H., Gama, J.: Event labeling combining ensemble detectors and background knowledge. Prog. Artif. Intell. **2**, 113–127 (2014)
8. Huvsková, M., Meintanis, S.G.: Change point analysis based on empirical characteristic functions. Metrika **63**, 145–168 (2006)
9. Jiang, D., Huang, J.: Majorization minimization by coordinate descent for concave penalized generalized linear models. Stat. Comput. **24**, 871–883 (2014)
10. Jin, B., Shi, X., Wu, Y.: A novel and fast methodology for simultaneous multiple structural break estimation and variable selection for nonstationary time series models. Stat. Comput. **23**, 221–231 (2013)
11. Jin, B., Wu, Y., Shi, X.: Consistent two-stage multiple changepoint detection in linear models. Can. J. Stat. **44**, 161–179 (2016)

12. Lu, Q., Wang, X.L.: An extended cumulative logit model for detecting a shift in frequencies of sky-cloudiness conditions. J. Geophys. Res. **117**, D16210 (2012). https://doi.org/10.1029/2012JD017893
13. Matteson, D.S., James, N.A.: A nonparametric approach for multiple change point analysis of multivariate data. J. Am. Stat. Assoc. **109**, 334–345 (2014)
14. Page, E.S.: Continuous inspection schemes. Biometrika **41**, 100–115 (1954)
15. Page, E.S.: A test for a change in a parameter occurring at an unknown point. Biometrika **42**, 523–527 (1955)
16. Robbins, M.W., Lund, R.B., Gallagher, C.M., Lu, Q.: Changepoints in the North Atlantic tropical cyclone record. J. Am. Stat. Assoc. **106**, 89–99 (2011)
17. Tan, C., Shi, X., Sun, X., Wu, Y.: On nonparametric change point estimator based on empirical characteristic functions. Sci. China Math. **59**, 2463–2484 (2016)
18. Zhang, C.: Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. **38**, 894–942 (2010)

Chapter 23
Data-Based Priors for Bayesian Model Averaging

M. Ai, Y. Huang, and J. Yu

Abstract The uncertainty of models is now becoming one of the most important issues in the process of dealing with practical applications. In order to improve reliability and accuracy of inference, one usually adopts the model averaging method instead of selecting a single final model through a model selection procedure. Under the Bayesian framework, two upper bounds of the risk are derived and the posteriors are obtained by minimizing the bounds with a fixed prior. Then we propose two data-based algorithms to get proper priors for Bayesian model averaging in this paper. Simulations show that by using these priors, smaller mean squared prediction errors can be gotten both in synthetic data and real data studies, especially for the data of poor quality.

23.1 Introduction

It is common in practice that the observed data can be described by different models. A standard procedure to make inference is to choose a best model according to some criteria, such as model predictive ability, model fitting ability or many different information criteria like AIC and BIC. After selection, all the inferences and conclusions are made based on the assumption that the selected model is correct.

However, the drawbacks of this approach exist obviously. The selection of one particular model may lead to riskier decisions since it ignores the model uncertainty.

M. Ai (✉) · Y. Huang
LMAM, School of Mathematical Sciences and Center for Statistical Science,
Peking University, Beijing 100871, China
e-mail: myai@math.pku.edu.cn

Y. Huang
e-mail: huangyimin@pku.edu.cn

J. Yu
School of Mathematics and Statistics,
Beijing Institute of Technology, Beijing 100081, China
e-mail: yujunbeta@bit.edu.cn

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_23

In other words, if we choose a wrong model, the consequence will be disastrous. Moral-Benito already pointed out the concern in [8], "From a pure empirical viewpoint, model uncertainty represents a concern because estimates may well depend on the particular model considered." Therefore, combining multiple models to reduce the model uncertainty is very desirable.

As an alternative strategy, model averaging enables researchers to draw conclusions based on the whole universe of candidate models. In particular, researchers estimate all the candidate models and then compute a weighted average of all the estimates for the coefficient on X. There are two different approaches to model averaging in the literature including Frequentist Model Averaging (FMA) and Bayesian Model Averaging (BMA).

Frequentist approaches focus on improving prediction and use weighted mean of estimates from different models while Bayesian approaches focus on the probability that a model is true and consider priors and posteriors for different models. Ref. [4] suggested to use Bayesian inference to reduce the model uncertainty and pointed out the importance of the fragility of regression analysis to arbitrary decisions about the choice of control variables. Bayesian Model Averaging considers model uncertainty through the posterior distribution. The model posteriors are obtained by Bayes' theorem, and therefore allowing for combined estimation and prediction. Compared with the FMA approaches, there are a huge literature on the use of BMA in statistics.

Influenced by [4], most works were concentrated only on the linear models. Ref. [10] extended to generalized linear models by providing a straightforward approximation. For more details, refer to some landmark reviews such as [2, 8, 15] on BMA. Moreover, Refs. [6, 19] gave good estimators of the risk in linear mixed-effects models. For getting the posterior distribution of the weights, Ref. [17] gave a method called SOIL which can well separate the variables in the true model from the rest under some assumptions. However, they used a default prior for the procedure.

The Bayesian approaches have the advantage of using arbitrary domain knowledge through a proper prior. However, they can't guarantee the upper bound of the decision risk without assuming the truth of the prior. The Probably Approximately Correct (PAC) framework, first formulated by [7], was proposed to deal with this problem. It has been widely developed in recent years. Refs. [5, 11] gave tighter bounds in some specific cases. Ref. [1] provided an extended PAC-Bayes bound for learning the proper priors. But, they used the same data for learning the prior and the posterior simultaneously. This issue will make the ability of generalization worse.

There have been many recent developments in model averaging. Refs. [14, 18] presented two criteria, Mallows criterion and jackknife criterion, to determine the weights of model averaging. Their meanings are not as directly as minimizing the upper bound of the risk. They didn't build the relation between the risk and the criteria theoretically. Refs. [6, 19] gave good estimators of the risk in a certain type of models while our work doesn't specify the model type. For getting the posterior distribution of the weights, Ref. [17] gave a method without choosing a proper prior. Ref. [1] provided an extended PAC-Bayes bound for learning the proper

priors. Nevertheless, it involved reusing of the data which increased the probability of overfitting.

In this paper, we propose a specific risk bound under our settings and two data-based methods for adjusting the priors in PAC-Bayes framework. And, two practical algorithms are given accordingly. The main contributions of this work are the following. First, sequential batch sampling method is proposed to deal with the situation that there isn't historical data while the data can be sampled with the rules made by researchers. Second, when the historical data existed, we use similar old tasks to extract the mutual knowledge with the current task for adjusting the priors. Third, two theoretical risk bounds are provided for these two situations respectively. Fourth, empirical demonstration shows that the proposed meta-methods have excellent performances in the numerical studies.

The reminder of this paper is organized as follows. In Sect. 23.2, a standard risk bound and a practical sequential batch sampling method are established for obtaining a better prior in no previous data situation. Section 23.3 proposes the method to deal with historical similar data for the same purpose. Illustrative simulations given in Sect. 23.4 show that our algorithms will lead to more effective prediction and support our theoretical results. For real-world dataset, we apply the proposed methods to two real datasets and confirm the higher prediction accuracy of minimizing risk bound method. Section 23.5 concludes this paper with some discussions. Some proofs of theorems are delegated to the supplementary materials.

23.2 Sequential Adjustment of Priors

In a traditional supervised learning task, the learner needs to find an optimal *model* (or hypothesis) to fit the data, and then uses the learned model to make predictions. In the Bayesian approach, various models are allowed to fit the data. In particular, the learner needs to learn an optimal model *distribution* over the candidate models, and then uses the learned model distribution to make predictions.

More specifically, in a supervised learning task, we are given a set $S = \{(x_i, y_i)\}_{i=1}^n$ of i.i.d. samples drawn from an unknown distribution D over $\mathscr{X} \times \mathscr{Y}$, i.e., $(x_i, y_i) \sim D$. The goal is to find a model h in the candidate model set $\mathscr{H}$, a set of functions mapping features (feature vector) to responses, that minimizes the expected loss function $\mathbb{E}_{(x,y)\sim D} L(h, x, y)$, where L is a bounded loss function. Without loss of generality, we assume L is bounded by $[0, 1]$. In the Bayesian framework, a distribution Q over $\mathscr{H}$ is the purpose instead of searching a specific optimal model $h \in \mathscr{H}$. Therefore, the goal turns to finding the optimal model distribution Q, which minimizes $\mathbb{E}_{h\sim Q}\mathbb{E}_{(x,y)\sim D} L(h, x, y)$. Then one could use the weighted average of the models over $\mathscr{H}$ to make predictions, namely, $\widehat{y} = \mathbb{E}_{h\sim Q} h(x)$. More generally, we further assume that the candidate model set $\mathscr{H}$ consists of K classes of models $\mathscr{M}_1, \mathscr{M}_2, \ldots, \mathscr{M}_K$ with $\mathscr{H} = \bigcup_{k=1}^K \mathscr{M}_k$. Each model class $\mathscr{M}_k$ is associated with a probability w_k, and for each model class $\mathscr{M}_k$, there is a distribution Q_k over $\mathscr{M}_k$. For example, a model class $\mathscr{M}_k$ could be a group of models obtained from the Lasso

method, and the hyper-parameter λ in Lasso follows a distribution Q_k. Another common example is that $\mathscr{M}_k$ is a group of neural networks with a certain architecture, and the weights of neural networks follow a joint distribution Q_k. In this way, the total distribution over $\mathscr{H}$ can be written as $\xi = (w, Q_1, \ldots, Q_K)$, where w consists of $w_1, \ldots, w_K$ with $\|w\|_1 = 1$. The goal of the learning task is to find an optimal distribution ξ which minimizes the expected risk $R(\xi, D) := \mathbb{E}_{h\sim\xi}\mathbb{E}_{(x,y)\sim D}L(h, x, y)$, and then the prediction is made by $\widehat{y} = \mathbb{E}_{h\sim\xi}h(x) = \sum_{k=1}^{K}[w_k \cdot \mathbb{E}_{h\sim Q_k}h(x)]$.

Since sample distribution D is unknown, the expected risk $R(\xi, D)$ cannot be computed directly. Therefore, it is usually be approximated by the empirical risk $\widehat{R}(\xi, S) := \mathbb{E}_{h\sim\xi}\sum_{(x_i,y_i)\in S}L(h, x_i, y_i)/|S|$ in practice, and ξ is learned by minimizing the empirical risk $\widehat{R}(\xi, S)$. When the sample size is large enough, it would be a good approximation. However, in many situations, we don't have so much data, which may lead to large difference between them. Thus, using the empirical risk $\widehat{R}(\xi, S)$ to approximate the expected risk $R(\xi, D)$ is not appropriate any longer.

We first study the difference between the empirical risk $\widehat{R}(\xi, S)$ and the expected risk $R(\xi, D)$. Based on the literature [7], we can obtain an upper bound of their difference which is stated as the following theorem.

Theorem 23.1 *Let ξ^0 be a prior distribution over $\mathscr{H}$ that must be chosen before observing the samples, and let $\delta \in (0, 1)$. Then with probability at least $1 - \delta$, the following inequality holds for all posterior distributions ξ over $\mathscr{H}$,*

$$R(\xi, D) \leq \widehat{R}(\xi, S) + \sqrt{\frac{\mathrm{KL}(w||w^0) + \sum_{k=1}^{K} w_k \mathrm{KL}(Q_k||Q_k^0) + \ln\frac{n}{\delta}}{2(n-1)}}, \tag{23.1}$$

where n is the cardinality of sample set S, and $\mathrm{KL}(\cdot||\cdot)$ denotes the Kullback-Leibler (KL) divergence between two distributions.[1]

According to the above theorem, it is clear that only when the sample size n is large, the difference $R(\xi, D) - \widehat{R}(\xi, S)$ can be guaranteed to be small. Thus, minimizing $\widehat{R}(\xi, S)$ may not lead to the minimizer of $R(\xi, D)$, which matches our intuition. To avoid the risk of the approximation, one can minimizes the upper bound of the expected risk $R(\xi, D)$ in stead of using the empirical risk $\widehat{R}(\xi, S)$ as an approximation. In particular, we denote the right hand side of Eq.(23.1) by $\overline{R}(\xi, \xi^0, S)$. Then one can learn the model distribution ξ by minimizing $\overline{R}(\xi, \xi^0, S)$. Intuitively, such choice of ξ for the learning task makes the worst case best.

Theorem 23.1 also indicates that the prior ξ^0 plays an important role. Since the choice of ξ balances the tradeoff between the empirical risk $\widehat{R}(\xi, S)$ and the regularization term, if the prior ξ^0 is far away from the true optimal model distribution ξ^*, the posterior ξ will also be bad. The best situation for optimizing the posterior ξ is that the prior ξ^0 exactly equals to the true optimal model distribution ξ^*. Then, the regularization term disappears. In other words, if there is a good prior ξ^0 which is close to ξ^*, the upper bound $\overline{R}(\xi, \xi^0, S)$ will be small. However, without

[1] $\mathrm{KL}(P||P^0)$ is defined as $\mathbb{E}_{x\sim P}\ln\frac{P(x)}{P^0(x)}$.

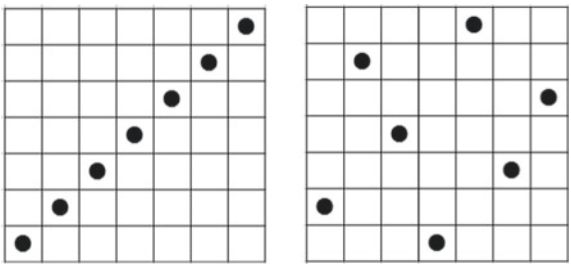
any prior knowledge, one can only use data to help obtain a better prior. The naive method is directly using the non-informative prior as ξ^0 for minimizing $\overline{R}(\xi, \xi^0, S)$ to get the posterior ξ. In this paper, we propose a more carefully designed method to get a better posterior than the naive method. In the following, we consider two different scenarioes for learning the prior. First, the data can be collected adaptively. The learner is allowed to do sampling in rounds and updates the prior distribution after each sampling. In each round, the learner can sample the data according to the prior distribution in the current round. Such iterative procedure updates the prior step by step. Ultimately, compared with dealing the whole data at once, this procedure of adjusting prior leads to a smaller upper bound. Moreover, it also gives an opportunity to choose some good sample sets for reducing the volatility of the estimators which is measured by $v(\xi, D) = \mathbb{E}_x \mathbb{E}_h (h(x) - \mathbb{E}_h h(x))^2$. The function $\widehat{v}(\xi, B) = \frac{1}{|B|} \sum_{x \in B} \mathbb{E}_h (h(x) - \mathbb{E}_h h(x))^2$ is defined to measure the volatility of the posterior ξ at the sample set B. The complete algorithm for sequential batch sampling is shown in Algorithm 6. Second, the data including the new task and other similar old tasks which have been already collected. The sequential sampling method can not be adopted in this scenario. Since the previous tasks are similar with the new task, we could use these old data to learn the prior for the new task. The details will be discussed in Sect. 23.3.

Algorithm 6 Sequential Batch Sampling Algorithm

1: Obtain a sample set B_1 from the sample space $\mathscr{X} \times \mathscr{Y}$ by a initial space-filling design.
2: Get the posterior ξ_1 based on the sample set B_1 by minimizing the risk bound with non-informative prior.
3: **for** $i = 2$ to b **do**
4: Search next sample set B_i ($|B_i| = n_b$) with the large volatility under the current posterior ξ_{i-1}, i.e., $\widehat{v}(\xi_{i-1}, B_i) > \gamma_i$ where γ is a given constant vector.
5: Get the posterior ξ_i based on the sample set B_i by minimizing the risk bound with the prior ξ_{i-1}.
6: **end for**
7: The final posterior is ξ_b.

For Algorithm 6, the data is processed in b steps. First, a space-filling design is used as initial experiment points to reduce the probability of overfitting caused by the unbalanced sampling. Traditional space-filling design aims to fill the input space with design points that are as "uniform" as possible in the input space. The uniformity of space-filling design is illustrated in Fig. 23.1. For next steps, uncertain points are needed to be explored. And, the uncertainty is measured by the volatility v. Hence, the batch with large volatility will be chosen. Note that if we set a huge γ, we will just explore a small region of the input space.

The setting of γ refers to [20]. However, in practice, it is found that this parameter γ does not matter much, since the results are similar with a wide range of γ. This procedure helps to reduce the variance of the estimator which is proved in [20] by sequential sampling. Furthermore, it also helps to adjust the prior in each step which is called learning the prior. The proposition is stated as below.



(a) Nonuniform design (b) Uniform design

Fig. 23.1 The illustration for uniform space-filling design

Proposition 23.1 *For $i = 1, 2, \ldots, b$, let $B_i = S$, ξ^* is the minimizer of the RHS of Eq. (23.1) with non-informative prior ξ^0 and ξ_i are obtained by Algorithm 6, then we have $\overline{R}(\xi_b, \xi_{b-1}, S) \leq \overline{R}(\xi^*, \xi^0, S)$.*

The above proposition can be understood straightforwardly. First, since we adjust the prior through the data step by step, the final prior ξ_{b-1} is better than the non-informative prior. Consequently, it receives the smaller expected risk. Second, we choose the sample sets sequentially with large volatility to do experiments in order to reduce uncertainty. The property is also confirmed in Sect. 23.4.

23.3 Priors Based on Historical Data

As mentioned in Sect. 23.2, when the data of historical tasks and the new tasks have already collected, sampling method can not be used any longer. Still, the learner needs a good prior for the reliable inferences. In order to get a good prior, it is helpful to extract the mutual knowledge from similar tasks. In particular, there are m sample tasks $T_1, \ldots, T_m$ i.i.d. generated from an unknown task distribution τ. For each sample task T_i, a sample set S_i with n_i samples is generated from an unknown distribution D_i. Without ambiguity, we use notation $\xi(\xi^0, S)$ to denote the posterior under the prior ξ^0 after observing the sample set S. The quality of a prior ξ^0 is measured by $\mathbb{E}_{D_i \sim \tau} \mathbb{E}_{S_i \sim D_i^{n_i}} R(\xi(\xi^0, S_i), D_i)$. Thus, the expected loss we want to minimize is

$$R(\xi^0, \tau) = \mathbb{E}_{D_i \sim \tau} \mathbb{E}_{S_i \sim D_i^{n_i}} R(\xi(\xi^0, S_i), D_i).$$

Similar to the single-task case, the above expected risk cannot be computed directly, thus the following empirical risk is used to estimate it:

$$\widehat{R}(\xi^0, S_1, \ldots, S_m) = \frac{1}{m} \sum_{i=1}^{m} \widehat{R}(\xi(\xi^0, S_i^{train}), S_i^{validation}),$$

where each sample set S_i is divided into a training set S_i^{train} and a validation set $S_i^{validation}$.

Consider the regression setting for task T. Suppose the true model is

$$y_T = f_T(x_T) + \sigma_T(x_T) \cdot \varepsilon_T,$$

where $f_T : \mathbb{R}^d \to \mathbb{R}$ is the function to be learned, the error term ε_T is assumed to be independent of X and has a known probability density $q(t), t \in \mathbb{R}$ with mean 0 and a finite variance. The unknown function $\sigma_T(x_T)$ controls the variance of the error at $X = x_T$. There are n_T i.i.d. samples $\{(x_{T,i}, y_{T,i})\}_{i=1}^{n_T}$ drawn from an unknown joint distribution of (x_T, y_T). Assume that there is a candidate model set $\mathscr{H}$. Each of them is a function mapping features (feature vector) to response, i.e., $h \in \mathscr{H} : \mathbb{R}^d \to \mathbb{R}$. To take the information of the old tasks, which can reflect the importance of each $h \in \mathscr{H}$, the following Algorithm 7 is proposed.

Algorithm 7 Historical Data Related Algorithm

1: **for** $i = 1$ to m **do**
2: Using T_i to obtain ξ_i by minimizing the risk bound with non-informative prior.
3: **end for**
4: **for** $i = 1$ to m **do**
5: Randomly split the data S_i into two parts $S_{i,n_i'}^{(1)} = (x_{i,\alpha}, y_{i,\alpha})_{\alpha=1}^{n_i'}$ for training and $S_{i,n_i'}^{(2)} = (x_{i,\alpha}, y_{i,\alpha})_{\alpha=n_i'+1}^{n_i}$ for validation.
6: **for** each $j \neq i$ **do**
7: Obtain estimates $\widehat{f}_{j,n_i'}(x, S_{i,n_i'}^{(1)}), \widehat{\sigma}_{j,n_i'}(x, S_{i,n_i'}^{(1)})$ with prior ξ_j.
8: Evaluate predictions on $S_{i,n_i'}^{(2)}$ and compute

$$E_j^i = \frac{\Pi_{\alpha=n_i'+1}^{n_i} q\left(\frac{y_\alpha - \widehat{f}_{j,n_i'}(x_{i,\alpha})}{\widehat{\sigma}_{j,n_i'}(x_{i,\alpha})}\right)}{\Pi_{\alpha=n_i'+1}^{n_i} \widehat{\sigma}_{j,n_i'}(x_{i,\alpha})}.$$

9: **end for**
10: **end for**
11: Repeat the random data segmentation more times and average the weights E_j^i after normalization to get $w_j^{(i)} (j \neq i)$.
12: Average all the $w_j^{(i)} (j \neq i)$ from $i = 1$ to m to obtain the final weights w_j.
13: The prior learned for a new task is $\xi^* = \sum_{i=1}^m w_i \xi_i$.

This algorithm is based on the cross-validation framework. First, using T_i to obtain the candidate priors ξ_i by minimizing the risk bound with non-informative prior. Cross-validation determines the importance of the priors. The j-th task is divided into two parts randomly. The first part is used to learn the posterior with the prior ξ_i. The second part is to evaluate the performance of the posterior by its likelihood function. This evaluation is inspired by [9]. To simplify the determination of the

weights, Ref. [9] proposed a frequentist approach to BMA. The Bayes' theorem was replaced by the Schwarz asymptotic approximation which could be viewed as using maximized likelihood function as the weights of the candidate models. The $\widehat{\sigma}$ on the denominator of E_j^i makes the weight larger if the model is accurate. This procedure repeats many times for each pair (i, j). Their averages reveal the importance of the priors. In the end, the ξ^* is obtained by weighted averaging them all. the property of this algorithm can be guaranteed by the following theorem.

The following regularization conditions are assumed for the results. First, q is assumed to be a known distribution with 0 and variance 1.

(C1) The functions f and σ are uniformly bounded, i.e., $\sup_x |f(x)| \leq A < \infty$ and $0 < c_m \leq \sigma(x) \leq c_M < \infty$ for constants A, c_m and c_M.

(C2) The error distribution q satisfies that for each $0 < s_0 < 1$ and $c_T > 0$, there exists a constant B such that

$$\int q(x) \ln \frac{q(x)}{\frac{1}{s} q(\frac{x-t}{s})} \mu(dx) \leq B((1-s)^2 + t^2)$$

for all $s_0 \leq s \leq s_0^{-1}$ and $-c_T \leq t \leq c_T$.

(C3) The risks of the estimators for approximating f and σ^2 decrease as the sample size increases.

For the condition (C1), note that, when we deal with k-way classification tasks, the responses belong to $\{1, 2, \ldots, k\}$ which is bounded obviously. Moreover, if the input space is a finite region which often happens in real datasets, most common functions are bounded uniformly. The constants A, c_m, c_M are involved in the derivation of the risk bounds, but they can be unknown in practice when we implement the Algorithm 7. The condition (C2) is satisfied by Gaussian, t (with degrees of freedom larger than two), double-exponential, and so on. The condition (C3) usually holds for a good estimating procedure, like consistent estimators. A model has consistency if the expected risk tends to zero when experimental size tends to infinity. Note that the conditions are satisfied in most situations.

Theorem 23.2 *Assume (C1)–(C3) are satisfied and σ_{T_i} is known. Then, the combined posterior ξ^* as given above satisfies*

$$R(\xi^*, \tau) \leq \inf_j \left(\frac{C_1}{\sum_{i \neq j}(n_i - n_i')} + \frac{C_2}{\sum_{i \neq j}(n_i - n_i')} \sum_{i \neq j} (n_i - n_i') \left[\widehat{R}(\xi_j^*, S_{i,n_i'}^{(2)}) + \sqrt{\frac{\mathrm{KL}(w_j{}^* || w_j) + \sum_{k=1}^K w_{j,k} \mathrm{KL}(Q_{j,k}^* || Q_{j,k}) + \ln \frac{n_i}{\delta}}{2(n_i - 1)}} \right] \right)$$

with probability at least $1 - \delta$, where the constant C_1, C_2 depend on the regularization conditions, π is the initial prior which should be non-informative prior and ξ_j^* is the minimizer of Eq. (23.1) with $\xi^0 = \xi_j$ and $S = S_{i,n_i'}^{(1)}$.

For simplify, we assume that the condition that σ_{T_i} is known in Theorem 23.2. In fact it is not a necessary condition, a more general case and corresponding proof can be found in Appendix.

In this general proof, it can be seen that variance estimation is also important for the Algorithm 7. Even if a procedure estimates f_T very well, a bad estimator of σ_T can substantially reduce its weight in the final estimator. Under the condition (C3), the risks of a good procedure for estimating f_T and σ_T usually decrease as the sample size increases. The influence of the number of testing points n_i' is quite clear. Smaller n_i' decreases the first penalty term but increases the main terms that involve the risks of each j. Moreover, Theorem 23.2 reveals the vital property that if one alternative model is consistent, the combined model will also have the consistency.

23.4 Simulations

In this section, some examples are shown to illustrate the procedure of Algorithms 6 and 7 and confirm Proposition 23.1. The method of minimizing the upper bound in Theorem 23.1 with non-informative prior is denoted by RBM (Risk Bound Method). Also, the SOIL method in [17] is under the comparison. The optimization for RHS of Eq. (23.1) in our algorithms is dealt by gradient descend method. R package "SOIL" is used to obtain the results of the SOIL method. First, we begin with linear models.

23.4.1 Synthetic Data Analysis

Example 23.1 The simulation data $\{(x_i, y_i)\}_{i=1}^n$ is generated for the RBM from the linear model $y_i = 1 + {x_i}^T\beta + \sigma\varepsilon_i$, where $\varepsilon_i \sim N(0, 1)$, $\sigma \in \{1, 5\}$ and $x_i \sim N_d(0, \Sigma)$. For each element Σ_{ij} of Σ, $\Sigma_{ij} = \rho^{|i-j|}$ $(i \neq j)$ or 1 $(i = j)$ with $\rho \in \{0, 0.9\}$. The sequential batch sampling has b steps, and each step uses n/b samples followed Algorithm 6.

All the specific settings for parameters are summarized in Table 23.1, and the confidence level δ in Theorem 23.1 is set to 0.01. The Mean Squared Prediction Error(MSPE) $\mathbb{E}_x|f(x) - \widehat{f}(x)|^2$ and volatility defined in Sect. 23.2 are compared. They are obtained by sampling 1000 samples from the same distribution and computing their empirical MSPE $\sum_x |f(x) - \widehat{f}(x)|^2/10^3$ and volatility. For each model setting with a specific choice of the parameters (ρ, σ), we repeat 100 times and compute the average empirical value. The comparison among RBM, SOIL and SBS(Sequential Batch Sampling) are shown in Table 23.2.

The volatility of SOIL method is the smallest and very close to zero. This phenomenon shows that SOIL is focused on a few models, even just one model when the volatility equals to zero. Consequently, its MSPE is larger than other two

Table 23.1 Simulation settings of Example 23.1

Model	n	d	b	β
1	50	8	5	$(3, 1.5, 0, 0, 2, 0, 0, 0)^T$
2	150	50	5	$(1, 2, 3, 2, 0.75, 0, \ldots, 0)^T$
3	50	50	5	$(1, 1/2, 1/3, 1/4, 1/5, 1/6, 0, \ldots, 0)^T$

Table 23.2 Comparison among RBM, SOIL and SBS of Example 23.1

Model 1	(ρ, σ)	(0, 1)	(0, 5)	(0.9, 1)	(0.9, 5)
MSPE	RBM	2.03	48.23	3.71	53.83
	SOIL	2.13	53.21	2.17	53.21
	SBS	**1.71**	**14.08**	**3.25**	**26.40**
Volatility	RBM	1.64	3.47	1.31	0.49
	SOIL	0	0	0.002	0
	SBS	1.61	7.41	1.03	0.42
Model 2					
MSPE	RBM	1.97	46.26	1.46	35.97
	SOIL	2.01	50.23	1.96	49.78
	SBS	**1.93**	**38.69**	**1.38**	**12.92**
Volatility	RBM	1.60	2.72	3.38	7.48
	SOIL	0	0	0.001	0.01
	SBS	1.46	8.67	3.35	6.74
Model 3					
MSPE	RBM	1.67	42.06	1.24	38.51
	SOIL	1.99	49.80	1.93	47.99
	SBS	**1.65**	**27.32**	**1.23**	**29.44**
Volatility	RBM	0.27	1.54	0.74	3.39
	SOIL	0	0	0.02	0.36
	SBS	0.29	0.47	0.77	4.06

methods. SBS as a modification of RBM has similar results with RBM when σ is small. However, when σ is large, SBS performs much better than RBM. In this situation, the information of data is easily covered by big noises. Hence, a good prior which can provide more information is vital for this procedure.

Next example considers the same comparison but in non-linear models. In last example, the alternative models include the true model, but now the true non-linear model is approximated by many linear models.

Example 23.2 The simulation data $\{(x_i, y_i)\}_{i=1}^{50}$ is generated for the RBM from the non-linear models

1. $y_i = 1 + \sin(x_{i,1}) + \cos(x_{i,2}) + \varepsilon_i$,
2. $y_i = 1 + \sin(x_{i,1} + x_{i,2}) + \varepsilon_i$,

Table 23.3 Comparison among RBM, SOIL and SBS of Example 23.2

		Model 1	Model 2
MSPE	RBM	1.26	1.54
	SOIL	1.42	1.80
	SBS	**1.23**	**1.47**
Volatility	RBM	0.1	0.11
	SOIL	0.07	0.02
	SBS	0.11	0.14

where $\varepsilon_i \sim N(0, 1)$, and $x_i \sim N_8(0, I)$. The sequential batch sampling has 5 steps, and each step uses 10 samples followed Algorithm 6.

The results of Example 23.2 is listed in Table 23.3. Mostly, it is similar with the results of Example 23.1. The difference is that the volatility of SOIL becomes large when the model is completely non-linear. Using linear models to fit non-linear model obviously increases the model uncertainty, since none of the fitting models is correct.

The final example is under the situation that the data has been already collected. Hence, we can't use the SBS method to get the data. However, we have the extra data of many old similar tasks. In particular, we have the data of Example 23.1. Now, the new task is to fit a new model.

Example 23.3 The data of Example 23.1 with $(\rho, \sigma) = (0, 1)$ is given. The new task data $\{(x_i, y_i)\}_{i=1}^{20}$ is generated from the linear model $y_i = 1 + {x_i}^T\beta + \sigma\varepsilon_i$, where $\varepsilon_i \sim N(0, 1)$, $\sigma \in \{1, 2, 3, 4, 5\}$, $\beta = \{1, -1, 0, 0, 0.5, 0, \ldots, 0\}$ and $x_i \sim N_{10}(0, I)$.

The method described in Algorithm 7 is denoted by HDR (Historical Data Related). The results in Fig. 23.2 show the high consistency with the last two examples. When σ is small, the different priors lead to similar result since the current data has key influence. However, when σ is large, the difference between RMB and HDR is huge. The reason is that the current data has been polluted by the strong noise. Hence, a good prior can provide the vital information about the model distribution.

23.4.2 Real Data Study

Here, we apply the proposed methods to two real datasets, BGS data and Bardet data, which are also used in [17].

First, the BGS data is with small d and from the Berkeley Guidance Study (BGS) by [13]. The dataset records 66 boys' physical growth measures from birth to eighteen

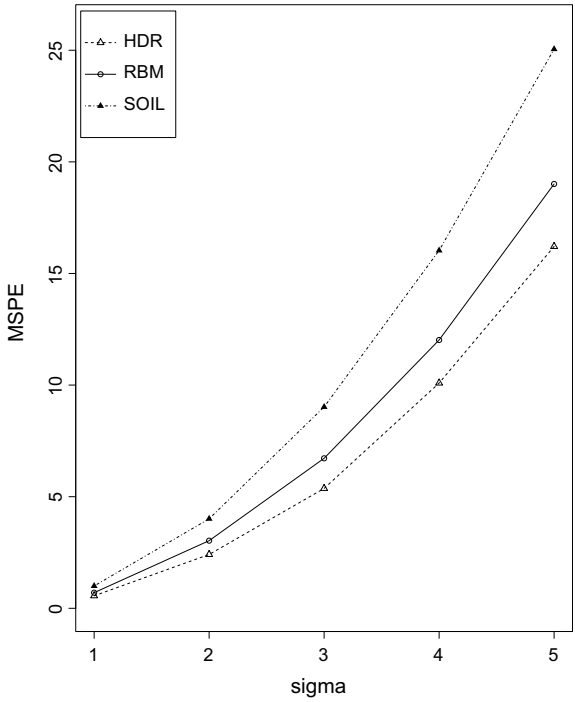


Fig. 23.2 Comparisons among the three methods in Example 23.3

years. Following [17], we consider the same regression model. The response is age 18 height and the factors include weights at ages two (WT2) and nine (WT9), heights at ages two (HT2) and nine (HT9), age nine leg circumference (LG9) and age 18 strength (ST18).

Second, for large d, the Bardet data collects tissue samples from the eyes of 120 twelve-week-old male rats. For each tissue, the RNAs of 31, 042 selected probes are measured by the normalized intensity valued. The gene intensity values are in log scale. Gene TRIM32, which causes the Bardet-Biedl syndrome, is the response in this study. The genes that are related to it are investigated. A screening method [3] is applied to the original probes. This screened dataset with 200 probes for each of 120 tissues is also used in [17].

Both cases are data-given cases that we can't use sequential batch sampling method. For the different setting of d, we assign corresponding similar historical data for two real datasets. The data of model 1 in Example 23.1 for the BGS data with small d. The data of model 3 in Example 23.1 for the Bardet data with large d.

We randomly sample 10 rows from the data as the test set to calculate empirical MSPE and volatility. The results are summarized in Table 23.4. From Table 23.4, we can see that both RBM and HDR have smaller MSPE than SOIL. However, HDR doesn't perform much better than RBM. This can be explained intuitively as follows. In theory, the historical tasks and the current task are assumed that they come from the

Table 23.4 Comparison among RBM, SOIL and HDR in real data

		BGS	Bardet
MSPE	RBM	13.54	0.0054
	SOIL	16.74	0.0065
	HDR	**13.06**	**0.0050**
Volatility	RBM	1.99	0.0013
	SOIL	0.43	0.0013
	HDR	1.84	0.0012

same task distribution. But in practice, how to measure the similarity between tasks is still a problem. Hence, an unrelated historical dataset may provide less information for the current prediction.

23.5 Concluding Remarks

This paper is based on the PAC-Bayes framework to study the model averaging problem. More concretely, the work is about how to assign the proper distribution on the candidate models. The work proposes specific upper bounds of the risks in different situations and aims to minimize them. In other words, it makes the worst situation best. For this purpose, two practical algorithms are provided to solve this optimization under two realistic situations respectively. One is that no previous data can be used, but the experimenters have the opportunity to design the sampling method before the collection of the data. The other one is that much historical data is given, the analysts should figure out a proper method to deal with these data. In the first case, the prior is adjusted step by step. Compared with dealing the whole data at once, this sequential method has the smaller upper bound of the risk. In the second case, using historical similar tasks to extract the information about the prior which is called meta-learning. The meta-learner is for the prior and the base-learner is for the posterior. Both methods are confirmed to be effective in our simulation and real data study.

However, some problems need to be investigated. First, in sequential batch sampling procedure, the volatility is used as a criterion to sample the data. This choice is based on our experience. There may exist other choices that have better results. Second, when a lot of historical data is available, many similar old tasks may be considered to extract more information for learning the new task better. How to define 'similar' is still an open problem. In practice, the similarity isn't measured by the data. Instead, it is judged by experts, which is not expected.

Acknowledgments The authors sincerely thank the editors and a referee for their valuable comments, which further improve this paper. The work is supported by NSFC grant 11671019, LMEQF and Beijing Institute of Technology Research Fund Program for Young Scholars.

Appendix

First, we review the classical PAC-Bayes bound [7, 12] with general notations.

Lemma 23.1 *Let $\mathscr{X}$ be a sample space and $\mathscr{F}$ be a function space over $\mathscr{X}$. Define a loss function $g(f, X) : \mathscr{F} \times \mathscr{X} \to [0, 1]$, and $S = \{X_1, \ldots, X_n\}$ be a sequence of n independent identical distributed random samples. Let π be some prior distribution over $\mathscr{F}$. For any $\delta \in (0, 1]$, the following bound holds for all posterior distributions ρ over $\mathscr{F}$,*

$$\mathbb{P}_S\left(\mathbb{E}_X\mathbb{E}_f g(f, X) \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_f g(f, X_i) + \sqrt{\frac{\mathrm{KL}(\rho||\pi) + \ln\frac{n}{\delta}}{2(n-1)}}\right) \ge 1 - \delta. \tag{23.2}$$

Proof of Theorem 23.1 We use Lemma 23.1 to bound the expected risk with the following substitutions. The n samples are $X_i \triangleq z_i$. The function $f \triangleq h$ where $h \in \mathscr{H}$. The loss function $g(f, X) \triangleq L(h, z) \in [0, 1]$. The prior π is defined by $\pi \triangleq \xi^0$, in which we first sample k from $\{1, \ldots, K\}$ according to corresponding weights $\{w_1, \ldots, w_K\}$ and then sample h from Q_k. The posterior is defined similarly, $\rho \triangleq \xi$.

The KL-divergence term is

$$\begin{aligned}\mathrm{KL}(\rho||\pi) &= \mathbb{E}_f \ln\frac{\rho(f)}{\pi(f)} = \mathbb{E}_{k\in\{1,\ldots,K\}}(\mathbb{E}_h \frac{Q_k(h)}{Q_k^0(h)}|h \in \mathscr{M}_k)\\ &= \sum_{k=1}^{K} w_k \mathbb{E}_{h\in\mathscr{M}_k} \ln\frac{w_k Q_k(h)}{w_{0,k} Q_k^0(h)}\\ &= \mathrm{KL}(w||w^0) + \sum_{k=1}^{K} w_k \mathrm{KL}(Q_k||Q_k^0).\end{aligned} \tag{23.3}$$

Substituting the above into Eq. (23.2), it follows that

$$\begin{aligned}&\mathbb{P}_S\Bigg(\mathbb{E}_z\mathbb{E}_{k\in\{1,\ldots,K\}}\mathbb{E}_{h\in\mathscr{M}_k} L(h, z) \le \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{k\in\{1,\ldots,K\}}\mathbb{E}_{h\in\mathscr{M}_k} L(h, z)\\ &+ \sqrt{\frac{\mathrm{KL}(w||w^0) + \sum_{k=1}^{K} w_k \mathrm{KL}((Q_k||Q_k^0) + \ln\frac{n}{\delta}}{2(n-1)}}\Bigg) \ge 1 - \delta.\end{aligned} \tag{23.4}$$

Using the notations in Sect. 23.2, we can rewrite the above as below,

$$\mathbb{P}_S\left(R(\xi, D) \le \widehat{R}(\xi, S) + \sqrt{(\mathrm{KL}(w||w^0) + \sum_{k=1}^{K} w_k \mathrm{KL}(Q_k||Q_k^0) + \ln\frac{n}{\delta})/(2n-2)}\right) \ge 1 - \delta. \tag{23.5}$$

Proof of Proposition 23.1 First, we proof that for $i = 2, \ldots, b$,

$$\overline{R}(\xi_i, \xi_{i-1}, B_i) \leq \overline{R}(\xi_{i-1}, \xi_{i-2}, B_{i-1}).$$

By definition of ξ_i,

$$\begin{aligned}
&\overline{R}(\xi_i, \xi_{i-1}, B_i) \leq \overline{R}(\xi_{i-1}, \xi_{i-1}, B_i) \\
&= \widehat{R}(\xi_{i-1}, B_i) + \sqrt{\ln \frac{n}{\delta}/(2n-2)} \\
&\leq \overline{R}(\xi_{i-1}, \xi_{i-2}, B_i) = \overline{R}(\xi_{i-1}, \xi_{i-2}, B_{i-1}).
\end{aligned}$$

Following these inequalities,

$$\overline{R}(\xi_b, \xi_{b-1}, S) = \overline{R}(\xi_b, \xi_{b-1}, B_b) \leq \overline{R}(\xi_1, \xi^0, B_1) = \overline{R}(\xi^*, \xi^0, S).$$

This finishes the proof.

Proof of Theorem 23.2 According to Theorem 1 in [16], we have

$$\begin{aligned}
R(\xi^*, \tau) \leq \inf_j \Bigg(&\frac{C_1}{\sum_{i \neq j}(n_i - n_i^{'})} \\
&+ \frac{C_2}{\sum_{i \neq j}(n_i - n_i^{'})} \sum_{i \neq j} \sum_{\alpha = n_i^{'}+1}^{n_i} \left[\mathbb{E}||\sigma_{T_i}^2 - \widehat{\sigma}_{j,\alpha}^2||^2 + R(\xi_j^*, D_i) \right] \Bigg),
\end{aligned} \tag{23.6}$$

where ξ_j^* is the minimizer of Eq. (23.1) with $\xi_0 = \xi_j$ and $S = S_{i,\alpha}^{(1)}$ denoted by $\xi_j^*(\xi_j, S_{i,\alpha}^{(1)})$.

For any $\alpha \geq n_i^{'}$ and an estimator satisfied the condition (C3), the inequalities $\mathbb{E}||\sigma_{T_i}^2 - \widehat{\sigma}_{j,n_i^{'}}^2||^2 \geq \mathbb{E}||\sigma_{T_i}^2 - \widehat{\sigma}_{j,\alpha}^2||^2$ and $R(\xi_j^*(\xi_j, S_{i,n_i^{'}}^{(1)}), D_i) \geq R(\xi^*(\xi_j, S_{i,\alpha}^{(1)}), D_i)$ hold. Plugging into Eq. (23.6) for $\alpha = n_i^{'} + 1, \ldots, n_i$, it follows that

$$R(\xi^*, \tau) \leq \inf_j \left(\frac{C_1}{\sum_{i \neq j}(n_i - n_i^{'})} + \frac{C_2}{\sum_{i \neq j}(n_i - n_i^{'})} \sum_{i \neq j} \left[\mathbb{E}||\sigma_{T_i}^2 - \widehat{\sigma}_{j,n_i^{'}}^2||^2 + R(\xi_j^*, D_i) \right] \right),$$

where ξ_j^* is the minimizer of Eq. (23.1) with $\xi^0 = \xi_j$ and $S = S_{i,n_i^{'}}^{(1)}$.

Then, the result follows by the above inequality combined with Eq. (23.5). In order to obtain the form in Theorem 23.2, one only needs to note that if σ_{T_i} is known, the term $\mathbb{E}||\sigma_{T_i}^2 - \widehat{\sigma}_{j,n_i^{'}}^2||^2$ vanishes.

References

1. Amit, R., Meir, R.: Meta-learning by adjusting priors based on extended pac-bayes theory. In: International Conference on Machine Learning, pp. 205–214 (2018)
2. Hoeting, J.A., Madigan, D., Raftery, A.E., Volinsky, C.T.: Bayesian model averaging: A tutorial. Stat. Sci. **14**(4), 382–401 (1999)
3. Huang, J., Ma, S., Zhang, C.H.: Adaptive lasso for sparse high-dimensional regression. Stat. Sin. **18**(4), 1603–1618 (2008)
4. Leamer, E.E.: Specification searches. Wiley, New York (1978)
5. Lever, G., Laviolette, F., Shawe-Taylor, J.: Tighter pac-bayes bounds through distribution-dependent priors. Theor. Comput. Sci. **473**(2), 4–28 (2013)
6. Liang, H., Zou, G., Wan, A.T.K., Zhang, X.: Optimal weight choice for frequentist model average estimators. J. Am. Stati. Assoc. **106**(495), 1053–1066 (2011)
7. Mcallester, D.A.: Pac-bayesian model averaging. In Proceedings of the Twelfth Annual Conference on Computational Learning Theory, pp. 164–170 (1999)
8. Moral-Benito, E.: Model averaging in economics: An overview. J. Econ. Surv. **29**(1), 46–75 (2015)
9. Raftery, A.E.: Bayesian model selection in social research. Soc. Methodol. **25**(25), 111–163 (1995)
10. Raftery, A.E.: Approximate bayes factors and accounting for model uncertainty in generalised linear models. Biometrika **83**(2), 251–266 (1996)
11. Seeger, M.: Pac-bayesian generalisation error bounds for gaussian process classification. J. Mach. Learn. Res. **3**(2), 233–269 (2002)
12. Shalev-Shwartz, S., Ben-David, S.: Understanding Machine Learning: From Theory to Algorithms. Cambridge University Press (2014)
13. Tuddenham, R.D., Snyder, M.M.: Physical growth of california boys and girls from birth to eighteen years. Publ. Child Dev. **1**, 183–364 (1954)
14. Wan, A.T.K., Zhang, X., Zou, G.: Least squares model averaging by mallows criterion. J Econ. **156**(2), 277–283 (2010)
15. Wasserman, L.: Bayesian model selection and model averaging. J. Math. Psychol. **44**(1), 92–107 (2000)
16. Yang, Y.: Adaptive regression by mixing. J. Am. Stat. Assoc. **96**(454), 574–588 (2001)
17. Ye, C., Yang, Y., Yang, Y.: Sparsity oriented importance learning for high-dimensional linear regression. J. Am. Stat. Assoc. **2**, 1–16 (2016)
18. Zhang, X., Wan, A.T.K., Zou, G.: Model averaging by jackknife criterion in models with dependent data. J. Econ. **174**(2), 82–94 (2013)
19. Zhang, X., Zou, G., Liang, H.: Model averaging and weight choice in linear mixed-effects models. Biometrika **1**(1), 205–218 (2014)
20. Zhou, Q., Ernst, P.A., Morgan, K.L., Rubin, D.B., Zhang, A.: Sequential rerandomization. Biometrika **105**(3), 745–752 (2018)

Chapter 24
Quantile Regression with Gaussian Kernels

Baobin Wang, Ting Hu, and Hong Yin

Abstract This paper aims at the error analysis of stochastic gradient descent (SGD) for quantile regression, which is associated with a sequence of varying ϵ-insensitive pinball loss functions and flexible Gaussian kernels. Analyzing sparsity and learning rates will be provided when the target function lies in some Sobolev spaces and a noise condition is satisfied for the underlying probability measure. Our results show that selecting the variance of the Gaussian kernel plays a crucial role in the learning performance of quantile regression algorithms.

Keywords Quantile regresion · Gaussian kernels · Reproducing kernel Hilbert spaces · Insensitive pinball loss · Learning rate

24.1 Introduction

Quantile regression has been investigated in machine learning and statistics, see [3, 4, 13–15] and references therein. Compared with the least squares regression, quantile regression provides more information about the conditional distributions of output variables such as stretching or compressing tails and multimodality [5, 6]. In the setting of learning problems, let X be a multivariate random variable with

B. Wang
School of Mathematics and Statistics, South-Central University for Nationalities, Wuhan 430074, People's Republic of China
e-mail: wbb@scuec.edu.cn

T. Hu
School of Mathematics and Statistics, Wuhan University, Wuhan 430072, People's Republic of China
e-mail: tinghu@whu.edu.cn

H. Yin (✉)
School of Mathematics, Renmin University of China, Beijing 100872, People's Republic of China
e-mail: yinhong@ruc.edu.cn

J. Fan and J. Pan (eds.), *Contemporary Experimental Design, Multivariate Analysis and Data Mining*,
https://doi.org/10.1007/978-3-030-46161-4_24

values in a compact subset of $\mathbb{R}^n$ and $Y \subset \mathbb{R}$ be a real valued response variable. The purpose of quantile regression is to study the quantile regression functions from a sample of T observations $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^T$ drawn independently according to the identical distribution ρ on $Z = X \times Y$. With a quantile parameter $0 < \tau < 1$, a *quantile regression function* $f_{\rho,\tau} : X \to Y$ is defined by its value $f_{\rho,\tau}(x)$ to be a τ-quantile of the conditional distribution $\rho(\cdot|x)$ of ρ at $x \in X$, that is, a value $u \in Y$ satisfying

$$\rho\left(\{y \in Y, y \leq u\}|x\right) \geq \tau, \text{ and } \rho\left(\{y \in Y, y \geq u\}|x\right) \geq 1 - \tau.$$

Gaussian kernels are one of the most often used kernels in modern machine learning methods such as support vector machines (SVMs) [12, 15]. The Gaussian kernel with variance $\sigma > 0$ is the function on $X \times X$ defined by

$$K_\sigma(x, u) := \exp\left\{-\frac{|x-u|^2}{2\sigma^2}\right\}.$$

Let $\mathscr{H}_\sigma(X)$ be the RKHS [1] on X associated with the kernel K_σ and the inner product $\langle\cdot,\cdot\rangle_{\mathscr{H}_\sigma(X)}$. Its reproducing property takes the form

$$\langle K_\sigma(x,\cdot), f(\cdot)\rangle_{\mathscr{H}_\sigma(X)} = f(x), \forall\, x \in \mathscr{X},\ f \in \mathscr{H}_\sigma(X). \tag{24.1}$$

Quantile regression has been studied by means of kernel-based regularization schemes in a vast literature, see [7, 11, 15]. Its associated loss function is the pinball loss ϕ_τ defined by

$$\phi_\tau(u) = \begin{cases} (1-\tau)u, & \text{if } \ u \geq 0, \\ -\tau u, & \text{if } \ u < 0, \end{cases}$$

and the regularization scheme takes the form

$$f_{\mathbf{z},\lambda} := \arg\min_{f \in \mathscr{H}_\sigma(X)} \frac{1}{T}\sum_{i=1}^{T} \phi_\tau(y_i - f(x_i)) + \frac{\lambda}{2}\|f\|^2_{\mathscr{H}_\sigma(X)}. \tag{24.2}$$

In this paper, SGD method (or called online learning) is taken to solve the scheme (24.2) for its low complexity and good practical performance. Inspired by the work in [15, 19], we consider the below SGD algorithm for quantile regression associated with a varying ϵ-insensitive pinball loss $\phi_\tau^\epsilon(u)$ with an insensitive parameter $\epsilon \geq 0$, given as

$$\phi_\tau^\epsilon(u) = \begin{cases} (1-\tau)(u-\epsilon), & \text{if } \ u > \epsilon, \\ -\tau(u+\epsilon), & \text{if } \ u < -\epsilon, \\ 0, & \text{otherwise.} \end{cases}$$

Definition 24.1 The SGD algorithm for (24.2) is defined by $f_1 = 0$ and

$$f_{t+1} = f_t - \eta_t \left\{ \left(\phi_\tau^{\epsilon_t}\right)'_- (f_t(x_t) - y_t) K_\sigma(x_t, \cdot) + \lambda f_t \right\} \tag{24.3}$$

where $\{\eta_t\} > 0$ is the positive stepsize sequence, $\lambda = \lambda(T)$ is the regularization parameter, $\{\epsilon_t\} > 0$ is the varying insensitive parameters and $\left(\phi_\tau^{\epsilon_t}\right)'_-$ is the left (one-side) derivative of $\phi_\tau^{\epsilon_t}$.

This algorithm is a generalization for the pinball loss ϕ_τ with $\epsilon = 0$ and the ϵ-insensitive loss with $\tau = \frac{1}{2}$ (median). The initial form of quantile regression with a fixed insensitive parameter $\epsilon > 0$ was introduced by [11, 16], which aims at producing possible sparsity of support vectors for the median. Then this idea was developed to τ-quantile regression with any $0 < \tau < 1$ and the ϵ-insensitive pinball loss $\phi_\tau^\epsilon(u)$ was proposed in [15, 17]. In the previous work [2, 17], the corresponding mathematical analysis in the batch learning has been conducted when ϵ change with the sample size T and $\epsilon = \epsilon(T) \to 0$ as the sample size T goes to infinity.

Here the insensitive parameters $\{\epsilon_t\} > 0$ used in the algorithm (24.3) form a decreasing sequence converging to zero when the learning step t increases. In the work [15], Hu et al. derived the learning rate of (24.3) with flexible insensitive parameters $\{\epsilon_t\}$ under the suitable choices of the parameters (λ, η_t) for balancing the approximation and sparsity. Their convergence rate is closely related to the strong assumption on the approximation power of RKHS. Actually, for a Gaussian RKHS $\mathscr{H}_\sigma$ with the fixed variance $\sigma > 0$, the approximation error decays logarithmically with respect to the range of $\mathscr{H}_\sigma$, which has been proved in [8]. So, putting this decay into their analysis leads that the learning rate for quantile regression is rather slow, which is unaccepted in real applications. In simulations, the variance σ of $\mathscr{H}_\sigma$ usually serves as a tuned parameter for a good learning performance in training processes and can be chosen in a data-dependent way such as cross-validation. Since the variance of a Gaussian kernel reflexes the specific structure of RKHS induced by the Gaussian or other important features of learning problems such as the frequency of function components, choosing the variance σ of $\mathscr{H}_\sigma$ is related to the model selection problem, which adjusts the complexity or the capacity of learning problems according to the learning time or sample size. The selecting rule of σ has been studied in various learning settings [7, 12, 18], SVM, least squares, etc.

The goal of this paper is to study the convergence behavior of the algorithm (24.3) with flexible Gaussians and investigate the effects of parameters in keeping sparsity and nice learning power for quantile problems. Our results show that the online quantile regression is feasible in the framework of the Gaussian RKHS, in which the variance of Gaussian serves as a trade-off between the approximation ability and sparsity of the algorithm. We present a selection rule for the variance $\sigma = \sigma(T)$ to avoid over-fitting or under-fitting in the iteration process. The performance of the iterates $\{f_t\}$ is usually measured by the convergence in terms of the excess generalization error. In this work, under the noise condition, we can obtain the convergence result in Banach spaces, which implies that $\{f_t\}$ is closed to the target function $f_{\rho,\tau}$ in a strong sense.

24.2 Main Results and Effects of Parameters

For conceptual simplicity, we assume throughout this paper that the support of the conditional distribution $\rho(\cdot|x)$ is $[-1, 1]$ and our results below is applicable for the support $[-M, M]$ with any $M > 0$. Moreover, let the value of $f_{\rho,\tau}(x)$ be unique at each $x \in X$. To demonstrate our main result in the general case, we first shall give the following learning rate in the special case if the quantile regression function $f_{\rho,\tau}$ lies in some smooth functional space. Its regularity is usually measured in terms of Sobolev spaces. Recall the Sobolev space $H^r(\mathbb{R}^n)$ with index $r > 0$ consisting of all functions in $L^2(\mathbb{R}^n)$ with the semi-norm $|f|_{H^r(\mathbb{R}^n)} = \left\{(2\pi)^{-n} \int_{\mathbb{R}^n} |\xi|^{2r} |\widehat{f}(\xi)|^2\right\}^{\frac{1}{2}}$ finite where $\widehat{f}$ is the Fourier transform of f defined as $\widehat{f}(\xi) = \int_{\mathbb{R}^n} f(x) e^{-i\xi\cdot x} dx$. In the sequel, ρ_X denotes the marginal distribution of ρ on X and $\widehat{f}$ denotes the projection operation on any measurable function $f : X \to \mathbb{R}$, given as

$$\widehat{f}(x) = \begin{cases} 1, & \text{if } \ f(x) \geq 1, \\ f(x), & \text{if } \ -1 < f(x) < 1, \\ -1, & \text{if } \ f(x) \leq -1. \end{cases}$$

Theorem 24.1 *Let $X \subset \mathbb{R}^n$ be a domain with Lipschitz boundary and ρ_X be the uniform distribution on X. Assume that $f_{\rho,\tau} \in H^r(X)$ for some $r > 0$, $\|f_{\rho,\tau}\|_\infty \leq 1$ and the conditional distributions $\{\rho(\cdot|x), x \in X\}$ have density functions given with $\zeta > 0$,*

$$\frac{d\rho(y|x)}{dy} = \begin{cases} \frac{\zeta+1}{2}|y - f_{\rho,\tau}(x)|^\zeta, & \text{if } \ |y - f_{\rho,\tau}(x)| \leq 1; \\ 0, & \text{otherwise.} \end{cases}$$

Take $\eta_t = -\frac{n+3r}{2n+5r}$, $\lambda = T^{-\frac{n+r}{2n+5r}}$, $\sigma = T^{-\frac{1}{2n+5r}}$ and $\epsilon_t = t^{-\beta}$ with $\beta \geq \frac{1}{2}$ then

$$\mathbb{E}_{z_1,\cdots,z_T}\left[\|\widehat{f}_{T+1} - f_{\rho,\tau}\|_{L^2_{\rho_X}}\right] \leq C^* T^{-\frac{r}{(2n+5r)(\zeta+2)}} \tag{24.4}$$

where C^* is a constant independent of T, and will be given in the proof.

Remark 24.1 Notice that the larger the index r is, the faster the projected function f_{T+1} in (24.3) converges to $f_{\rho,\tau}$. In addition, the choice of parameters λ, σ, η_t is closely related to r. Thus, the regularity of the quantile function $f_{\rho,\tau}$ is important in the learning process. The index β of the insensitive parameter characterizes the sparsity and the learning rate will not be affected if $\beta \geq \frac{1}{2}$. As the index β increases, the value of the insensitive parameter ϵ_t will decrease at each iteration t. So, it is suitable to choose $\beta = \frac{1}{2}$ in this case. Here the variance σ of the Gaussian kernel K_σ changes with the learning time T. This is reasonable since a small σ will lead to over-fitting and a large σ to under-fitting. In the above example, we are considering the quantile regression problems on a domain of $\mathbb{R}^n$, so the learning rate is poor if the dimension n is large. However, in many situations, the input space X is a

low-dimensional manifold embedded in the large-dimensional space $\mathbb{R}^n$. In such a situation, the learning rates may be greatly improved.

Now we are in a position of stating our main result in the general case. First, a noise condition on the measure ρ is given, which was introduced in [13].

Definition 24.2 Let $0 < p \leq \infty$ and $w > 0$. We say that ρ has a τ-quantile of p-average type w if there exist two functions b and a from X to $\mathbb{R}$ such that $\{ba^w\}^{-1} \in L^p_{\rho_X}$ and for any $x \in X$ and $q \in (0, a(x)]$, there hold

$$\rho(\{y : f_{\rho,\tau}(x) < y < f_{\rho,\tau}(x) + q\}|x) \geq b(x)q^w$$

and

$$\rho(\{y : f_{\rho,\tau}(x) - q < y < f_{\rho,\tau}(x)\}|x) \geq b(x)q^w. \tag{24.5}$$

This assumption can be satisfied with many common conditional distributions such as Guassian, students' t distributions and uniform distributions. In the following, we will give an example to illustrate it. More examples can be found in [2, 13].

Example 24.1 Let the conditional distributions $\{\rho(\cdot|x)\}_{x\in X}$ be a family of Gaussian distributions with a uniform variance $\tilde{\sigma} > 0$, i.e. $\frac{d\rho(y|x)}{dy} = \frac{1}{\sqrt{2\pi}\tilde{\sigma}} \exp\left\{-\frac{(y-\mu_x)^2}{2\tilde{\sigma}^2}\right\}$ where $\{\mu_x\}_{x\in X}$ are expectations of the Gaussian distributions $\{\rho(\cdot|x)\}_{x\in X}$. It is direct to calculate that $f_{\rho,\tau}(x)$ can take the value of μ_x at each $x \in X$. We also find that for any $q \in (0, \tilde{\sigma}]$, there holds

$$\begin{aligned}&\rho(\{y : f_{\rho,\tau}(x) < y < f_{\rho,\tau}(x) + q\}|x) = \rho(\{y : \mu_x < y < \mu_x + q\}|x)\\ &= \frac{1}{\sqrt{2\pi}\tilde{\sigma}} \int_{\mu_x}^{\mu_x+q} \exp\left\{-\frac{(y-\mu_x)^2}{2\tilde{\sigma}^2}\right\} dy = \frac{1}{\sqrt{2\pi}\tilde{\sigma}} \int_0^q \exp\left\{-\frac{y^2}{2\tilde{\sigma}^2}\right\} dy \geq \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}\tilde{\sigma}} q.\end{aligned}$$

Similarly, we have that $\rho(\{y : f_{\rho,\tau}(x) - q < y < f_{\rho,\tau}(x)\}|x) \geq \frac{e^{-\frac{1}{2}}}{\sqrt{2\pi}\tilde{\sigma}} q$. Thus, the measure ρ has a ∞-average type 1.

In addition, we need a condition about the continuity of the conditional distributions $\{\rho(\cdot|x)\}_{x\in X}$.

Definition 24.3 Let $s > 0$. We say that the family of conditional distributions $\{\rho(\cdot|x)\}_{x\in X}$ is Lipschitz-s if there exists a constant C_ρ such that

$$\rho(\{y : u \leq y \leq v\}|x) \leq C_\rho |u - v|^s, \quad \forall u < v \in Y,\ x \in X. \tag{24.6}$$

With these preliminaries in place, we present the following learning rates whose proof will be provided in the next section.

Theorem 24.2 *Suppose that for some $r > 0$, the quantile regression function $f_{\rho,\tau}$ is the restriction of some $\tilde{f}_{\rho,\tau} \in H^r(\mathbb{R}^n) \bigcap L^\infty(\mathbb{R}^n)$ over X, and the density function $\frac{d\rho_X}{dx}$ lies in $L^2(X)$. Let the parameters $\eta_t, \epsilon_t, \lambda, \sigma$ be of the form*

$$\eta_t = t^{-\frac{n+3r}{2n+5r}},\ \epsilon_t = t^{-\beta},\ \lambda = T^{-\frac{n+r}{2n+5r}},\ \sigma = T^{-\frac{1}{2n+5r}} \tag{24.7}$$

with $\beta \geq \max\left\{\frac{3(n+2r)}{s(2n+5r)} - 1, \frac{n+2r}{s(2n+5r)}\right\}$.

Denote $\mu := \frac{p(w+1)}{p+1}$. *If the measure* ρ *satisfies (24.5) and (24.6), then*

$$\mathbb{E}_{z_1,\cdots,z_T}\left[\|\widehat{f}_{T+1} - f_{\rho,\tau}\|_{L^{\mu}_{\rho_X}}\right] \leq C^* T^{-\frac{r}{(2n+5r)(w+1)}}. \tag{24.8}$$

Here the constant C^* *is independent of* T *and will be given in the proof.*

This theorem investigates the learning ability of the learned function $\widehat{f}_{T+1}$ that approximates the quantile regression function $f_{\rho,\tau}$ with suitable chosen parameters including the variance parameter σ and the insensitive parameters $\{\epsilon_t\}$. It shows how to adapt the variance σ in the learning process while keeping the sparsity and the learning power for the algorithm (24.3). It is also worth noticing that our leaning rate is given in a weighted L^{μ}-space by the noise condition (24.5). Our rate still holds for the generalization error (see Sect. 24.3) if the condition (24.5) is not imposed on ρ. At the end of this section, we would like to remark that the quantile regression problem considered here is fully nonparametric, so the parameters in (3) are usually unknown in advance and tuned in training processes according to various quantile regression problems. They can be chosen by a data-dependent way in training processes, e.g. cross-validation.

24.3 Error Analysis and Proofs of Main Results

In learning theory, the performance of learning algorithms is often measured by the generalization error. For the quantile regression, we define the *generalization error* for $f : X \to \mathbb{R}$ associated with the pinball loss ϕ_τ as

$$\mathscr{E}(f) = \int_Z \phi_\tau(f(x) - y)d\rho$$

and the quantile regression function $f_{\rho,\tau}$ is a minimizer of $\mathscr{E}(f)$. Meanwhile, we define the ϵ-insensitive generalization error $\mathscr{E}^\epsilon(f)$, given as $\mathscr{E}^\epsilon(f) := \int_Z \phi^\epsilon_\tau(f(x) - y)d\rho$. Our error analysis is conducted based on an error decomposition for the *excess generalization error* $\mathscr{E}(f_{T+1}) - \mathscr{E}(f_{\rho,\tau})$. To this end, we introduce the below approximation error with respect to the approximation ability of $\mathscr{H}_\sigma(X)$. In the sequel, we denote the norm $\|\cdot\|_{\mathscr{H}_\sigma(X)}$ by $\|\cdot\|_\sigma$ and $\mathscr{H}_\sigma(X)$ by $\mathscr{H}_\sigma$ for simplicity.

Definition 24.4 For any regularization parameter $\lambda > 0$, the approximation error $\mathscr{D}(\sigma, \lambda)$ of the triple (K_σ, ρ, τ) is defined by

$$\mathscr{D}(\sigma, \lambda) = \min_{f \in \mathscr{H}_\sigma}\left\{\mathscr{E}(f) - \mathscr{E}(f_{\rho,\tau}) + \frac{\lambda}{2}\|f\|^2_\sigma\right\}.$$

The regularization function is defined as

$$f_\lambda = \arg\min_{f\in\mathscr{H}_\sigma}\left\{\mathscr{E}(f)-\mathscr{E}(f_{\rho,\tau})+\frac{\lambda}{2}\|f\|_\sigma^2\right\} \quad or \quad f_\lambda = \arg\min_{f\in\mathscr{H}_\sigma}\left\{\mathscr{E}(f)+\frac{\lambda}{2}\|f\|_\sigma^2\right\}. \tag{24.9}$$

Then its associated insensitive regularization function for any $\epsilon > 0$ is

$$f_\lambda^\epsilon = \arg\min_{f\in\mathscr{H}_\sigma}\left\{\mathscr{E}^\epsilon(f)+\frac{\lambda}{2}\|f\|_\sigma^2\right\}. \tag{24.10}$$

Now, the error decomposition for $\mathscr{E}(f_{T+1})-\mathscr{E}(f_{\rho,\tau})$ can be displayed as

$$\mathscr{E}(f_{T+1})-\mathscr{E}(f_{\rho,\tau}) = \left\{\mathscr{E}(f_{T+1})-\mathscr{E}(f_\lambda)\right\}+\left\{\mathscr{E}(f_\lambda)-\mathscr{E}(f_{\rho,\tau})\right\} \le \left\{\mathscr{E}(f_{T+1})-\mathscr{E}(f_\lambda)\right\}+\mathscr{D}(\sigma,\lambda). \tag{24.11}$$

Notice the Lipschitz continuity of ϕ_τ and the property of RKHS with $\|f\|_\infty \le \|f\|_\sigma,\ \forall f\in\mathscr{H}_\sigma$. It yields that $|\mathscr{E}(f_{T+1})-\mathscr{E}(f_\lambda)| \le \|f_{T+1}-f_\lambda\|_\infty \le \|f_{T+1}-f_\lambda\|_\sigma$. So, the first term on the right-hand side of (24.11) will be handled in the sequel by means of the *sample error* $\|f_{T+1}-f_\lambda\|_\sigma$.

24.3.1 Approximation Error

For the second term $\mathscr{D}(\sigma,\lambda)$, it is associated with the approximation powers of the RKHSs induced by Gaussians with variance $\sigma > 0$. The following polynomial decay of $\mathscr{D}(\sigma,\lambda)$ under some Sobolev smoothness conditions on the function $f_{\rho,\tau}$ can be found in [18].

Lemma 24.1 *Suppose that for some $r > 0$, the quantile regression function $f_{\rho,\tau}$ is the restriction of some $\tilde{f}_{\rho,\tau}\in H^r(\mathbb{R}^n)\bigcap L^\infty(\mathbb{R}^n)$ over X, and the density function $\frac{d\rho_X}{dx}$ lies in $L^2(X)$. Then*

$$\mathscr{D}(\sigma,\lambda) \le C'(\sigma^r+\lambda\sigma^{-n}),\quad \forall\, 0<\sigma<1,\quad \lambda>0, \tag{24.12}$$

where C' is a constant independent of σ,λ.

24.3.2 Insensitive Analysis

According to the above error analysis, we need to estimate $\|f_{t+1}-f_\lambda\|_\sigma$ by iteration on $t = 1,\cdots,T$. In the iteration procedure, the function f_{t+1} is generated by updating f_t according to the sample (x_t, y_t). Here, the technical difficulty lies in the

change of the insensitive parameters ϵ_t. This can be handled by the following lemma in [15] for varying $\{\epsilon_t\}$.

Lemma 24.2 *Suppose that the family of conditional distributions* $\{\rho(\cdot|x)\}_{x\in X}$ *is Lipschitz-s satisfying (24.6). Then for any* $0 \le u < v$, *we have*

$$\|f_\lambda^u - f_\lambda^v\|_\sigma \le C\lambda^{-1}|u-v|^s. \tag{24.13}$$

If the insensitive parameters $\epsilon_t = \epsilon_1 t^{-\beta}$ *with* $\epsilon_1, \beta > 0$, *then*

$$\|f_\lambda^{\epsilon_{t-1}} - f_\lambda^{\epsilon_t}\|_\sigma \le C\lambda^{-1}t^{-(\beta+1)s}, \ \forall t \ge 2. \tag{24.14}$$

Here C *is a constant independent of* λ *and insensitive parameters.*

24.3.3 One Step-Iteration

Denote $h_t := \|f_\lambda^{\epsilon_{t-1}} - f_\lambda^{\epsilon_t}\|_\sigma$. We can get the one step iteration result as follows. To obtain optimal error bounds, we shall use the flexibility caused by some free parameters $0 < d < 2$ and $c_1 > 0$.

Lemma 24.3 *Define* $\{f_t\}$ *by (24.3). Let some constants* $0 < d < 2$ *and* $c_1 > 0$, *then*

$$\mathbb{E}_{z_t}\left(\|f_{t+1} - f_\lambda^{\epsilon_t}\|_\sigma^2\right) \le \left(1 + c_1 h_t^d - \lambda\eta_t\right)\|f_t - f_\lambda^{\epsilon_{t-1}}\|_\sigma^2 + h_t^{2-d}/c_1 + h_t^2 + 4\eta_t^2. \tag{24.15}$$

Proof First, we claim that $\|f_t\|_\sigma \le \frac{1}{\lambda}, \forall t \ge 2$. It can be easily derived from $f_1 = 0$ and the following induction by (24.3) that

$$\|f_{t+1}\|_\sigma \le (1-\lambda\eta_t)\|f_t\|_\sigma + \eta_t \le (1-\lambda\eta_t)\frac{1}{\lambda} + \eta_t = \frac{1}{\lambda}. \tag{24.16}$$

Denote $B_t := \left(\phi_\tau^{\epsilon_t}\right)'_-(f_t(x_t) - y_t)K_\sigma(x_t, \cdot) + \lambda f_t$. The online algorithm (24.3) can be written as $f_{t+1} = f_t - \eta_t B_t$. Then

$$\|f_{t+1} - f_\lambda^{\epsilon_t}\|_\sigma^2 = \|f_t - f_\lambda^{\epsilon_t}\|_\sigma^2 + \eta_t^2\|B_t\|_\sigma^2 - 2\eta_t\langle f_t - f_\lambda^{\epsilon_t}, B_t\rangle_\sigma \tag{24.17}$$

Applying the reproducing property (24.1) to part of the last term of (24.17), we have that

$$\langle f_t - f_\lambda^{\epsilon_t}, \left(\phi_\tau^{\epsilon_t}\right)'_-(f_t(x_t) - y_t)K_\sigma(x_t, \cdot)\rangle_\sigma = \left(\phi_\tau^{\epsilon_t}\right)'_-(f_t(x_t) - y_t)\left(f_t(x_t) - f_\lambda^{\epsilon_t}(x_t)\right).$$

The convexity of $\phi_\tau^{\epsilon_t}$ implies that

$$\left(\phi_\tau^{\epsilon_t}\right)'_-\left(f_t(x_t)-y_t\right)\left(f_t(x_t)-f_\lambda^{\epsilon_t}(x_t)\right)\geq \phi_\tau^{\epsilon_t}(f_t(x_t)-y_t)-\phi_\tau^{\epsilon_t}(f_\lambda^{\epsilon_t}(x_t)-y_t).$$

For the other part of the last term of (24.17), we have that

$$\langle f_t-f_\lambda^{\epsilon_t},f_t\rangle_\sigma\geq\|f_t\|_\sigma^2-\frac{1}{2}\|f_t\|_\sigma^2-\frac{1}{2}\|f_\lambda^{\epsilon_t}\|_\sigma^2=\frac{1}{2}\|f_t\|_\sigma^2-\frac{1}{2}\|f_\lambda^{\epsilon_t}\|_\sigma^2.$$

Thus, the last term of (24.17) can be bounded as

$$\langle f_t-f_\lambda^{\epsilon_t},B_t\rangle_\sigma\geq\left[\phi_\tau^{\epsilon_t}(f_t(x_t)-y_t)+\frac{\lambda}{2}\|f_t\|_\sigma^2\right]-\left[\phi_\tau^{\epsilon_t}(f_\lambda^{\epsilon_t}(x_t)-y_t)\frac{\lambda}{2}\|f_\lambda^{\epsilon_t}\|_\sigma^2\right].$$

Since f_t only depends on $z_1,\cdots,z_{t-1}$, then

$$\mathbb{E}_{z_t}\langle f_t-f_\lambda^{\epsilon_t},B_t\rangle_\sigma\geq\left[\mathscr{E}(f_t)+\frac{\lambda}{2}\|f_t\|_\sigma^2\right]-\left[\mathscr{E}(f_\lambda^{\epsilon_t})+\frac{\lambda}{2}\|f_\lambda^{\epsilon_t}\|_\sigma^2\right]$$

This together with Theorem 2 in [19], implies that $\mathbb{E}_{z_t}\langle f_t-f_\lambda^{\epsilon_t},B_t\rangle_\sigma\geq\frac{\lambda}{2}\|f_t-f_\lambda^{\epsilon_t}\|_\sigma^2$. Putting it into (24.17), then

$$\mathbb{E}_{z_t}\left(\|f_{t+1}-f_\lambda^{\epsilon_t}\|_\sigma^2\right)\leq(1-\lambda\eta_t)\|f_t-f_\lambda^{\epsilon_t}\|_\sigma^2+\eta_t^2\mathbb{E}_{z_t}\|B_t\|_\sigma^2.\tag{24.18}$$

Now we estimate $\|f_t-f_\lambda^{\epsilon_t}\|_\sigma$. It is decomposed as

$$\|f_t-f_\lambda^{\epsilon_t}\|_\sigma=\|f_t-f_\lambda^{\epsilon_{t-1}}+f_\lambda^{\epsilon_{t-1}}-f_\lambda^{\epsilon_t}\|_\sigma\leq\|f_t-f_\lambda^{\epsilon_{t-1}}\|_\sigma+h_t.$$

Applying the elementary inequality $2xy\leq c_1x^2y^d+y^{2-d}/c_1$ with any $0<d<2$ and $c_1>0$, to $x=\|f_t-f_\lambda^{\epsilon_{t-1}}\|_\sigma$ and $y=h_t$, then

$$\|f_t-f_\lambda^{\epsilon_t}\|_\sigma^2\leq\|f_t-f_\lambda^{\epsilon_{t-1}}\|_\sigma^2+2\|f_t-f_\lambda^{\epsilon_{t-1}}\|_\sigma h_t+h_t^2\leq(1+c_1h_t^d)\|f_t-f_\lambda^{\epsilon_{t-1}}\|_\sigma^2+h_t^{2-d}/c_1+h_t^2.$$

Plugging it into (24.18) and noticing that $(1-\lambda\eta_t)(1+c_1h_t^d)\leq1+c_1h_t^d-\lambda\eta_t$, we get

$$\mathbb{E}_{z_t}\left(\|f_{t+1}-f_\lambda^{\epsilon_t}\|_\sigma^2\right)\leq(1+c_1h_t^d-\lambda\eta_t)\|f_t-f_\lambda^{\epsilon_t}\|_\sigma^2+h_t^{2-d}/c_1+h_t^2+\eta_t^2\mathbb{E}_{z_t}\|B_t\|_\sigma^2.$$

We now only need to estimate $\|B_t\|_\sigma^2$. Note that $\|\left(\phi_\tau^{\epsilon_t}\right)'_-\|_\infty\leq1$ and the bound (24.16)holds for the learning sequence $\{f_t\}$. Using the reproducing property $\|K_\sigma(x_t,\cdot)\|_\sigma^2=\langle K_\sigma(x_t,\cdot),K_\sigma(x_t,\cdot)\rangle_\sigma=K_\sigma(x_t,x_t)=1$, then

$$\|B_t\|_\sigma\leq\left\|\left(\phi_\tau^{\epsilon_t}\right)'_-(f_t(x_t)-y_t)K_\sigma(x_t-\cdot)\right\|_\sigma+\lambda\|f_t\|_\sigma\leq\|\left(\phi_\tau^{\epsilon_t}\right)'_-\|_\infty\|K_\sigma(x_t,\cdot)\|_\sigma++\lambda\|f_t\|_\sigma\leq2.$$

Based on the above analysis, we can get the desired conclusion (24.15).

24.3.4 Sample Error Estimate

We are in a position to present the estimate of the sample error $\|f_{T+1} - f_\lambda\|_\sigma$, which is the key analysis in our study. For simplicity, denote $\prod_{j=T+1}^{T}\left(1-\frac{1}{2}\lambda\eta_j\right) := 1$, $\sum_{j=T+1}^{T}\lambda\eta_j := 0$ and $f_\lambda^0 := f_\lambda$.

Lemma 24.4 *Let the parameters $\eta_t, \epsilon_t, \lambda$ be of the form as $\eta_t = \eta_1 t^{-\alpha}$, $\epsilon_t = \epsilon_1 t^{-\beta}$ and $\lambda = T^{-(1-\alpha-\varepsilon)}$ for any $1-2\alpha < \varepsilon < 1-\alpha$, $\eta_1 > 0, \epsilon_1 > 0$ satisfying*

$$\max\{1-\beta s-\varepsilon, 2-(\beta+1)s-2\varepsilon\} < \alpha < \min\{2(\beta+1)s, 1\}. \tag{24.19}$$

Then we have

$$\begin{aligned}\mathbb{E}_{z_1,\cdots,z_T}\left(\|f_{T+1}-f_\lambda\|_\sigma\right) \le \tilde{C}T^{-\min\left\{(\beta+1)s+\alpha-2+2\varepsilon, \alpha-\frac{1}{2}+\frac{\varepsilon}{2}, \beta s-1+\alpha+\varepsilon\right\}} \\ + \sqrt{\frac{2\mathscr{D}(\sigma,\lambda)}{\lambda}}\exp\left\{-\frac{\lambda\eta_1}{8(1-\alpha)}(T+1)^{1-\alpha}\right\}\end{aligned} \tag{24.20}$$

where $\tilde{C}$ is a constant independent of T, given in the proof.

Proof We split $\|f_{T+1}-f_\lambda\|_\sigma$ into two parts as $\|f_{T+1}-f_\lambda^{\epsilon_T}\|_\sigma$ and $\|f_\lambda - f_\lambda^{\epsilon_T}\|_\sigma$. For the first term $\|f_{T+1}-f_\lambda^{\epsilon_T}\|_\sigma$, we shall apply the conclusion in Lemma 24.3. By (24.14), $h_t \le C\lambda^{-1}t^{-(\beta+1)s}$. We take $d = \frac{\alpha}{(\beta+1)s}$ and $c_1 = \frac{1}{2}\eta_1 C^{-d}T^{-(d+1)(1-\alpha-\varepsilon)}$ for any $1-2\alpha < \varepsilon < 1-\alpha$. The restriction (24.19) of parameters implies that $c_1 h_t^d \le \frac{1}{2}\lambda\eta_t$ and $1 + c_1 h_t^d - \lambda\eta_t \le 1 - \frac{1}{2}\lambda\eta_t$. With (24.15), it yields that

$$\mathbb{E}_{z_t}\left(\|f_{t+1}-f_\lambda^{\epsilon_t}\|_\sigma^2\right) \le \left(1-\frac{1}{2}\lambda\eta_t\right)\|f_t - f_\lambda^{\epsilon_{t-1}}\|_\sigma^2 + 2h_t^{2-d}/c_1 + 4\eta_t^2.$$

Applying the relation above iteratively for $t = t_0, \cdots, T$, we obtain that

$$\begin{aligned}&\mathbb{E}_{z_1,\cdots,z_T}\left(\|f_{T+1}-f_\lambda^{\epsilon_T}\|_\sigma^2\right) \le \left(1-\frac{1}{2}\lambda\eta_T\right)(1-\lambda\eta_{T-1})\,\mathbb{E}_{z_1,\cdots,z_{T-1}}\left(\|f_T - f_\lambda^{\epsilon_{T-1}}\|_\sigma^2\right) \\ &+ 2h_T^{2-d}/c_1 + 4\eta_T^2 + \left(1-\frac{1}{2}\lambda\eta_T\right)\left(2h_{T-1}^{2-d}/c_1 + 4\eta_{T-1}^2\right) \\ &= \prod_{t=t_0}^{T}\left(1-\frac{1}{2}\lambda\eta_t\right)\mathbb{E}_{z_1,\cdots,z_{t_0-1}}\left(\|f_{t_0} - f_\lambda^{t_0-1}\|_\sigma^2\right) + \sum_{t=t_0}^{T}\left(2h_t^{2-d}/c_1 + 4\eta_t^2\right)\prod_{j=t+1}^{T}\left(1-\frac{1}{2}\lambda\eta_j\right).\end{aligned}$$

Using the above inequality with $t_0 = 1$ and noting that $\|f_\lambda\|_\sigma^2 \le 2\mathscr{D}(\sigma,\lambda)/\lambda$, with the elementary inequality $1 - x \le e^{-x}$ for any $x > 0$, then we have

$$\mathbb{E}_{z_1,\cdots,z_T}\left(\|f_{T+1}-f_\lambda^{\epsilon_T}\|_\sigma^2\right) \le \exp\left\{-\frac{\lambda}{2}\sum_{t=1}^{T}\eta_t\right\}\|f_\lambda\|_\sigma^2 + \sum_{t=1}^{T}\left(2h_t^{2-d}/c_1+4\eta_t^2\right)\exp\left\{-\frac{\lambda}{2}\sum_{j=t+1}^{T}\eta_t\right\}$$

$$\le 2\exp\left\{-\frac{\lambda}{2}\sum_{t=1}^{T}\eta_t\right\}\mathscr{D}(\sigma,\lambda)/\lambda + \sum_{t=1}^{T}\left(2h_t^{2-d}/c_1+4\eta_t^2\right)\exp\left\{-\frac{\lambda}{2}\sum_{j=t+1}^{T}\eta_t\right\}$$

$$= 2\exp\left\{-\frac{\lambda\eta_1}{2}\sum_{t=1}^{T}t^{-\alpha}\right\}\mathscr{D}(\sigma,\lambda)/\lambda + \sum_{t=1}^{T}\left(\frac{2C^{2-d}}{c_1\lambda^{2-d}}t^{-(2-d)(\beta+1)s}+4\eta_1^2t^{-2\alpha}\right)\exp\left\{-\frac{\lambda\eta_1}{2}\sum_{j=t+1}^{T}t^{-\alpha}\right\}$$

$$:= I_1 + I_2.$$

For I_1, using the elementary inequality in Lemma 4 of [19], that for any $0<\alpha<1$, there holds $\sum_{t=1}^{T}t^{-\alpha}\ge\frac{(T+1)^{1-\alpha}-1}{1-\alpha}$, we have

$$I_1 \le \frac{2\mathscr{D}(\sigma,\lambda)}{\lambda}\exp\left\{-\frac{\lambda\eta_1}{2(1-\alpha)}\left((T+1)^{1-\alpha}-1\right)\right\} \le \frac{2\mathscr{D}(\sigma,\lambda)}{\lambda}\exp\left\{-\frac{\lambda\eta_1}{4(1-\alpha)}(T+1)^{1-\alpha}\right\}.$$

For I_2, we apply the following elementary inequality valid for $t\in\mathbb{N}, 0<q_1<1$ and $c, q_2>0$:

$$\sum_{i=1}^{t-1}i^{-q_2}\exp\left\{-c\sum_{j=i+1}^{t}j^{-q_1}\right\} \le \frac{2^{q_1+q_2}}{c}t^{q_1-q_2}+\frac{t}{2}\exp\left\{-\frac{c(1-2^{q_1-1})}{1-q_1}(t+1)^{1-q_1}\right\}. \tag{24.21}$$

It can be derived in the proof procedure of Lemma 2 (b) of [9]. Here we omit it for simplicity.

Take $q_1=\alpha$, $q_2=(2-d)(\beta+1)s$ and $c=\frac{\lambda\eta_1}{2}$. Then the first part of I_2 is bounded as

$$I_{21} := \sum_{t=1}^{T}\left(\frac{2C^{2-d}}{c_1\lambda^{2-d}}t^{-(2-d)(\beta+1)s}\right)\exp\left\{-\frac{\lambda\eta_1}{2}\sum_{j=t+1}^{T}t^{-\alpha}\right\} \le 2C^{2-d}\Big[\frac{2^{(2-d)(\beta+1)s+\alpha+1}}{\eta_1c_1\lambda^{3-d}}T^{-(2-d)(\beta+1)s+\alpha}$$

$$+\frac{T}{2c_1\lambda^{2-d}}\exp\left\{-\frac{\eta_1(1-2^{\alpha-1})\lambda}{2(1-\alpha)}(T+1)^{1-\alpha}\right\}+\frac{T^{-(2-d)(\beta+1)s}}{c_1\lambda^{2-d}}\Big].$$

Note that $\lambda=T^{-(1-\alpha-\varepsilon)}$ implies that there exists a constant C_ε independent of T such that the middle term $\frac{T}{2c_1\lambda^{2-d}}\exp\left\{-\frac{\eta_1(1-2^{\alpha-1})\lambda}{2(1-\alpha)}(T+1)^{1-\alpha}\right\}\le C_\varepsilon T^{-(2(\beta+1)s+2\alpha-4+4\varepsilon)}$. Together with the choice of d, c_1, we have that

$$I_{21} \le A_1T^{-(2(\beta+1)s+2\alpha-4+4\varepsilon)}$$

where $A_1 := 2C^{2-d}\left(\frac{2^{(2-d)(\beta+1)s+\alpha+2}}{\eta_1^2}C^d+C_\varepsilon+\frac{2C^d}{\eta_1}\right)$.

For the second part of I_2, by similarity, applying (24.21) with $q_1=\alpha, q_2=2\alpha$ and $c=\frac{\lambda\eta_1}{2}$, we have that

$$I_{22} := \sum_{t=1}^{T} 4\eta_1^2 t^{-2\alpha} \exp\left\{-\frac{\lambda\eta_1}{2}\sum_{j=t+1}^{T} t^{-\alpha}\right\} \le A_2 T^{-2\alpha+1-\varepsilon}$$

where $A_2 := 4\eta_1^2\left(\frac{2^{3\alpha+1}}{\eta_1} + C_\varepsilon + 1\right)$. Based on the above analysis, we see that

$$\begin{aligned}&\mathbb{E}_{z_1,\cdots,z_T}\left(\|f_{T+1} - f_\lambda^{\epsilon_T}\|_\sigma^2\right)\\ \le&\frac{2\mathscr{D}(\sigma,\lambda)}{\lambda}\exp\left\{-\frac{\lambda\eta_1}{4(1-\alpha)}(T+1)^{1-\alpha}\right\} + (A_1+A_2)T^{-\min\{2(\beta+1)s+2\alpha-4+4\varepsilon,2\alpha-1+\varepsilon\}}.\end{aligned}$$

For the term $\|f_\lambda - f_\lambda^{\epsilon_T}\|_\sigma$, by (24.13), it can be bounded as $\|f_\lambda - f_\lambda^{\epsilon_T}\|_\sigma \le C\epsilon_1^s T^{-s\beta}$ $\lambda^{-1} = C\epsilon_1^s T^{-s\beta+1-\alpha-\varepsilon}$. Then we can get the conclusion (24.20) with $\tilde{C} = \sqrt{A_1 + A_2} + C\epsilon_1^s$.

24.3.5 Bounding the Total Error

In our analysis we shall make use of the following comparison theorem [2, 13]. Recall that $\mu := \frac{p(w+1)}{p+1}$.

Lemma 24.5 *Suppose that the measure ρ has a p-average type w satisfying (24.5). Then for any measurable function $f : X \to [-1, 1]$, we have*

$$\|f - f_{\rho,\tau}\|_{L^\mu_{\rho_X}} \le C_\mu\left(\mathscr{E}(f) - \mathscr{E}(f_{\rho,\tau})\right)^{\frac{1}{w+1}} \tag{24.22}$$

where the constant $C_\mu = 2(w+1)^{\frac{1}{w+1}}\|(ba^w)^{-1}\|_{L^p_{\rho_X}}^{\frac{1}{w+1}}$.

Now we can present the proof of our error estimate for the convergence of online algorithm (24.3) in a general form.

Proof of Theorem 24.2 Putting the explicit form (24.7) of $\eta_t, \epsilon_t, \lambda, \sigma$ into (24.20), we know that there exists a constant C'_ϵ independent of T or τ such that

$$\sqrt{\frac{2\mathscr{D}(\sigma,\lambda)}{\lambda}}\exp\left\{-\frac{\lambda\eta_1}{8(1-\alpha)}(T+1)^{1-\alpha}\right\} \le C'_\epsilon T^{-\frac{r}{2n+5r}}$$

and

$$\min\left\{(\beta+1)s+\alpha-2+2\varepsilon, \alpha-\frac{1}{2}+\frac{\varepsilon}{2}, \beta s-1+\alpha+\varepsilon\right\} = \frac{r}{2n+5r}.$$

This yields that

$$\mathbb{E}_{z_1,\cdots,z_T}\left[\|f_{T+1}-f_\lambda\|_\sigma\right] \le \left(\tilde{C}+C'_\epsilon\right)T^{-\frac{r}{2n+5r}}.$$

By (24.11), we know that

$$\mathbb{E}_{z_1,\cdots,z_T}\left[\mathscr{E}(f_{T+1})-\mathscr{E}(f_{\rho,\tau})\right] \le \mathbb{E}_{z_1,\cdots,z_t}\left[\|f_{T+1}-f_\lambda\|_\sigma\right]+\mathscr{D}(\sigma,\lambda) \le \left(\tilde{C}+C'_\epsilon+2C'\right)T^{-\frac{r}{2n+5r}}.$$

Since the support of $\rho(\cdot|x)$ is $[-1,1]$, we have that $\phi_\tau(\widehat{f}(x)-y) \le \phi_\tau(f(x)-y)$ for any measurable function $f: X \to \mathbb{R}$. It yields that $\mathscr{E}(\widehat{f}_{T+1}) \le \mathscr{E}(f_{T+1})$ and

$$\mathbb{E}_{z_1,\cdots,z_T}\left[\mathscr{E}(\widehat{f}_{T+1})-\mathscr{E}(f_{\rho,\tau})\right] \le \left(\tilde{C}+C'_\epsilon+2C'\right)T^{-\frac{r}{2n+5r}}.$$

Using the relation (24.22), we can complete the proof of Theorem 24.2 with $C^*=\left(\tilde{C}+C'_\epsilon+2C'\right)^{\frac{1}{w+1}}C_\mu$.

Proof of Theorem 24.1 We shall prove Theorem 24.1 by Theorem 24.2. Since X has a Lipschitz boundary, we know from [10] that there exists an extension function $\tilde{f}_{\rho,\tau} \in H^r(\mathbb{R}^n)$ such that $\tilde{f}_{\rho,\tau}|_X = f_{\rho,\tau}$. Next, we check the noise condition (24.5). Let the function $a(x)=1$ and $b(x)=\frac{1}{2}$, we have that for any $q \in [0,1]$

$$\rho(\{y: f_{\rho,\tau}(x) \le y \le f_{\rho,\tau}(x)+q\}|x) = \int_{f_{\rho,\tau}(x)}^{f_{\rho,\tau}(x)+q} \frac{d\rho(y|x)}{dy}dy = \frac{1}{2}q^{\zeta+1}.$$

By similarity, we have $\rho(\{y: f_{\rho,\tau}(x)-q \le y \le f_{\rho,\tau}(x)\}|x)=\frac{1}{2}q^{\zeta+1}$. Therefore, the measure ρ has a τ-quantile of ∞-average type $\zeta+1$. Meanwhile, we find that the family of conditional distributions $\{\rho(\cdot|x)\}_{x\in X}$ is Lipschitz-1 and (24.6) is satisfied with $C_\rho=\frac{\zeta+1}{2}$ and $s=1$ since the density function $\frac{d\rho(y|x)}{dy}$ is uniformly bounded by $\frac{\zeta+1}{2}$. Thus, we can apply (24.8) to get that

$$\mathbb{E}_{z_1,\cdots,z_T}\left[\|f_{T+1}-f_{\rho,\tau}\|_{L^2_{\rho_X}}\right] \le \mathbb{E}_{z_1,\cdots,z_T}\left[\|f_{T+1}-f_{\rho,\tau}\|_{L^{\zeta+2}_{\rho_X}}\right] \le C^*T^{-\frac{r}{(2n+5r)(\zeta+2)}}.$$

Then the proof is completed.

Acknowledgments The work described in this paper is partially supported by National Natural Science Foundation of China [Nos. 11671307 and 11571078], Natural Science Foundation of Hubei Province in China [No. 2017CFB523] and the Special Fund for Basic Scientific Research of Central Colleges, South-Central University for Nationalities [No. CZY18033].

References

1. Aronszajn, N.: Theory of reproducing kernels. Tran. Am. Math. Soc. **68**(3), 337–404 (1950)
2. Hu, T., Yuan, Y.: Learning rates of regression with q-norm loss and threshold. Anal. Appl. **14**(06), 809–827 (2016)
3. Hwang, c., Shim, J.: A simple quantile regression via support vector machine. In: International Conference on Natural Computation, Springer, pp. 512–520 (2005)
4. Koenker, R., Geling, O.: Reappraising medfly longevity: a quantile regression survival analysis. J. Am. Stat. Assoc. **96**(454), 458–468 (2001)
5. Koenker, R.: Quantile Regression. Cambridge University Press, New York (2005)
6. Rosset, S.: Bi-level path following for cross validated solution of kernel quantile regression. J. Mach. Learn. Res. **10**(11), 2473–2505 (2009)
7. Shi, L., Huang, X., Tian, Z., Suykens, J.A.: Quantile regression with $l1$-regularization and Gaussian kernels. Adv. Comput. Math. **40**(2), 517–551 (2014)
8. Smale, S., Zhou, D.X.: Estimating the approximation error in learning theory. Anal. Appl. **1**(01), 17–41 (2003)
9. Smale, S., Zhou, D.X.: Online Learning with Markov Sampling. Anal. Appl. **7**(01), 87–113 (2009)
10. Stein, E.M.: Singular Integrals and Differentiability Properties of Functions. In Bulletin of the London Mathematical Society (1973)
11. Steinwart, I., Christman, A.: Sparsity of SVMs that use the ϵ-insensitive loss. In: Advances in Neural Information Processing Systems, pp. 1569–1576 2008)
12. Steinwart, I., Scovel, C., et al.: Fast rates for support vector machines using Gaussian kernels. Ann. Stat. **35**(2), 575–607 (2007)
13. Steinwart, I., Christmann, A., et al.: Estimating conditional quantiles with the help of the pinball loss. Bernoulli. **17**(1), 211–225 (2011)
14. Takeuchi, I., Le, Q.V., Sears, T.D., Smola, A.J.: Nonparametric quantile estimation. J. Mach. Learn. Res. **7**, 1231–1264 (2006)
15. Ting, H., Xiang, D.H., Zhou, D.X.: Online learning for quantile regression and support vector regression. J. Stat. Plan. Inference **142**(12), 3107–3122 (2012)
16. Vapnik, V.: The nature of statistical learning theory. Springer science & business media (2013)
17. Xiang, D.H., Hu, T., Zhou, D.X.: Approximation analysis of learning algorithms for support vector regression and quantile regression. J. Appl. Math. (2012). https://doi.org/10.1155/2012/902139
18. Xiang, D.H., Zhou, D.X.: Classification with Gaussians and Convex Loss. J. Mach. Learn. Res. **10**(10), 1447–1468 (2009)
19. Ying, Y., Zhou, D.X.: Online regularized classification algorithms. IEEE Trans. Inf. Theory. **52**(11), 4775–4788 (2006)