Springer Proceedings in Mathematics & Statistics

Michele La Rocca
Brunero Liseo
Luigi Salmaso *Editors*

Nonparametric Statistics

4th ISNPS, Salerno, Italy, June 2018

Springer

Springer Proceedings in Mathematics & Statistics

Volume 339

Springer Proceedings in Mathematics & Statistics

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

More information about this series at http://www.springer.com/series/10533

Michele La Rocca · Brunero Liseo ·
Luigi Salmaso
Editors

Nonparametric Statistics

4th ISNPS, Salerno, Italy, June 2018

Springer

Editors
Michele La Rocca
Department of Economics and Statistics
University of Salerno
Salerno, Italy

Luigi Salmaso
Department of Management
and Engineering
University of Padua
Vicenza, Italy

Brunero Liseo
Department of Methods and Models
for Economics, Terrority and Finance
Sapienza University of Rome
Rome, Italy

ISSN 2194-1009 ISSN 2194-1017 (electronic)
Springer Proceedings in Mathematics & Statistics
ISBN 978-3-030-57305-8 ISBN 978-3-030-57306-5 (eBook)
https://doi.org/10.1007/978-3-030-57306-5

Mathematics Subject Classification: 62Gxx, 62G05, 62G08, 62G09, 62G10, 62G15, 62G20, 62G35

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

Preface

This book provides a selection of papers developed from talks presented at the Fourth Conference of the International Society for Nonparametric Statistics (ISNPS), held in Salerno (Italy) June 11–15, 2018. The papers cover a wide spectrum of subjects within nonparametric and semiparametric statistics, including theory, methodology, applications, and computational aspects. Among the most common and relevant topics in the volume, we mention nonparametric curve estimation, regression smoothing, models for time series and more generally dependent data, varying coefficient models, symmetry testing, robust estimation, rank-based methods for factorial design, nonparametric and permutation solution for several different data, including ordinal data, spatial data, survival data and the joint modeling of both longitudinal and time-to-event data, permutation and resampling techniques, and practical applications of nonparametric statistics.

ISNPS was founded in 2010 "to foster the research and practice of nonparametric statistics, and to promote the dissemination of new developments in the field via conferences, books, and journal publication". ISNPS had a distinguished Advisory Committee that included R. Beran, P. Bickel, R. Carroll, D. Cook, P. Hall. R. Johnson, B. Lindsay, E. Parzen, P. Robinson, M. Rosen-blatt, G. Roussas, T. SubbaRao, and G. Wahba; an Executive Committee that comprised of M. Akritas, A. Delaigle, S. Lahiri and D. Politis; and a Council that included P. Bertail, G. Claeskens, R. Cao, M. Hallin, H. Koul, J.-P. Kreiss, T. Lee, R. Liu, W. Gonzáles Maintega, G. Michailidis, V. Panaretos, S. Paparoditis, J. Racine, J. Romo, and Q. Yao.

The 4th ISNPS conference focused on recent advances and trends in several areas of nonparametric statistics. It included 12 plenary and special invited sessions, 69 invited sessions, 30 contributed sessions, with about 450 participants from all over the world, thus promoting and facilitating the exchange of research ideas and collaboration among scientists and contributing to the further development of the field.

We would like to thank Dr. Veronika Rosteck and Dr. Tatiana Plotnikova of Springer for their support in this project. Finally, we are also extremely grateful to all Referees who reviewed the papers included in this volume, giving a constructive

feedback on a tight schedule for timely publication of the proceedings. Their valuable contribution and their efforts significantly improved the quality of this volume.

Co-editors also wish to thank Chiara Brombin for her great commitment and support in coordinating and managing the referring and editorial process.

Salerno, Italy
Rome, Italy
Vicenza, Italy

Michele La Rocca
Brunero Liseo
Luigi Salmaso
Co-Editors of the book
and Co-Chairs of the Fourth ISNPS Conference

Contents

Portfolio Optimisation via Graphical Least Squares Estimation

Saeed Aldahmani, Hongsheng Dai, Qiao-Zhen Zhang, and Marialuisa Restaino

Abstract In this paper, an unbiased estimation method called GLSE (proposed by Aldahmani and Dai [1]) for solving the linear regression problem in high-dimensional data ($n < p$) is applied to portfolio optimisation under the linear regression framework and compared to the ridge method. The unbiasedness of method helps in improving the portfolio performance by increasing its expected return and decreasing the associated risk when $n < p$, thus leading to a maximisation of the Sharpe ratio. The verification of this method is achieved through conducting simulation and data analysis studies and comparing the results with those of ridge regression. It is found that GLSE outperforms ridge in portfolio optimisation when $n < p$.

Keywords Graphical model · Linear regression · Ridge regression

1 Introduction

In the world of finance, investors usually seek to construct a portfolio that maximises expected returns and minimises their risk through diversifying and computing the correct weights of the assets in that portfolio. This weights computation can be

S. Aldahmani (✉)
Department of Statistics, College of Business and Economics, United Arab Emirates University, Al Ain, UAE
e-mail: saldahmani@uaeu.ac.ae

H. Dai
Department of Mathematical Sciences, University of Essex, Colchester CO4 3SQ, UK
e-mail: hdaia@essex.ac.uk

Q.-Z. Zhang
Institute of Statistics, Nankai University, Tianjin, China
e-mail: zhangqz@nankai.edu.cn

M. Restaino
University of Salerno, Fisciano, Italy
e-mail: mlrestaino@unisa.it

M. La Rocca et al. (eds.), *Nonparametric Statistics*, Springer Proceedings in Mathematics & Statistics 339, https://doi.org/10.1007/978-3-030-57306-5_1

achieved by what is technically known as portfolio optimisation, a problem that was addressed by Markowitz [14] through utilising a model known as Markowitz theory. The Markowitz theory for portfolio optimisation stipulates selecting portfolio weights $\boldsymbol{w}$ that minimise the risk (variance) of the portfolio return for a predetermined target return. This idea assumes that the future performance of asset returns' mean $\boldsymbol{\mu}$ and variance are known. However, in practice, these two factors are unknown and should be estimated using a historical dataset. To select an optimal portfolio, investors need to estimate the covariance matrix $\boldsymbol{\Sigma}$ of the returns and take its inverse. This is a typical inverse problem if the number of assets p is too large in relation to the return observations n; i.e. the inverse of the covariance matrix of the returns is singular. Therefore, many regularisation methods have been proposed in the literature to find covariance matrices and their inverses, such as in Bickel and Levina [2], Huang et al. [10], Wong et al. [19]. However, the estimates of these methods are biased, which might give undesirable weights for some higher return assets in portfolio.

Britten-Jones [3] utilised regression in order to find the portfolio weights as follows:

$$\boldsymbol{w} = \frac{\hat{\boldsymbol{\beta}}}{\hat{\boldsymbol{\beta}}' \mathbf{1}_v}, \tag{1}$$

where $\hat{\boldsymbol{\beta}}$ is the ordinary least squares (OLS) estimate of the coefficient parameter $\boldsymbol{\beta}$ for the linear regression model

$$\boldsymbol{y} = \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon}, \tag{2}$$

where the response $\boldsymbol{y} = \mathbf{1}_v$.

When $n < p$, the popular ordinary least square method (OLS) becomes ineffective, and this has triggered the proposal of many methods to solve this issue, such as Least Absolute Shrinkage and Selection Operator (LASSO) [18], Least Angle Regression (LARS) [7] and ridge regression [9]. However, all these methods suffer from the limitation of giving biased estimates. In addition, LASSO and LARS suffer from the problem of not selecting more than n covariates [20] and giving a sparse portfolio. Another problem with some of these methods is over-shrinking the final regression coefficients [16], which might lead to inaccuracy in portfolio weights. Apart from these methods, some other related approaches could be found in Candes and Tao [4], Meinshausen and Yu [15], DeMiguel et al. [6], Still and Kondor [17], Carrasco and Noumon [5], Fastrich et al. [8] and Lin et al. [12]. These methods, however, still give biased estimates and perhaps produce inaccurate weights for some higher return and less risk assets in the portfolio.

Aldahmani and Dai [1] proposed an unbiased estimation method called GLSE which can provide unbiased estimates for regression coefficients in high-dimensional data ($n < p$). The GLSE method is closely related to the theory of graphical models,

where least square estimation in conjunction with undirected Gaussian graphical models is implemented.

GLSE can give unbiased coefficient estimates for all assets, which helps the low-risk and high return assets maintain their correct weights in the portfolio and consequently assists in maximising their expected returns and lower the associated risk. Such an advantage will lead to increasing the Sharpe ratio and the expected rate of returns and decreasing the risk of the portfolio for both in-and-out-of-sample periods. This is particularly important upon comparison with other regularisation methods such as ridge, where the weights of low-risk and high return assets may be sharply reduced due to the method's biasedness, thus causing the portfolio's expected returns to fall down and its risk to rise. Moreover, unlike other regularisation methods which produce sparse portfolios(such as LASSO and LARS), GLSE and ridge share the advantage of generating diversified portfolios across a large number of stocks, as they produce non-sparse portfolios. This diversification of the portfolio leads to lowering the risk [13] due to the fact that when one or more sectors of the economy fail or decline, the rest of the sectors can then mitigate the significant impact of the loss caused by market fluctuations. However, due to the biasedness of ridge regression, the weights of some low-risk and high return assets may be sharply reduced, which may deprive ridge of its ability to reduce the risk through diversifying the assets. This limitation can clearly be overcome by GLSE due to its unbiasedness feature.

In the rest of the paper, graph theory and Matrices are given in Sect. 2. Section 3 presents the main methodology of GLSE and its properties. Section 4 provides the algorithm of graph structure selection. Simulation studies are given in Sect. 5, and a real data analysis is presented in Sect. 6. The study is concluded in Sect. 6.

2 Graph Theory and Matrices

2.1 Graph Theory

An undirected graph G consists of two sets, a set P and a set $\mathcal{E}$. The set P denotes the vertices representing variables and $\mathcal{E}$ is the set of edges (a subset of $P \times P$) connecting the vertices [11]. The elements in the set P are usually natural numbers, i.e. $P = 1, 2, \ldots, p$, representing the labels of random variables. If all the pairs of vertices in P in a graph G are joined by an edge, then the graph is complete. If $A \subseteq P$, the subset A induces a subgraph $G_A = (A, \mathcal{E}_A)$, where $\mathcal{E}_A = \mathcal{E} \cap (A \times A)$. The subset graph G_A is complete if it induces a complete subgraph from G. This subgraph is maximal if it cannot be extended by including one more neighbouring vertex. A complete subset that is maximal is called a clique.

2.1.1 Decomposition of a Graph

A triple (A, B, C) of disjoint subsets of the vertex set P of an undirected graph G is said to form a *decomposition* of G if $P = A \cup B \cup C$ and the following conditions hold [11]:

- B separates A from C;
- B is a complete subset of P.

An undirected graph G is considered as decomposable if it holds one of the following:

- Graph g is complete.
- There is a proper decomposition (A, B, C) into decomposable subgraphs g_{AB} and g_{BC} where B separates A from C.

Consider a sequence of sets $C_1, \ldots, C_q$ that are the subsets of the vertex set P of an undirected graph g such that $C_1 \cup \cdots \cup C_q = P$. If the following holds, then the given sequence is said to be a perfect sequence [11]:

$$S_j = C_j \cap (C_1 \cup C_2 \cup \cdots \cup C_{j-1}) \subseteq C_i,$$

where $j = 2, \ldots, q$ and $i \in \{1, \ldots, j-1\}$. The sets S_j are the separators. These orderings, if they exist, might not be unique.

2.2 Matrices

A $p \times p$ matrix $\mathbf{F}$ can be written as $(F_{kj})_{k,j\in P}$. $F \in R^p$ represent a vector. Denote $\mathbf{F}_{AB} = (F_{kj})_{k\in A, j\in B}$, a submatrix of $\mathbf{F}$. Denote $[\mathbf{F}_{AB}]^{\Gamma}$ as a $p \times p$-dimensional matrix obtained by filling up 0s, with

$$\left([\mathbf{F}_{AB}]^{\Gamma}\right)_{jk} = \begin{cases} F_{jk} & \text{if } j \in A,\ k \in B \\ 0 & \text{otherwise.} \end{cases} \tag{3}$$

Similarly, let $\mathbf{x}_A$ is a matrix only having covariates with indices in set A and $ssd_A = \mathbf{x}'_A\mathbf{x}_A$. Then $[(ssd_A)^{-1}]^{\Gamma}$ represents a $p \times p$-dimensional matrix obtained by filling up 0s, with

$$\left([(ssd_A)^{-1}]^{\Gamma}\right)_{jk} = \begin{cases} \left((ssd_A)^{-1}\right)_{jk} & \text{if } j,k \in A \\ 0 & \text{otherwise.} \end{cases} \tag{4}$$

3 The Idea of GLSE

Suppose that the graph G is decomposable and let $\mathcal{C}$ denote the set of cliques and $\mathcal{S}$ denote the set of separators [1]. Then the GLSE is given as follows:

$$\hat{\boldsymbol{\beta}} = \left[\sum_{C \in \mathcal{C}} [(ssd_C)^{-1}]^{\Gamma} - \sum_{S \in \mathcal{S}} [(ssd_S)^{-1}]^{\Gamma} \right] \mathbf{x}' \boldsymbol{y}. \tag{5}$$

For the existence of the GLSE, the following condition must hold

Condition 3.1 *The sample size* $n > \max_{C \in \mathcal{C}}\{|C|\}$.

For unbiasedness of $\hat{\boldsymbol{\beta}}$, based on Aldahmani and Dai [1], the following condition is imposed:

Condition 3.2 *Write the cliques and separators of* g *in the perfect ordering, as* $C_1, \cdots, C_q$ *and* $S_2, \cdots, S_q$, *such that*

$$\begin{aligned} \mathbf{x}_{C_1 \setminus S_2} &= \mathbf{x}_{S_2} \cdot \mathbf{r}_{S_2, C_1 \setminus S_2} + \boldsymbol{\xi}_1, \quad E(\boldsymbol{\xi}_1) = \mathbf{0}, \\ \mathbf{x}_{C_k \setminus S_k} &= \mathbf{x}_{S_k} \cdot \mathbf{r}_{S_k, C_k \setminus S_k} + \boldsymbol{\xi}_k, \quad E(\boldsymbol{\xi}_k) = \mathbf{0}, \quad k = 2, \cdots, q, \end{aligned}$$

where $\mathbf{r}_{S_k, C_k \setminus S_k}$ *are constant matrices with dimensions* $|s_k \times (c_k - s_k)|$;

Under Conditions 3.1 and 3.2, Aldahmani and Dai [1] show that the above estimator is unbiased;

$$\mathbb{E}(\hat{\boldsymbol{\beta}}) = \boldsymbol{\beta}.$$

4 Model Selection

A stepwise selection algorithm has been used by Aldahmani and Dai [1] to find which graph G is the best for the data. The method considers adding/deleting edges one by one to/from the current graph. When an edge under consideration is not in the current graph, it will be added if the addition makes an improvement in terms of the predetermined criteria; otherwise it will not be added. The same applies to the case of edge deletion. According to Aldahmani and Dai [1], the best graph is given by minimising a target function $\mathbb{T}(\boldsymbol{\beta}, g, \lambda_g)$:

$$(\hat{\boldsymbol{\beta}}, \hat{g}, \hat{\lambda}_g) = \arg \min_{\boldsymbol{\beta}, g \in \mathcal{G}, \lambda_g} \mathbb{T}(\boldsymbol{\beta}, g, \lambda_g) \tag{6}$$

$$\mathbb{T}(\boldsymbol{\beta}, g, \lambda_g) = ||\boldsymbol{y} - \mathbf{x}\boldsymbol{\beta}||^2 + \lambda_g |\mathcal{E}_g|, \tag{7}$$

where $\mathcal{G}$ is the set of all possible graphs, λ_g is a penalty term and $|\mathcal{E}_g|$ is the number of edges in graph G. The following pseudocode is the algorithm used by Adlahmani and Dai [1] to find the optimal graph that best fits the data:

Algorithm 1 Pseudocode of the GLSE graph selection

1: Start graph $g = (P, \mathcal{E})$, which can be an empty (or a given decomposable) graph such that $n > \max_{C \in \mathcal{C}} |C|$.
2: Generate all possible graphs, g_i, such that there is only one edge difference between g_i and the current graph g. All such g_i are decomposable and $n > \max_{C \in \mathcal{C}} |C|$.
3: Find the graph g_i^* and the associated $\hat{\boldsymbol{\beta}}$ such that g_i^* minimises the target function $\mathbb{T}(.)$ (given in (7)).
4: Go to step 2 with the selected graph g_i^* and iterate until the best one is found.
5: Output g and $\hat{\boldsymbol{\beta}}$.

It is worth noting that step 2 of Algorithm 1 can be improved significantly via parallel computation.

5 Simulation Study

The aims of this simulation study are to investigate the performance of GLSE in constructing a saturated optimal portfolio compared to ridge. The graph structure for the covariates used in generating the dataset under this simulation study is presented in Fig. 1.

This simulation involves a total of $n = 48$ observations corresponding to $p = 60$ variables derived from multivariate normal distribution, with mean 0.01 and variance covariance matrix Σ, where 36 observations are used for the in-sample period through estimating the portfolio's weight and performance (Sharpe ratios, expected returns and risk), and the remaining observations are used to find the performance of the portfolio for the out-of-sample period. The true weight of the portfolio $\boldsymbol{w}$ is derived based on the true covariance matrix Σ.

Table 1 gives the means of 500 simulated data for the in-and-out-of-sample portfolio's Sharpe ratios, expected returns and risk. It shows that out of the 500 simulated data, the GLSE yields higher means of the portfolio's Sharpe ratio and lower risk than the ridge does for the out-of-sample period. However, for the in-sample period, the ridge gives higher means of the portfolio's expected returns than the GLSE does. It should be noted that the ridge portfolio's risk is very high compared to this under the GLSE. In addition, the in-sample portfolio's Sharpe ratio is negative for the ridge but positive for the GLSE, which is desirable in finance.

The computational burden for the proposed algorithm is not too heavy with modern parallel computing technology. The computational times for one run of the above

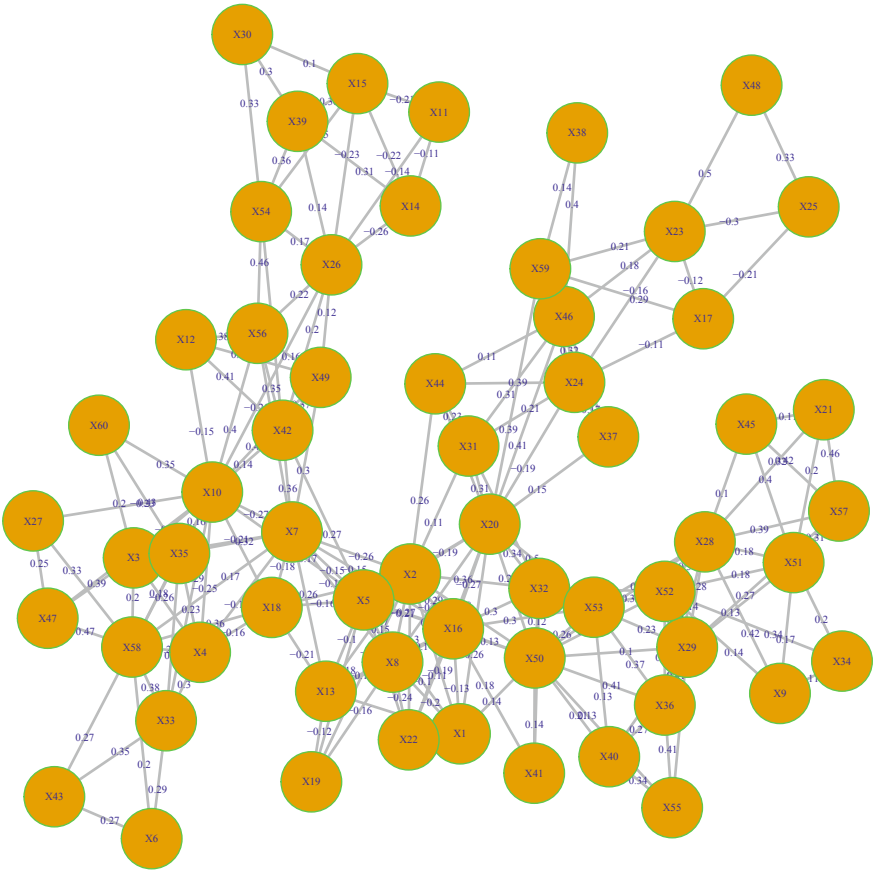

Fig. 1 Graph structure for covariates under the simulation study

Table 1 The in-and-out-of-sample portfolio's Sharpe ratios, expected returns and risk from the simulated data

	Ridge		GLSE	
	In sample	Out of sample	In sample	Out of sample
Sharpe ratio	−0.007	0.005	0.733	0.570
Expected returns	0.149	0.030	0.127	0.107
Portfolio's risk	1.282	1.236	0.526	0.516

Table 2 Portfolio size and in- and out-of sample portfolio's Sharpe ratios, expected returns and risk find by the ridge and GLSE

Portfolio size	Methods	Sharpe ratio	Expected returns	Portfolio's risk
150 stocks (in sample)	Ridge	0.719	0.097	0.134
	GLSE	2.023	0.061	0.030
150 stocks (out of sample)	Ridge	−0.074	−0.010	0.135
	GLSE	0.117	0.015	0.130
200 stocks (in sample)	Ridge	0.792	0.056	0.071
	GLSE	0.963	0.046	0.047
200 stocks (out of sample)	Ridge	0.150	0.013	0.086
	GLSE	0.224	0.015	0.068

generated datasets for both serial and parallel computing are considered. It is noted that on a machine with 8 GB of memory and 3.60 GHz processor, the time taken is approximately 20 min. When the parallel processing was used, with 5 cores, the computational time reduced to approximately 2 min.

6 Data Analysis

Monthly returns of 875 stocks listed on the New York Stock Exchange (NYSE) covering the period from 02/12/2007 to 02/12/2017 are downloaded from Datastream. Out of these stocks, 150 and 200 stocks are selected at random. Then, ridge and GLSE are applied to construct two portfolios for the selected stocks. The in-sample period for the above constructed portfolios is from 02/12/2007 to 01/12/2016. The out-of-sample period, on the other hand, is from 02/12/2016 to 01/12/2017. For ridge, cross validation is used for obtaining the penalty parameter. The in-and-out-of sample average returns, risk and Sharpe ratio are used to evaluate the performance of the obtained portfolios. The results are shown in Table 2 and they reveal that the GLSE method performs better than ridge in term of average returns, risk and the Sharpe ratio of portfolios for both in-and-out-of-sample periods.

7 Conclusion

The unbiased GLSE method was used in this paper to construct a saturated optimal portfolio in high-dimensional data ($n < p$). The results of applying this method were compared to those of ridge and they showed that GLSE outperforms ridge in terms of its ability to reduce the portfolio's risk and increase its expected returns, consequently maximising the Sharpe ratio. While both ridge and GLSE have practical implications

in the world of finance in that they both lead to a non-sparse portfolio with diversified assets, the GLSE overcomes ridge's shortcoming where the weights of low-risk and high return assets may be reduced due to its biasedness. Due to its unbiasedness, GLSE thus maintains the higher weights of low-risk and high return assets, which, as a result, minimises the chances of risk increase and income reduction in the portfolio.

References

1. Aldahmani, S., Dai, H.: Unbiased estimation for linear regression when $n<v$. Int. J. Stat. Probab. **4**(3), p61 (2015)
2. Bickel, P.J., Levina, E.: Regularized estimation of large covariance matrices. Ann. Stat. 199–227 (2008)
3. Britten-Jones, M.: The sampling error in estimates of mean-variance efficient portfolio weights. J. Finan. **54**(2), 655–671 (1999)
4. Candes, E., Tao, T.: The dantzig selector: statistical estimation when p is much larger than n. Ann. Stat. **35**(6), 2313–2351 (2007)
5. Carrasco, M., Noumon, N.: Optimal portfolio selection using regularization. Technical report, University of Montreal (2011)
6. DeMiguel, V., Garlappi, L., Nogales, F.J., Uppal, R.: A generalized approach to portfolio optimization: improving performance by constraining portfolio norms. Manage. Sci. **55**(5), 798–812 (2009)
7. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R., et al.: Least angle regression. Ann. Stat. **32**(2), 407–499 (2004)
8. Fastrich, B., Paterlini, S., Winker, P.: Constructing optimal sparse portfolios using regularization methods. Comput. Manage. Sci. **12**(3), 417–434 (2015)
9. Hoerl, A.E., Kennard, R.W.: Ridge regression: Biased estimation for nonorthogonal problems. Technometrics **12**(1), 55–67 (1970)
10. Huang, J.Z., Liu, N., Pourahmadi, M., Liu, L.: Covariance matrix selection and estimation via penalised normal likelihood. Biometrika **93**(1), 85–98 (2006)
11. Lauritzen, S.L.: Graphical Models. Oxford University Press, Oxford (1996)
12. Lin, W., Shi, P., Feng, R., Li, H.: Variable selection in regression with compositional covariates. Biometrika asu 031 (2014)
13. Malkiel, B.G., Xu, Y.: Risk and return revisited. J. Portfolio Manage. **23**(3), 9–14 (1997)
14. Markowitz, H.: Portfolio selection. J. Finan. **7**(1), 77–91 (1952)
15. Meinshausen, N., Yu, B.: Lasso-type recovery of sparse representations for high-dimensional data. Ann. Stat. 246–270 (2009)
16. Radchenko, P., James, G.M., et al.: Improved variable selection with forward-lasso adaptive shrinkage. Ann. Appl. Stat. **5**(1), 427–448 (2011)
17. Still, S., Kondor, I.: Regularizing portfolio optimization. New J. Phys. **12**(7), 075034 (2010)
18. Tibshirani, R.: Regression shrinkage and selection via the lasso: a retrospective. J. R. Stat. Soc. Ser. B (Statistical Methodology) **58**(1), 267–288 (2011)
19. Wong, F., Carter, C.K., Kohn, R.: Efficient estimation of covariance selection models. Biometrika **90**(4), 809–830 (2003)
20. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Stat. Soc. Ser. B (Statistical Methodology) **67**(2), 301–320 (2005)

Change of Measure Applications in Nonparametric Statistics

Mayer Alvo

Abstract Neyman [7] was the first to propose a change in measure in the context of goodness of fit problems. This provided an alternative density to the one for the null hypothesis. Hoeffding introduced a change of measure formula for the ranks of the observed data which led to obtaining locally most powerful rank tests. In this paper, we review these methods and propose a new approach which leads on the one hand to new derivations of existing statistics. On the other hand, we exploit these methods to obtain Bayesian applications for ranking data.

Keywords Ranks · Change of measure · Bayesian methods

Mathematics Subject Classification (2010) 62F07 · 62G86 · 62H11

1 Introduction

In a landmark paper, [7] considered the nonparametric goodness of fit problem and introduced the notion of smooth tests of fit by proposing a parametric family of alternative densities to the null hypothesis. In this article, we describe a number of applications of this change of measure. Hence, we obtain a new derivation of the well-known Friedman statistic as the locally most powerful test in an embedded family of distributions.

2 Smooth Models

Suppose that the probability mass function of a discrete k-dimensional random vector $\boldsymbol{X}$ is given by

M. Alvo (✉)
Department of Mathematics, Statistics University of Ottawa, Ottawa, ON, Canada
e-mail: malvo@uottawa.ca

M. La Rocca et al. (eds.), *Nonparametric Statistics*, Springer Proceedings in Mathematics & Statistics 339, https://doi.org/10.1007/978-3-030-57306-5_2

$$\pi\left(\boldsymbol{x}_j;\boldsymbol{\theta}\right)=\exp\left(\boldsymbol{\theta}'\boldsymbol{x}_j-K\left(\boldsymbol{\theta}\right)\right)p_j,\ \ j=1,\ldots,m, \tag{1}$$

where $\boldsymbol{x}_j$ is the jth value of $\boldsymbol{X}$ and $\boldsymbol{p}=\left(p_j\right)'$ denotes the vector of probabilities when $\boldsymbol{\theta}=\boldsymbol{\theta}_0$. Here $K\left(\boldsymbol{\theta}\right)$ is a normalizing constant for which

$$\sum_j \pi\left(\boldsymbol{x}_j;\boldsymbol{\theta}\right)=1.$$

We see that the model in (1) prescribes a change of measure from the null to the alternative hypothesis. Let $\boldsymbol{T}=[\boldsymbol{x}_i,\ldots,\boldsymbol{x}_m]$ be the $k\times m$ matrix of possible vector values of $\boldsymbol{X}$. Then under the distribution specified by $\boldsymbol{p}$,

$$\begin{aligned}\boldsymbol{\Sigma}\equiv Cov_{\boldsymbol{p}}\left(\boldsymbol{X}\right)&=E_{\boldsymbol{p}}\left[\left(\boldsymbol{X}-E\left[\boldsymbol{X}\right]\right)\left(\boldsymbol{X}-E\left[\boldsymbol{X}\right]\right)'\right] &(2)\\ &=\boldsymbol{T}\left(diag\left(\boldsymbol{p}\right)\right)\boldsymbol{T}'-\left(\boldsymbol{T}\boldsymbol{p}\right)\left(\boldsymbol{T}\boldsymbol{p}\right)', &(3)\end{aligned}$$

where the expectations are with respect to the model (1). This particular situation arises often when dealing with the nonparametric randomized block design. Define

$$\boldsymbol{\pi}\left(\boldsymbol{\theta}\right)=\left(\pi\left(\boldsymbol{x}_1;\boldsymbol{\theta}\right),\ldots,\pi\left(\boldsymbol{x}_m;\boldsymbol{\theta}\right)\right)'$$

and suppose that we would like to test

$$H_0:\boldsymbol{\theta}=\boldsymbol{0}\ \text{vs}\ H_1:\boldsymbol{\theta}\neq\boldsymbol{0}.$$

Letting $\boldsymbol{N}$ denote a multinomial random vector with parameters $\left(n,\boldsymbol{\pi}\left(\boldsymbol{\theta}\right)\right)$, we see that the log likelihood as a function of $\boldsymbol{\theta}$ is, apart from a constant, proportional to

$$\begin{aligned}\sum_{j=1}^{m} n_j\log\left(\pi\left(\boldsymbol{x}_j;\boldsymbol{\theta}\right)\right)&=\sum_{j=1}^{m} n_j\left(\boldsymbol{\theta}'\boldsymbol{x}_j-K\left(\boldsymbol{\theta}\right)\right)\\ &=\boldsymbol{\theta}'\left(\sum_{j=1}^{m} n_j\boldsymbol{x}_j\right)-nK\left(\boldsymbol{\theta}\right).\end{aligned}$$

The score vector under the null hypothesis is then given by

$$\begin{aligned}U\left(\boldsymbol{\theta};\boldsymbol{X}\right)&=\sum_{j=1}^{m} N_j\left(\frac{1}{\pi_j\left(\boldsymbol{\theta}\right)}\frac{\partial\pi_j\left(\boldsymbol{\theta}\right)}{\partial\boldsymbol{\theta}}\right)\\ &=\boldsymbol{T}\left(\boldsymbol{N}-n\boldsymbol{p}\right).\end{aligned}$$

Under the null hypothesis,

$$E\left[U\left(\boldsymbol{\theta};\boldsymbol{X}\right)\right]=0,$$

and the score statistic is given by

$$\frac{1}{n}\left[\boldsymbol{T}(\boldsymbol{N}-n\boldsymbol{p})\right]'\boldsymbol{\Sigma}^{-1}\left[\boldsymbol{T}(\boldsymbol{N}-n\boldsymbol{p})\right]=\frac{1}{n}(\boldsymbol{N}-n\boldsymbol{p})'\left(\boldsymbol{T}'\boldsymbol{\Sigma}^{-1}\boldsymbol{T}\right)(\boldsymbol{N}-n\boldsymbol{p})\xrightarrow{\mathcal{L}}\chi_r^2, \tag{4}$$

where $r=rank\left(\boldsymbol{T}'\boldsymbol{\Sigma}^{-1}\boldsymbol{T}\right)$.

In the one-sample ranking problem whereby a group of judges are each asked to rank a set of t objects in accordance with some criterion, let $\mathcal{P}=\left\{\boldsymbol{\nu}_j,\, j=1,\ldots,t!\right\}$ be the space of all $t!$ permutations of the integers $1,2,\ldots,t$ and let the probability mass distribution defined on $\mathcal{P}$ be given by

$$\boldsymbol{p}=(p_1,\ldots,p_{t!}),$$

where $p_j=\Pr\left(\boldsymbol{\nu}_j\right)$. Conceptually, each judge selects a ranking $\boldsymbol{\nu}$ in accordance with the probability mass distribution $\boldsymbol{p}$. In order to test the null hypothesis that each of the rankings are selected with equal probability, that is,

$$H_0:\boldsymbol{p}=\boldsymbol{p}_0 \text{ vs } H_1:\boldsymbol{p}\neq\boldsymbol{p}_0, \tag{5}$$

where $\boldsymbol{p}_0=\frac{1}{t!}\mathbf{1}$, define a k-dimensional vector score function $\boldsymbol{X}(\boldsymbol{\nu})$ on the space $\mathcal{P}$ and following (1), let its smooth probability mass function be given as

$$\pi(\boldsymbol{x}_j;\boldsymbol{\theta})=\exp\left(\boldsymbol{\theta}'\boldsymbol{x}_j-K(\boldsymbol{\theta})\right)\frac{1}{t!},\quad j=1,\ldots,t! \tag{6}$$

where $\boldsymbol{\theta}$ is a t-dimensional vector, $K(\boldsymbol{\theta})$ is a normalizing constant and $\boldsymbol{x}_j$ is a t-dimensional score vector to be specified in (8). Since

$$\sum_{j=1}^{t!}\pi\left(\boldsymbol{x}_j;\boldsymbol{\theta}\right)=1$$

it can be seen that $K(\mathbf{0})=0$ and hence the hypotheses in (5) are equivalent to testing

$$H_0:\boldsymbol{\theta}=\mathbf{0} \text{ vs } H_1:\boldsymbol{\theta}\neq\mathbf{0}. \tag{7}$$

It follows that the log likelihood function is proportional to

$$l(\boldsymbol{\theta})\sim n\left[\boldsymbol{\theta}'\hat{\boldsymbol{\eta}}-K(\boldsymbol{\theta})\right],$$

where

$$\hat{\boldsymbol{\eta}}=\left[\sum_{j=1}^{t!}\boldsymbol{x}_j\hat{p}_{nj}\right],\ \hat{p}_{nj}=\frac{n_j}{n}$$

and n_j represents the number of observed occurrences of the ranking $\boldsymbol{\nu}_j$. The Rao score statistic evaluated at $\boldsymbol{\theta} = \mathbf{0}$ is

$$\begin{aligned} U(\boldsymbol{\theta}; \boldsymbol{X}) &= n \frac{\partial}{\partial \theta} \left[\boldsymbol{\theta}' \hat{\boldsymbol{\eta}} - K(\mathbf{0}) \right] \\ &= n \left[\hat{\boldsymbol{\eta}} - \frac{\partial}{\partial \theta} K(\mathbf{0}) \right], \end{aligned}$$

whereas the information matrix is

$$\boldsymbol{I}(\boldsymbol{\theta}) = -n \left[\frac{\partial^2}{\partial \theta^2} K(\mathbf{0}) \right].$$

The test then rejects the null hypothesis whenever

$$n^2 \left[\hat{\boldsymbol{\eta}} - \frac{\partial}{\partial \boldsymbol{\theta}} K(\mathbf{0}) \right]' \boldsymbol{I}^{-1}(\mathbf{0}) \left[\hat{\boldsymbol{\eta}} - \frac{\partial}{\partial \boldsymbol{\theta}} K(\mathbf{0}) \right] > \chi_f^2(\alpha),$$

where $\chi_f^2(\alpha)$ is the upper $100(1-\alpha)\,\%$ critical value of a chi square distribution with $f = \text{rank}(\boldsymbol{I}(\boldsymbol{\theta}))$ degrees of freedom. We note that the test just obtained is the locally most powerful test of H_0.

Specializing this test statistic to the Spearman score function of adjusted ranks

$$\boldsymbol{x}_j = \left(\nu_j(1) - \frac{t+1}{2}, \ldots, \nu_j(t) - \frac{t+1}{2} \right)', \; j = 1, \ldots, t!, \tag{8}$$

we can show that the Rao score statistic is the well-known Friedman test [5].

$$W = \frac{12n}{t(t+1)} \sum_{i=1}^{t} \left[\bar{R}_i - \frac{t+1}{2} \right]^2, \tag{9}$$

where $\bar{R}_i$ is the average of the ranks assigned to the ith object.

2.1 The Two-Sample Ranking Problem

The approach just described can be used to deal with the two-sample ranking problem assuming again the Spearman score function. Let $\boldsymbol{X}_1, \boldsymbol{X}_2$ be two independent random vectors whose distributions as in the one sample case are expressed for simplicity as

$$\pi\left(\boldsymbol{x}_j; \boldsymbol{\theta}_l\right) = \exp\left\{ \boldsymbol{\theta}_l' \boldsymbol{x}_j - K(\boldsymbol{\theta}_l) \right\} p_l(j), \; j = 1, \ldots, t!, l = 1, 2,$$

where $\boldsymbol{\theta}_l = (\theta_{l1}, \ldots, \theta_{lt})'$ represents the vector of parameters for population l. We are interested in testing

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 \text{ vs } H_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2.$$

The probability distribution $\{p_l\,(j)\}$ represents an unspecified null situation. Define

$$\hat{\boldsymbol{p}}_l = \left(\frac{n_{l1}}{n_l}, \ldots, \frac{n_{lt!}}{n_l}\right)',$$

where n_{ij} represents the number of occurrences of the ranking $\boldsymbol{\nu}_j$ in sample l.

Also, for $l = 1, 2$, set $\sum_j n_{ij} \equiv n_l$, $\boldsymbol{\gamma} = \boldsymbol{\theta}_1 - \boldsymbol{\theta}_2$ and

$$\boldsymbol{\theta}_l = \boldsymbol{m} + b_l\boldsymbol{\gamma},$$

where

$$\boldsymbol{m} = \frac{n_1\boldsymbol{\theta}_1 + n_2\boldsymbol{\theta}_2}{n_1 + n_2}, b_1 = \frac{n_{2,}}{n_1 + n_2}, b_2 = -\frac{n_1}{n_1 + n_2}.$$

Let $\boldsymbol{\Sigma}_l$ be the covariance matrix of $\boldsymbol{X}_l$ under the null hypothesis defined as

$$\boldsymbol{\Sigma}_l = \boldsymbol{\Pi}_l - \boldsymbol{p}_l\boldsymbol{p}_l',$$

where $\boldsymbol{\Pi}_l = diag\,(p_l\,(1), \ldots, p_l\,(t!))$ and $\boldsymbol{p}_l = (p_l\,(1), \ldots, p_l\,(t!))'$. The logarithm of the likelihood L as a function of $(\boldsymbol{m}, \boldsymbol{\gamma})$ is proportional to

$$\log\, L\,(\boldsymbol{m}, \boldsymbol{\gamma}) \sim \sum_{l=1}^{2}\sum_{j=1}^{t!} n_{lj}\left\{(\boldsymbol{m} + b_l\boldsymbol{\gamma})'\,\boldsymbol{x}_j - K\,(\boldsymbol{\theta}_l)\right\}.$$

In order to test

$$H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2 \text{ vs } H_1 : \boldsymbol{\theta}_1 \neq \boldsymbol{\theta}_2$$

we calculate the Rao score test statistic which is given by

$$n\left(\boldsymbol{T}_S\hat{\boldsymbol{p}}_1 - \boldsymbol{T}_S\hat{\boldsymbol{p}}_2\right)'\hat{\boldsymbol{D}}\left(\boldsymbol{T}_S\hat{\boldsymbol{p}}_1 - \boldsymbol{T}_S\hat{\boldsymbol{p}}_2\right). \tag{10}$$

It can be shown to have asymptotically a χ_f^2 whenever $n_l/n \to \lambda_l > 0$ as $n \to \infty$, where $n = n_1 + n_2$. Here $\hat{\boldsymbol{D}}$ is the Moore–Penrose inverse of $\boldsymbol{T}_S\hat{\boldsymbol{\Sigma}}\boldsymbol{T}_S'$ and $\hat{\boldsymbol{\Sigma}}$ is a consistent estimator of $\boldsymbol{\Sigma} = \frac{\boldsymbol{\Sigma}_1}{\lambda_1} + \frac{\boldsymbol{\Sigma}_2}{\lambda_2}$ and f is the rank of $\hat{\boldsymbol{D}}$, as required.

2.2 The Use of Penalized Likelihood

In the previous sections, it was possible to derive test statistics for the one and two-sample ranking problems by means of the change of measure paradigm. This paradigm may be exploited to obtain new results for the ranking problems. Specifically, we consider a negative penalized likelihood function defined to be the negative log likelihood function subject to a constraint on the parameters which is then minimized with respect to the parameter. This approach yields further insight into ranking problems.

For the one-sample ranking problem, let

$$\Lambda(\boldsymbol{\theta}, c) = -\boldsymbol{\theta}' \left[\sum_{j=1}^{t!} n_j \boldsymbol{x}_j \right] + nK(\boldsymbol{\theta}) + \lambda \left(\sum_{i=1}^{t} \theta_i^2 - c \right) \tag{11}$$

represent the penalizing function for some prescribed values of the constant c. We shall assume for simplicity that $\|\boldsymbol{x}_j\| = 1$. When t is large (say $t \geq 10$), the computation of the exact value of the normalizing constant $K(\boldsymbol{\theta})$ involves a summation of $t!$ terms. [6] noted the resemblance of (6) to the continuous von Mises-Fisher density

$$f(\boldsymbol{x}; \boldsymbol{\theta}) = \frac{\|\boldsymbol{\theta}\|^{\frac{t-3}{2}}}{2^{\frac{t-3}{2}} t! I_{\frac{t-3}{2}}(\|\boldsymbol{\theta}\|) \Gamma(\frac{t-1}{2})} \exp\left(\boldsymbol{\theta}' \boldsymbol{x}\right),$$

where $\|\boldsymbol{\theta}\|$ is the norm of $\boldsymbol{\theta}$ and $\boldsymbol{x}$ is on the unit sphere and $I_\upsilon(z)$ is the modified Bessel function of the first kind given by

$$I_\upsilon(z) = \sum_{k=0}^{\infty} \frac{1}{\Gamma(k+1)\Gamma(\upsilon + k + 1)} \left(\frac{z}{2}\right)^{2k+\nu}.$$

This seems to suggest the approximation of the constant $K(\boldsymbol{\theta})$ by

$$\exp(-K(\boldsymbol{\theta})) \approx \frac{1}{t!} \cdot \frac{\|\boldsymbol{\theta}\|^{\frac{t-3}{2}}}{2^{\frac{t-3}{2}} I_{\frac{t-3}{2}}(\|\boldsymbol{\theta}\|) \Gamma(\frac{t-1}{2})}.$$

In [1], penalized likelihood was used in ranking situations to obtain further insight into the differences between groups of rankers.

3 Bayesian Models for Ranking Data

The fact that the model in (1) is itself parametric in nature leads one to consider an extension to Bayesian considerations. Let $\boldsymbol{R} = (R(1), \ldots, R(t))'$ be a ranking t items, labeled $1, \ldots, t$ and define the standardized rankings as

$$\boldsymbol{y} = \left(\boldsymbol{R} - \frac{t+1}{2}\mathbf{1}\right) / \sqrt{\frac{t(t^2-1)}{12}},$$

where $\boldsymbol{y}$ is the $t \times 1$ vector with $\|\boldsymbol{y}\| \equiv \sqrt{\boldsymbol{y}'\boldsymbol{y}} = 1$. We consider the following more general ranking model:

$$\pi(\boldsymbol{y}|\kappa, \boldsymbol{\theta}) = C(\kappa, \boldsymbol{\theta}) \exp\left\{\kappa \boldsymbol{\theta}' \boldsymbol{y}\right\},$$

where the parameter $\boldsymbol{\theta}$ is a $t \times 1$ vector with $\|\boldsymbol{\theta}\| = 1$, parameter $\kappa \geq 0$, and $C(\kappa, \boldsymbol{\theta})$ is the normalizing constant. This model has a close connection to the distance-based models considered in [3]. Here, $\boldsymbol{\theta}$ is a real-valued vector, representing a consensus view of the relative preference of the items from the individuals. Since both $\|\boldsymbol{\theta}\| = 1$ and $\|\boldsymbol{y}\| = 1$, the term $\boldsymbol{\theta}'\boldsymbol{y}$ can be seen as $\cos\phi$ where ϕ is the angle between the consensus score vector $\boldsymbol{\theta}$ and the observation $\boldsymbol{y}$. The probability of observing a ranking is proportional to the cosine of the angle from the consensus score vector. The parameter κ can be viewed as a concentration parameter. For small κ, the distribution of rankings will appear close to a uniform whereas for larger values of κ, the distribution of rankings will be more concentrated around the consensus score vector. We call this new model an *angle-based ranking model*.

To compute the normalizing constant $C(\kappa, \boldsymbol{\theta})$, let P_t be the set of all possible permutations of the integers $1, \ldots, t$. Then

$$(C(\kappa, \boldsymbol{\theta}))^{-1} = \sum_{\boldsymbol{y} \in \mathcal{P}} \exp\left\{\kappa \boldsymbol{\theta}^T \boldsymbol{y}\right\}. \tag{12}$$

Notice that the summation is over the $t!$ elements in $\mathcal{P}$. When t is large, say greater than 15, the exact calculation of the normalizing constant is prohibitive. Using the fact that the set of $t!$ permutations lie on a sphere in $(t-1)$-space, our model resembles the continuous von Mises-Fisher distribution, abbreviated as $vMF(\boldsymbol{x}|\boldsymbol{m}, \kappa)$, which is defined on a $(p-1)$ unit sphere with mean direction $\boldsymbol{m}$ and concentration parameter κ:

$$p(\boldsymbol{x}|\kappa, \boldsymbol{m}) = V_p(\kappa) \exp(\kappa \boldsymbol{m}'\boldsymbol{x}),$$

where

$$V_p(\kappa) = \frac{\kappa^{\frac{p}{2}-1}}{(2\pi)^{\frac{p}{2}} I_{\frac{p}{2}-1}(\kappa)},$$

and $I_a(\kappa)$ is the modified Bessel function of the first kind with order a. Consequently, we may approximate the sum in (12) by an integral over the sphere:

$$C(\kappa, \boldsymbol{\theta}) \simeq C_t(\kappa) = \frac{\kappa^{\frac{t-3}{2}}}{2^{\frac{t-3}{2}} t! I_{\frac{t-3}{2}}(\kappa) \Gamma(\frac{t-1}{2})},$$

where $\Gamma(.)$ is the gamma function. In ([9], it is shown that this approximation is very accurate for values of κ ranging from 0.01 to 2 and t ranging from 4 to 11. Moreover, the error drops rapidly as t increases. Note that this approximation allows us to approximate the first and second derivatives of $\log\ C$ which can facilitate our computation in what follows.

3.1 Maximum Likelihood Estimation (MLE) of Our Model

Let $\boldsymbol{Y} = \left\{\boldsymbol{y}_1, \ldots, \boldsymbol{y}_N\right\}$ be a random sample of N standardized rankings drawn from $p(\boldsymbol{y}|\kappa, \boldsymbol{\theta})$. The log likelihood of $(\kappa, \boldsymbol{\theta})$ is then given by

$$l(\kappa, \boldsymbol{\theta}) = n \log C_t(\kappa) + \sum_{i=1}^{n} \kappa \boldsymbol{\theta}' \boldsymbol{y}_i. \tag{13}$$

Maximizing (13) subject to $\|\boldsymbol{\theta}\| = 1$ and $\kappa \geq 0$, we find that the maximum likelihood estimator of $\boldsymbol{\theta}$ is given by $\hat{\boldsymbol{\theta}}_{MLE} = \frac{\sum_{i=1}^{N} \boldsymbol{y}_i}{\left\|\sum_{i=1}^{N} \boldsymbol{y}_i\right\|}$, and $\hat{\kappa}$ is the solution of

$$A_t(\kappa) \equiv \frac{-C_t^{'}(\kappa)}{C_t(\kappa)} = \frac{I_{\frac{t-1}{2}}(\kappa)}{I_{\frac{t-3}{2}}(\kappa)} = \frac{\left\|\sum_{i=1}^{N} \boldsymbol{y}_i\right\|}{N} \equiv r. \tag{14}$$

A simple approximation to the solution of (14) following [4] is given by

$$\hat{\kappa}_{MLE} = \frac{r(t - 1 - r^2)}{1 - r^2}.$$

A more precise approximation can be obtained from a few iterations of Newton's method. Using the method suggested by [8], starting from an initial value κ_0, we can recursively update κ by iteration:

$$\kappa_{i+1} = \kappa_i - \frac{A_t(\kappa_i) - r}{1 - A_t(\kappa_i)^2 - \frac{t-2}{\kappa_i} A_t(\kappa_i)},\ i = 0, 1, 2, \ldots.$$

3.2 One-Sample Bayesian Method with Conjugate Prior

Taking a Bayesian approach, we consider the following conjugate prior for $(\kappa, \boldsymbol{\theta})$ as

$$p(\kappa, \boldsymbol{\theta}) \propto [C_t(\kappa)]^{\nu_0} \exp\left\{\beta_0 \kappa \boldsymbol{m}_0' \boldsymbol{\theta}\right\}, \tag{15}$$

where $\|\boldsymbol{m}_0\| = 1, \nu_0, \beta_0 \geq 0$. Given $\boldsymbol{y}$, the posterior density of $(\kappa, \boldsymbol{\theta})$ can be expressed by

$$p(\alpha, \boldsymbol{\theta}|\boldsymbol{y}) \propto \exp\left\{\beta\kappa \boldsymbol{m}'\boldsymbol{\theta}\right\} V_t(\beta\kappa) \cdot \frac{[C_t(\kappa)]^{N+\nu_0}}{V_t(\beta\kappa)},$$

where $\boldsymbol{m} = \left(\beta_0 \boldsymbol{m_0} + \sum_{i=1}^N \boldsymbol{y}_i\right)\beta^{-1}, \beta = \left\|\beta_0 \boldsymbol{m}_0 + \sum_{i=1}^N \boldsymbol{y}_i\right\|$. The posterior density can be factored as

$$p(\kappa, \theta|\boldsymbol{y}) = p(\theta|\kappa, \boldsymbol{y})p(\kappa|\boldsymbol{y}), \tag{16}$$

where $p(\boldsymbol{\theta}|\kappa, \boldsymbol{y}) \sim vMF(\boldsymbol{\theta}|\boldsymbol{m}, \beta\kappa)$ and

$$p(\kappa|\boldsymbol{y}) \propto \frac{[C_t(\kappa)]^{N+\nu_0}}{V_t(\beta\kappa)} = \frac{\kappa^{\frac{t-3}{2}(\upsilon_0+N)} I_{\frac{t-2}{2}}(\beta\kappa)}{\left[I_{\frac{t-3}{2}}(\kappa)\right]^{\nu_0+N} (\beta\kappa)^{\frac{t-2}{2}}}.$$

The normalizing constant for $p(\kappa|\boldsymbol{y})$ is not available in closed form. For reasons explained in [9], we approximate the posterior distribution using the method of variational inference (abbreviated VI from here on). Variational inference provides a deterministic approximation to an intractable posterior distribution through optimization. We first adopt a joint vMF- Gamma distribution as the prior for $(\kappa, \boldsymbol{\theta})$:

$$\begin{aligned} p(\kappa, \boldsymbol{\theta}) &= p(\boldsymbol{\theta}|\kappa)p(\kappa) \\ &= vMF(\boldsymbol{\theta}|\boldsymbol{m}_0, \beta_0\kappa)\, Gamma(\kappa|a_0, b_0), \end{aligned}$$

where $Gamma(\kappa|a_0, b_0)$ is the Gamma density function with shape parameter a_0 and rate parameter b_0 (i.e., mean equal to $\frac{a_0}{b_0}$), and $p(\boldsymbol{\theta}|\kappa) = vMF(\boldsymbol{\theta}|\boldsymbol{m}_0, \beta_0\kappa)$. The choice of $Gamma(\kappa|a_0, b_0)$ for $p(\kappa)$ is motivated by the fact that for large values of κ, $p(\kappa)$ in (15) tends to take the shape of a Gamma density. In fact, for large values of κ, $I_{\frac{t-3}{2}}(\kappa) \simeq \frac{e^\kappa}{\sqrt{2\pi\kappa}}$, and hence $p(\kappa)$ becomes the Gamma density with shape $(\nu_0 - 1)\frac{t-2}{2} + 1$ and rate $\nu_0 - \beta_0$:

$$p(\kappa) \propto \frac{[C_t(\kappa)]^{\nu_0}}{V_t(\beta_0\kappa)} \propto \kappa^{(\nu_0-1)\frac{t-2}{2}} \exp(-(\nu_0 - \beta_0)\kappa).$$

Using the variational inference framework, [9] showed that the optimal posterior distribution of θ conditional on κ is a von Mises-Fisher distribution $vMF(\boldsymbol{\theta}|\boldsymbol{m}, \kappa\beta)$ where

$$\beta = \left\|\beta_0 \boldsymbol{m}_0 + \sum_{i=1}^N \boldsymbol{y}_i\right\| \quad \text{and} \quad \boldsymbol{m} = \left(\beta_0 \boldsymbol{m}_0 + \sum_{i=1}^N \boldsymbol{y}_i\right)\beta^{-1}.$$

The optimal posterior distribution of κ is a $Gamma(\kappa|a, b)$ with shape a and rate b with

$$a = a_0 + N\left(\frac{t-3}{2}\right) + \beta\bar{\kappa}\left[\frac{\partial}{\partial\beta\kappa}\ln I_{\frac{t-2}{2}}(\beta\bar{\kappa})\right], \tag{17}$$

$$b = b_0 + N\frac{\partial}{\partial\kappa}I_{\frac{t-3}{2}}(\bar{\kappa}) + \beta_0\left[\frac{\partial}{\partial\beta_0\kappa}\ln I_{\frac{t-2}{2}}(\beta_0\bar{\kappa})\right]. \tag{18}$$

Finally, the posterior mode $\bar{\kappa}$ can be obtained from the previous iteration as

$$\bar{\kappa} = \begin{cases} \frac{a-1}{b} & \text{if } a > 1, \\ \frac{a}{b} & \text{otherwise.} \end{cases} \tag{19}$$

3.3 Two-Sample Bayesian Method with Conjugate Prior

Let $\boldsymbol{Y}_i = \{\boldsymbol{y}_{i1}, \ldots, \boldsymbol{y}_{iN_i}\}$ for $i = 1, 2$, be two independent random samples of standardized rankings each drawn, respectively, from $p(\boldsymbol{y_i}|\kappa_i, \boldsymbol{\theta_i})$. Taking a Bayesian approach, we assume that conditional on κ, there are independent von Mises conjugate priors, respectively, for $(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)$ as

$$p(\boldsymbol{\theta_i}|\kappa) \propto [C_t(\kappa)]^{\nu_{i0}} \exp\left\{\beta_{i0}\kappa \boldsymbol{m}_{i0}^T\boldsymbol{\theta_i}\right\},$$

where $\|\boldsymbol{m}_{i0}\| = 1, \nu_{i0}, \beta_{i0} \geq 0$. We shall be interested in computing the Bayes factor when considering two models. Under model 1, denoted M_1, $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$ whereas under model 2, denoted M_2, equality is not assumed. The Bayes factor comparing the two models is defined to be

$$\begin{aligned} B_{21} &= \frac{\int p(\boldsymbol{y_1}|\kappa, \boldsymbol{\theta_1})p(\boldsymbol{y_2}|\kappa, \boldsymbol{\theta_2})p(\boldsymbol{\theta_1}|\kappa)p(\boldsymbol{\theta_2}|\kappa)d\boldsymbol{\theta_1}d\boldsymbol{\theta_2}d\kappa}{\int p(\boldsymbol{y_1}|\kappa, \boldsymbol{\theta})p(\boldsymbol{y_2}|\kappa, \boldsymbol{\theta})p(\boldsymbol{\theta}|\kappa)d\boldsymbol{\theta}d\kappa} \\ &= \frac{\int\left[\int p(\boldsymbol{y_1}|\kappa, \boldsymbol{\theta_1})p(\boldsymbol{\theta_1}|\kappa)d\boldsymbol{\theta_1}\right]\left[\int p(\boldsymbol{y_2}|\kappa, \boldsymbol{\theta_2})p(\boldsymbol{\theta_2}|\kappa)d\boldsymbol{\theta_2}\right]d\kappa}{\int p(\boldsymbol{y_1}|\kappa, \boldsymbol{\theta})p(\boldsymbol{y_2}|\kappa, \boldsymbol{\theta})p(\boldsymbol{\theta}|\kappa)d\boldsymbol{\theta}d\kappa}. \end{aligned}$$

The Bayes factor enables us to compute the posterior odds of model 2 to model 1. We fist deal with the denominator in B_{21}. Under M_1, we assume a joint von Mises-Fisher prior on θ and a Gamma prior on κ:

$$p(\theta, \kappa) = vMF(\theta|m_0, \beta_0\kappa)\, G(\kappa|a_0, b_0).$$

Hence,

$$\begin{aligned} \int p(\boldsymbol{y_1}|\kappa, \boldsymbol{\theta})p(\boldsymbol{y_2}|\kappa, \boldsymbol{\theta})p(\boldsymbol{\theta}|\kappa)d\boldsymbol{\theta}d\kappa &= \int C_t^N(\kappa)\, exp\left\{\beta\kappa\boldsymbol{\theta}^T\boldsymbol{m}\right\} V_t(\beta_0\kappa)\, G(\kappa|a_0, b_0)\, d\boldsymbol{\theta}d\kappa \\ &= \int C_t^N(\kappa)\, V_t(\beta_0\kappa)\, V_t^{-1}(\beta\kappa)\, G(\kappa|a_0, b_0)\, d\kappa, \end{aligned}$$