Statistical Methods for Data With Latent Structures

Statistical Methods for Data with Latent Structures

A DISSERTATION PRESENTED
BY
DINGDONG YI
TO
THE DEPARTMENT OF STATISTICS

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY
IN THE SUBJECT OF
STATISTICS

HARVARD UNIVERSITY
CAMBRIDGE, MASSACHUSETTS
FEBRUARY 2018

Thesis advisor: Professor Jun S. Liu Dingdong Yi

Statistical Methods for Data with Latent Structures

ABSTRACT

This dissertation develops statistical methods to study and utilize the latent structure of data. Here, the latent structure of our interest include but are not limited to latent heterogeneity of rank data, latent seasonal components of univariate time series data, as well as latent factors and change-points of multivariate time series data. We build all the models from a Bayesian perspective, and develop different types of statistical inferences tailed for different motivations and purposes of real data applications. This dissertation contains three self-contained chapters.

Chapter 1 studies rank aggregation problem with covariates and heterogeneous rankers. We propose the Bayesian Aggregation of Rank-data with Covariates (BARC) and its extensions not only to obtain a complete aggregated ranking list, but also to study individual reliability and overall consistency of rankers. In specific, the two extensions consider varying qualities and heterogeneous ranking opinions of rankers, respectively. We developed efficient full Bayesian inference via parameter-expanded Gibbs sampler. Simulation studies show the superior performance of our methods to other existing methods in a variety of scenarios. We finally exploit our proposed method to solve real-data problems in sports and medical studies.

Chapter 2 studies the forecasting of unemployment initial claims with the help of Internet search data. We presents a novel statistical method, Penalized Regression with Inferred Seasonality Module (PRISM) to better forecast (including nowcast) unemployment initial claims weeks into future. Our method PRISM is semi-parametric, as it collectively considers a wide range of parametric time series models. We introduce a general state space formulation that contains a variety of widely used time series models as special cases, and a joint model with Internet search data to put all contemporaneous time series into a same system. We then derive a universal predictive model for forecasting initial claim data from our general formulation, and develop a

two-stage estimation procedure using nonparametric seasonal decomposition and L_1 penalized regression. PRISM outperforms all alternatives in out-of-sample testing.

Chapter 3 introduces a Bayesian factor model with multiple change-points in the quest for estimating time-varying covariance of high-dimensional time series. Under the high-dimensional setting, we exploit spike-and-slab LASSO prior on factor loadings such that the estimated factor loading matrix is sparse and interpretable. On top of factor model, we consider piecewise stationary distributions for the factors to accommodate the change over time. We then proposed an efficient EM algorithm to estimate posterior mode of our proposed model by taking advantage of L_1 regularized regression and algorithms for exact change-point detection. The number of factors and the number of change-points are considered unknown and inferred coherently from observed data and our model specification. In the application to real data examples, our method delivers highly interpretable latent factor and meaningful change-points.

Contents

1 Bayesian Aggregation of Rank Data with Covariates and Heterogeneous Rankers **1**

1.1 Introduction . . . 2
1.2 Bayesian models for rank data with covariates . . . 9
1.3 MCMC computation with parameter expansion . . . 19
1.4 Rank aggregation via MCMC samples . . . 23
1.5 Simulation studies . . . 25
1.6 Analyses of the two real data sets . . . 31
1.7 Discussion . . . 37

2 Forecasting unemployment using Internet search data **40**

2.1 Introduction . . . 40
2.2 State space formulation for time series with seasonal pattern . . . 48
2.3 Joint model with exogenous time series . . . 54
2.4 Forecasting with PRISM . . . 56
2.5 Application to unemployment initial claim data . . . 64
2.6 Summary . . . 71

3 Bayesian Factor Model with Multiple Change-points **73**

3.1 Introduction . . . 73
3.2 Bayesian change-point model for covariance matrix . . . 76
3.3 Bayesian factor model with multiple change-points . . . 80
3.4 EM approach to factor analysis with change-points . . . 84

3.5 Simulation studies . . . 89
3.6 Real data examples . . . 92
3.7 Discussion . . . 97

APPENDIX A APPENDIX TO CHAPTER 1 **99**
A.1 Proof for the consistency of MLE . . . 99
A.2 Validity of the parameter-expanded Gibbs sampler . . . 102
A.3 Gibbs sampler for BARCW . . . 104
A.4 Detailed step 2 in Gibbs sampling of BARCM . . . 104
A.5 Rank aggregation methods in comparison . . . 105
A.6 MCMC Diagnostic . . . 107

APPENDIX B APPENDIX TO CHAPTER 2 **110**
B.1 Proof of propositions . . . 110
B.2 Robustness to seasonal decomposition method choice . . . 114
B.3 Effect of the discount factor . . . 115
B.4 Coefficient heatmap . . . 115

APPENDIX C APPENDIX TO CHAPTER 3 **117**
C.1 The detailed E-step . . . 118
C.2 The detailed M-step . . . 119

REFERENCES **131**

Author list

Chapter 1. Xinran Li and Prof. Jun S. Liu contributed to this chapter. I was the lead on all aspects of writing and research for this chapter.

Chapter 2. Prof. Samuel S.C. Kou contributed to this chapter. I was the lead on all aspects of writing and research for this chapter.

Chapter 3. Prof. Jun S. Liu contributed to this chapter. I was the lead on all aspects of writing and research for this chapter.

TO MY PARENTS.

Acknowledgments

First, I want to thank the members of my dissertation committee, Jun Liu, Samuel Kou, and Mark Glickman, for their invaluable mentorship and support through my doctoral study. I am particularly grateful to Jun Liu and Samuel Kou, as they lead as great academic role models, motivate me to think bigger, and push me forward when I laid back.

In addition, I am thankful to all other Harvard statistics faculty, especially Carl Morris, Joseph Blitzstein, Luke Miratrix and Tirthankar Dusgupta, for teaching me, offering me guidance and encouragement. I appreciate all the help of the Harvard Statistics staff, especially Betsey Cogswell, James Matejek, Madeleine Straubel and Kathleen Cloutier, who make the department a lovely place.

During my graduate study, I have great friendship developed inside this department. Through numerous thoughtful discussions, I gained knowledge and ideas from my fellow PhD students, especially my longtime friend and exceptional collaborator Xinran Li. I also want to thank all my friends who I share the ups and downs with, especially Shaoyang Ning, Espen Bernton, Luis Campos, Haoxin Li and Peter Xu.

Last but not least, I want to thank my classmate, friend and partner Ruobin Gong for the love and support.

1

Bayesian Aggregation of Rank Data with Covariates and Heterogeneous Rankers

1.1 Introduction

Combining ranking results from different sources is a common problem. Well-known rank aggregation problems range from the election problem back in 18th century (Borda, 1781) to search engine results aggregation in modern days (Dwork et al., 2001). In this paper, we tackle the problem of rank aggregation with relevant covariates of the ranked entities, as explained in detail in the following two applications.

Example 1 (NFL Quarterback Ranking) *During the National Football League (NFL) season, experts from different websites, such as* `espn.com` *and* `nfl.com`*, provide weekly ranking lists of players by position. For example, Table 1.1 shows the ranking lists of the NFL starting quarterbacks from 13 experts in week 12 of season 2014. The ranking lists can help football fans better predict the performance of the quarterbacks in the coming week and even place bets in online fantasy sports games. After collecting ranking lists from the experts, the websites mostly aggregate them using arithmetic means. Besides rankings, the summary statistics of the NFL players are also available online. For example, Table 1.2 shows the statistics of the ranked*

quarterbacks prior to week 12 of season 2014. Not surprisingly, in addition to watching football games, the experts may also use these summary statistics when ranking quarterbacks.

Table 1.1: Ranking lists of NFL starting quarterbacks from 13 different experts, as of week 12 in the 2014 season.

Player	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_8	τ_9	τ_{10}	τ_{11}	τ_{12}	τ_{13}
Andrew Luck	1	1	1	3	3	1	1	1	1	1	1	1	1
Aaron Rodgers	2	3	4	2	1	2	3	3	2	2	3	4	3
Peyton Manning	3	2	5	4	2	3	2	2	3	4	4	2	2
Tom Brady	4	7	3	5	4	5	4	6	4	3	6	8	4
Tony Romo	9	5	6	1	5	4	5	4	5	5	7	6	6
Drew Brees	10	4	2	8	9	7	7	5	7	6	2	3	5
Ben Roethlisberger	6	8	7	7	7	6	6	10	6	7	5	7	7
Ryan Tannehill	5	6	13	6	11	8	8	7	9	9	8	5	8
Matthew Stafford	8	9	11	13	8	9	9	8	8	8	9	9	9
Mark Sanchez	22	10	9	9	16	10	10	9	10	10	12	12	12
Russell Wilson	12	13	17	10	10	12	11	12	11	12	11	14	15
Philip Rivers	7	14	15	20	6	17	17	11	16	15	14	10	10
Cam Newton	18	12	8	17	19	11	14	14	14	16	21	13	14
Eli Manning	17	–	18	19	14	19	12	13	12	13	16	23	11
Matt Ryan	21	17	19	15	20	15	15	15	13	11	20	21	13
Andy Dalton	15	–	14	–	17	14	16	20	15	14	19	22	16
Alex Smith	16	11	21	16	18	18	18	16	20	21	13	11	17
Colin Kaepernick	11	16	16	11	12	16	21	17	19	18	22	16	21
Joe Flacco	24	15	12	14	24	13	13	18	18	20	15	15	19
Jay Culter	13	18	10	12	13	21	19	19	17	17	23	20	18
Josh McCown	14	19	22	18	15	22	22	21	21	19	18	17	23
Drew Stanton	20	20	–	22	22	20	20	23	22	22	10	19	20
Teddy Bridgewater	23	21	20	21	23	23	23	22	23	24	17	18	22
Brian Hoyer	19	–	–	–	21	24	24	24	24	23	24	24	24

Source: http://fantasy.nfl.com/research/rankings, http://www.fantasypros.com/nfl/rankings/qb.php.

In Example 1, the primary goal is to obtain an aggregated ranking list of all players, which is hoped to be more precise than the simple method using arithmetic means. In particular, we want to incorporate the covariates (i.e., the summary statistics here) of the players to improve the accuracy of rank aggregation. Moreover, according to Table 1.1, most of the experts give very similar ranking lists, with a few exceptions such as experts 4 and 5. Therefore, it is also important to discern the varying qualities of the rankers, in order to diminish the effect of low-quality rankers and make the aggregation results more robust.

Table 1.2: Relevant statistics of the ranked quarterbacks, prior to week 12 of the 2014 NFL season. From left to right, the statistics stand for: number of games played; pass completion percentage; passing attempts per game; average passing yards per attempt; touchdown percentage; intercept percentage; running attempts per game; running yards per attempt; running first down percentage.

Player	G	Pct	Att	Avg	Yds	TD	Int	RAtt	RAvg	RYds	R1st
Andrew Luck	11	63.40	42.20	7.80	331.00	6.30	2.20	4.20	4.20	17.50	30.40
Aaron Rodgers	11	66.70	31.10	8.60	268.80	8.80	0.90	2.50	6.40	16.20	50.00
Peyton Manning	11	68.10	40.20	8.00	323.50	7.70	2.00	1.50	-0.50	-0.70	0.00
Tom Brady	11	65.00	37.90	7.20	272.50	6.20	1.40	1.70	0.70	1.30	21.10
Tony Romo	10	68.80	29.50	8.50	251.90	7.50	2.00	1.50	2.50	3.70	20.00
Drew Brees	11	70.30	42.00	7.60	317.40	4.80	2.40	1.70	2.80	4.90	26.30
Ben Roethlisberger	11	68.30	37.50	7.90	297.30	5.80	1.50	1.90	1.10	2.10	19.00
Ryan Tannehill	11	66.10	35.40	6.60	234.70	5.10	2.10	3.70	6.70	25.10	36.60
Matthew Stafford	11	58.80	37.70	7.10	267.50	3.10	2.40	2.80	2.00	5.60	16.10
Mark Sanchez	4	62.30	36.50	8.10	296.80	4.80	4.10	3.50	0.60	2.00	7.10
Russell Wilson	11	63.60	28.50	7.10	202.70	4.50	1.60	7.60	7.70	58.50	45.20
Philip Rivers	11	68.30	33.00	7.80	257.70	6.10	2.50	2.50	2.50	6.40	25.00
Cam Newton	10	58.60	33.30	7.20	239.20	3.60	3.00	6.40	4.60	29.30	37.50
Eli Manning	11	62.30	36.90	7.00	257.50	5.20	3.00	0.80	3.80	3.10	33.30
Matt Ryan	11	65.10	38.50	7.20	278.70	4.50	2.10	1.60	4.30	7.10	33.30
Andy Dalton	11	62.40	30.70	7.10	219.40	3.60	3.00	3.80	2.50	9.50	33.30
Alex Smith	11	65.10	29.70	6.80	201.00	4.00	1.20	3.20	5.50	17.40	25.70
Colin Kaepernick	11	61.70	31.50	7.50	237.70	4.30	1.70	6.80	4.50	30.50	22.70
Joe Flacco	11	63.20	34.10	7.40	251.30	4.80	2.10	2.00	1.70	3.40	45.50
Jay Cutler	11	66.80	36.40	7.10	256.80	5.50	3.00	2.90	3.90	11.30	28.10
Josh McCown	6	60.40	30.30	7.40	225.00	3.80	4.40	2.70	5.80	15.30	50.00
Drew Stanton	6	53.60	25.20	7.10	178.20	3.30	2.00	3.00	2.00	6.00	22.20
Teddy Bridgewater	8	60.30	32.80	6.40	211.10	2.30	2.70	3.50	4.60	16.10	32.10
Brian Hoyer	11	55.90	33.20	7.80	260.40	3.00	2.20	1.80	0.90	1.50	20.00

Source: http://www.nfl.com/stats.

Example 2 (Orthodontics treatment evaluation ranking) *In 2009, 69 orthodontics experts were invited by the School of Stomatology at Peking University to evaluate the post-treatment conditions of 108 medical cases (Song et al., 2015). In order to make the evaluation easier for experts, cases were divided into 9 groups, each containing 12 cases. For each group of the cases, each expert evaluated the conditions of all cases and provided a within-group ranking list, mostly based on their personal experiences and judgments of the patients' teeth records. In the meantime, using each case's plaster model, cephalometric radiograph and photograph, the School of Stomatology located key points, measured their distances and angles that are considered to be relevant features for diagnosis, and summarized these features in terms of peer*

assessment rating (PAR) index (Richmond et al., 1992). Table 1.3 shows 15 of the 69 ranking lists for two groups, and Table 1.4 shows the corresponding features for these two groups.

Table 1.3: Ranking lists for Groups A and H, two of the 9 groups in Example 2

	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_8	τ_9	τ_{10}	τ_{11}	τ_{12}	τ_{13}	τ_{14}	τ_{15}
A1	1	3	5	2	4	1	1	2	5	5	10	8	2	4	2
A2	11	5	10	9	9	12	9	7	11	12	4	7	5	6	5
A3	6	10	8	11	11	8	11	8	12	9	6	11	12	11	11
A4	3	2	4	3	1	4	2	10	1	6	8	2	1	1	1
A5	9	4	7	5	6	6	6	5	3	3	2	5	11	7	9
A6	10	9	3	6	5	11	5	9	6	7	3	1	6	8	7
A7	8	8	11	7	12	9	12	11	8	10	7	9	8	12	12
A8	4	1	1	4	3	2	4	4	2	1	1	6	3	2	6
A9	2	12	9	8	8	5	7	3	9	8	11	12	7	5	8
A10	7	11	6	10	10	7	8	6	7	11	9	3	10	9	4
A11	5	7	2	1	2	3	10	1	10	2	5	4	9	3	3
A12	12	6	12	12	7	10	3	12	4	4	12	10	4	10	10
H1	4	8	5	8	4	11	4	3	8	9	4	4	3	11	8
H2	1	2	4	5	2	7	2	2	1	2	1	1	2	2	1
H3	2	3	2	2	1	4	1	1	2	1	6	5	5	3	3
H4	3	4	3	4	3	3	3	4	3	4	7	7	1	1	2
H5	12	12	12	12	12	12	12	12	12	12	10	12	12	9	12
H6	6	5	1	1	6	2	7	5	7	3	5	3	7	4	6
H7	8	11	6	9	10	9	11	11	10	11	11	11	6	7	10
H8	11	6	8	3	7	1	6	6	6	6	8	8	4	8	9
H9	5	7	10	11	5	10	10	10	11	8	2	6	10	12	4
H10	10	9	9	7	9	5	5	7	5	7	12	9	11	5	7
H11	9	10	7	10	11	8	9	8	9	10	9	10	8	6	11
H12	7	1	11	6	8	6	8	9	4	5	3	2	9	10	5

The rank aggregation problem emerges naturally in Example 2 because the average perception of experienced orthodontists is considered the cornerstone of systems for the evaluation of orthodontic treatment outcome as described in Song et al. (2014). However, Example 2 contains many "local" rankings among non-overlapping subgroups, and thus differs from Example 1 and most prevailing rank aggregation applications. Having been demonstrated to be associated with ranking outcomes by Song et al. (2015), the covariates information not only helps in improving ranking accuracy, but also is crucial for generating full ranking lists.

Table 1.4: Below are 11 covariates measured based on peer assessment rating (PAR) index. From left to right, the statistics stand for: Upper right segment; Upper anterior segment; Upper left segment; Lower right segment; Lower anterior segment; Lower left segment; Right buccal occlusion; Left buccal occlusion; Overjet; Overbit; Centerline.

	d1m	d2m	d3m	d4m	d5m	d6m	rbom	lbom	ojmm	obm	clm
A1	1.56	0.22	1.44	1.00	0.00	1.22	0.00	0.33	0.00	0.00	0.00
A2	1.33	0.22	1.00	0.33	0.00	0.33	0.00	0.33	0.00	0.33	0.00
A3	1.22	0.33	1.00	0.67	0.11	1.44	0.00	0.00	0.00	0.00	0.00
A4	0.00	0.00	0.11	1.78	0.22	1.89	0.33	0.67	0.00	0.00	0.00
A5	1.33	0.22	0.78	1.22	0.11	1.67	0.33	0.00	0.78	0.00	0.00
A6	1.11	0.56	1.78	0.89	0.22	0.89	0.67	1.00	0.78	0.00	0.00
A7	1.22	0.67	1.89	0.89	0.11	1.00	0.67	0.33	0.67	0.00	0.00
A8	1.44	0.22	1.56	0.89	0.22	0.56	2.00	2.00	0.00	0.00	0.00
A9	1.11	0.33	1.22	0.44	0.00	1.00	2.33	0.67	0.00	0.00	0.00
A10	0.67	0.11	0.89	0.11	0.00	0.00	0.67	1.00	0.00	0.67	0.00
A11	0.67	0.89	1.00	0.67	1.33	2.44	1.33	1.00	0.11	0.00	0.67
A12	0.67	0.11	0.22	1.00	0.00	0.56	0.33	1.33	0.00	0.33	0.00
H1	0.67	0.22	0.78	1.67	0.56	0.78	0.67	0.00	0.78	0.00	0.00
H2	1.56	0.56	0.22	0.44	0.00	0.11	0.00	0.67	0.00	0.00	0.00
H3	0.56	0.22	1.00	0.33	0.11	0.78	0.00	0.67	0.00	0.33	0.00
H4	0.56	0.22	0.67	0.44	0.11	0.44	0.67	1.00	0.00	0.00	0.00
H5	1.22	0.33	0.67	0.44	0.00	0.33	1.00	0.67	0.33	0.00	0.00
H6	0.56	0.11	1.33	1.22	0.00	1.33	1.00	0.67	0.22	0.00	0.00
H7	0.56	0.33	0.78	0.78	0.00	1.22	2.00	1.33	0.44	0.33	0.00
H8	0.78	0.22	1.56	0.89	0.00	0.33	1.67	2.00	0.00	0.00	0.00
H9	0.44	0.22	1.00	0.00	0.11	0.11	1.00	0.00	0.00	0.00	0.00
H10	1.11	0.33	1.78	0.22	0.22	0.33	1.33	1.67	0.00	0.00	0.00
H11	0.67	0.67	1.00	0.67	0.56	0.56	1.00	1.00	0.11	0.00	0.00
H12	1.22	0.78	1.00	0.33	0.33	0.67	1.00	0.67	0.56	0.00	0.00

Moreover, the individual reliability and overall consistency of these orthodontists (or rankers) are critical concerns prior to rank aggregation (Liu et al., 2012; Song et al., 2014). There could be heterogeneous quality or opinions among rankers as evidenced by the ranking discrepancies in Table 1.3. For example, the ranking position of case A9 from the listed 15 experts ranges from 2 to 12. Therefore, Example 2 presents a rank aggregation problem with covariates information and heterogeneous rankers.

Related work and main contributions

There are mainly two types of methods dealing with rank data. The first type tries to find an aggregated ranking list that is consistent with most input rankings according

to some criteria. For example, Borda (1781) aggregated rankings based on the arithmetic mean of ranking positions, commonly known as Borda count; and Van Erp & Schomaker (2000) studied several variants of Borda count. Dwork et al. (2001) proposed to aggregate rankings based on the stationary distributions of certain Markov chains, which are constructed heuristically based on the ranking lists; and DeConde et al. (2006) and Lin (2010) extended this approach to fit more complicated situations. Lin & Ding (2009) obtained the aggregated ranking list by minimizing its total distance to all the input ranking lists, an idea that can be traced back to the Mallows model (Mallows, 1957).

The second type of methods builds statistical models to characterize the data generating process of the rank data and uses the estimated models to generate the aggregated ranking list (Critchlow et al., 1991; Marden, 1996; Alvo & Yu, 2014). The most popular model for rank data is the Thurstone order statistics model, which includes the Thurstone–Mosteller–Daniels model (Thurstone, 1927; Mosteller, 1951; Daniels, 1950) and Plackett–Luce model (Luce, 1959; Plackett, 1975) as special cases. Together with variants and extensions (Benter, 1994; Böckenholt, 1992), the Thurston model family has been successfully applied to a wide range of problems (e.g., Gormley & Murphy, 2006, 2008a; Johnson et al., 2002; Gray-Davies et al., 2016). Briefly, the Thurstone model assumes that there is an underlying evaluation score for each entity, whose noisy version determines the rankings. In the Thurstone–Mosteller–Daniels and Plackett–Luce models, the noises are assumed to follow the normal and Gumbel distributions, respectively. The Plackett-Luce model can be equivalently viewed as a multistage model that models the ranking process sequentially, where each entity has a unique parameter representing its probability of being selected at each stage up to a normalizing constant.

Challenges arise in the analysis of ranking data when (a) rankers are of different qualities or belong to different opinion groups; (b) covariates information are available for either rankers or the ranked entities or both; and (c) there are incomplete ranking lists. Gormley & Murphy (2006, 2008a,b, 2010) developed the finite mixture of Plackett–Luce models and Benter models (Benter, 1994) to accommodate heterogeneous subgroups of rankers, where both the mixing proportion and group specific parameters can depend on the covariates of rankers. Böckenholt (1993) introduces the finite mixture of Thurstone models to allow for heterogeneous subgroups of rankers; Yu (2000) attempts to incorporate the covariates information for both ranked entities and rankers; Johnson et al. (2002) examines qualities of several known subgroups of rankers; and Lee et al. (2014) represents qualities of rankers by letting them have different noise levels. See Böckenholt (2006) for a review of developments in Thurstonian-based analysis, as well as some further extensions. Recently, Deng et al. (2014) proposed a Bayesian approach that can distinguish high-quality rankers from low-quality ones, and Bhowmik & Ghosh (2017) proposed a method that utilizes covariates of ranked entities to assess qualities of all rankers.

We here employ the Thurstone–Mosteller–Daniels model and its extensions because they are flexible enough to deal with incomplete ranking list and can provide a unified framework to accommodate covariate information of ranked entities, rankers with different qualities, and heterogeneous subgroups of rankers. In particular, we use the Dirichlet process prior for the mixture subgroups of rankers, which can automatically determine the total number of mixture components. Moreover, in contrast to focusing on inferring parameters of Thurstone models in most previous studies, we focus mainly on the rank aggregation and the uncertainty evaluation of the resulting aggregated ranking lists.

Computationally, the estimation for the Thurstone model is generally difficult due to the complicated form of the likelihood function, especially when there are a large number of ranked entities. To overcome the difficulty, Maydeu-Olivares (1999) transformed the estimation problem to a one involving mean and covariance structures with dichotomous indicators, Yao & Böckenholt (1999) proposed a Bayesian approach based on Gibbs sampler, and Johnson (2013) advocated the JAGS software to implement the Bayesian posterior sampling. We here develop a parameter-expanded Gibbs sampler (Liu & Wu, 1999), which facilitates group moves of the latent variables, to further improve the computational efficiency. As demonstrated in the numerical studies, the improvement of the new sampler over the standard one is significant.

The rest of this article is organized as follows. Section 2.2 elaborates on our Bayesian models for rank data with covariates. Section 2.3 provides details of our Markov Chain Monte Carlo algorithms. Section 1.4 introduces multiple analysis tools using MCMC samples. Section 2.4 displays simulation results to validate our approaches. Section 1.6 describes the two real-data applications using the proposed methods. Section 2.6 concludes with a short discussion.

1.2 Bayesian models for rank data with covariates

1.2.1 Notation and definitions

Let $\mathcal{U}$ be the set of all entities in consideration, and let $n = |\mathcal{U}|$ be the total number of entities in $\mathcal{U}$. We use $i_1 \succ i_2$ to denote that entity i_1 is ranked higher than entity i_2. A *ranking list* τ is a set of non-contradictory pairwise relations in $\mathcal{U}$, which gives rise to ordered preference lists for entities in $\mathcal{U}$. We call τ a *full ranking list* if τ identifies all pairwise relations in $\mathcal{U}$, otherwise a *partial ranking list*. When τ is a full ranking

list, we can equivalently write τ as $\tau = [i_1 \succ i_2 \succ \ldots \succ i_n]$ for notational simplicity, and further define $\tau(i)$ as the position of an entity $i \in \mathcal{U}$. Specifically, a high ranked element has a small-numbered position in the list, i.e. $\tau(i_1) < \tau(i_2)$ if and only if $i_1 \succ i_2$. Furthermore, for any vector $\boldsymbol{z} = (z_1, \ldots, z_n)' \in \mathbb{R}^n$, we use $\text{rank}(\boldsymbol{z}) = [i_1 \succ i_2 \succ \ldots \succ i_n]$ to denote the full ranking list of z_i's in a decreasing order, i.e., $z_{i_1} \geq \ldots \geq z_{i_n}$.

As introduced in Examples 1 and 2, we also observe some covariates of ranked entities. Let $\boldsymbol{x}_i \in \mathbb{R}^p$ be the p dimensional covariate vector of ranked entity i, and let $\boldsymbol{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n)' \in \mathbb{R}^{n \times p}$ be the covariate matrix for all n entities. For clarification, in the following discussion we use index i for ranked entities and index j for rankers, with n and m denoting the total numbers of ranked entities and rankers, respectively.

1.2.2 Full ranking lists without covariates

Suppose we have m full ranking lists $\tau_1, \tau_2, \ldots, \tau_m$ for entities in $\mathcal{U} = \{1, 2, \ldots, n\}$. Thurstone (1927) postulated that the ranking outcome τ_j is determined by n latent variables Z_{ij}'s, for $1 \leq i \leq n$, where Z_{ij} represents ranker j's evaluation score of the ith entity, and $Z_{i_1 j} > Z_{i_2 j}$ if and only if $i_1 \succ i_2$ for ranker j. Define $\boldsymbol{Z}_j = (Z_{1j}, \ldots, Z_{nj})'$ as ranker j's evaluations of all entities, and $\text{rank}(\boldsymbol{Z}_j)$ as the associated full ranking list based on $\boldsymbol{Z}_j$. Similar to Thurstone's assumption, we assume that $\boldsymbol{Z}_j$ follows a multivariate Gaussian distribution with mean $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_n)'$ representing the underlying true score of the ranked entities:

$$\begin{aligned} Z_{ij} &= \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, \sigma^2) \quad (1 \leq i \leq n; 1 \leq j \leq m) \\ \tau_j &= \text{rank}(\boldsymbol{Z}_j), \qquad\qquad\qquad\qquad\qquad\quad (1 \leq j \leq m) \end{aligned} \tag{1.2.1}$$

where ϵ_{ij}'s are jointly independently. Because we only observe the ranking lists τ_j's, multiplying $(\boldsymbol{\mu}, \sigma)$ by a constant or adding a constant to all the μ_i's does not influence the likelihood function. Therefore, to ensure identifiability of the parameters, we fix $\sigma^2 = 1$ and impose the constraint that $\boldsymbol{\mu}$ lies in the space $\boldsymbol{\Theta} = \{\boldsymbol{\mu} \in \mathbb{R}^n : \mathbf{1}'\boldsymbol{\mu} = 0\}$.

Model (1.2.1) implies that the τ_j's are independent and identically distributed (i.i.d.) conditional on $\boldsymbol{\mu}$, so the likelihood function is

$$p(\tau_1, \cdots, \tau_m \mid \boldsymbol{\mu}) = \prod_{j=1}^{m} p(\tau_j \mid \boldsymbol{\mu}) = \prod_{j=1}^{m} \int_{\mathbb{R}^n} p(\tau_j \mid \boldsymbol{Z}_j, \boldsymbol{\mu}) p(\boldsymbol{Z}_j \mid \boldsymbol{\mu}) \mathrm{d}\boldsymbol{Z}_j,$$

where $p(\tau_j \mid \boldsymbol{Z}_j, \boldsymbol{\mu}) = 1_{\{\mathrm{rank}(\boldsymbol{Z}_j) = \tau_j\}}$. Specifically, for any possible full ranking list τ on $\mathcal{U} = \{1, 2, \ldots, n\}$, the probability mass function is

$$P(\tau_j = \tau \mid \boldsymbol{\mu}) = \int_{\mathbb{R}^n} 1\{\mathrm{rank}(\boldsymbol{Z}_j) = \tau\} \cdot (2\pi)^{-n/2} e^{-\frac{1}{2}\|\boldsymbol{Z}_j - \boldsymbol{\mu}\|^2} \mathrm{d}\boldsymbol{Z}_j.$$

Our goal is to generate an aggregated rank based on an estimate of $\boldsymbol{\mu}$ in model (1.2.1). One approach is to use the maximum likelihood estimate (MLE) $\hat{\boldsymbol{\mu}}_m$ defined as

$$\hat{\boldsymbol{\mu}}_m = \arg\max_{\boldsymbol{\mu}} \frac{1}{m} \sum_{j=1}^{m} \log p(\tau_j \mid \boldsymbol{\mu}).$$

We have the following consistency result for $\hat{\boldsymbol{\mu}}_m$ with the proof deferred to the Supplementary Material.

Theorem 1.2.1 *Let true parameter value of model (1.2.1) be $\boldsymbol{\mu}_0 \in \boldsymbol{\Theta}$, and we observe $\tau_1, \ldots, \tau_m$ generated from model (1.2.1). Let $\hat{\boldsymbol{\mu}}_m$ be the MLE of $\boldsymbol{\mu}_0$. Then, for any $\epsilon > 0$ and any compact set $K \subset \boldsymbol{\Theta}$, we have*

$$P\left(\{\|\hat{\boldsymbol{\mu}}_m - \boldsymbol{\mu}_0\|_2 \geq \epsilon\} \cap \{\hat{\boldsymbol{\mu}}_m \in K\}\right) \to 0,$$

as n is fixed and $m \to \infty$.

Alternatively, we can employ a Bayesian procedure, which is more convenient to incorporate prior information, to quantify estimation uncertainties, and to utilize efficient Markov chain Monte Carlo (MCMC) algorithms including data augmentation (Tanner & Wong, 1987) and parameter expansion strategies (Liu & Wu, 1999). With a reasonable prior, the posterior mean of $\boldsymbol{\mu}$ is also a consistent estimator under the same setting as in Theorem 1.2.1. Denote the prior of $\boldsymbol{\mu}$ by $p(\boldsymbol{\mu})$. The posterior distribution of $\boldsymbol{\mu}$ and $(\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m)$ is

$$p(\boldsymbol{\mu}, \boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m \mid \tau_1, \cdots, \tau_m) = p(\boldsymbol{\mu}) \cdot \prod_{j=1}^{m} p(\boldsymbol{Z}_j \mid \boldsymbol{\mu}) \cdot \prod_{j=1}^{m} 1\{\tau_j = \text{rank}(Z_j)\}.$$

We can then generate the aggregated ranking list as

$$\rho = \text{rank}\,(\tilde{\boldsymbol{\mu}}) = \text{rank}\left((\tilde{\mu}_1, \tilde{\mu}_2, \ldots, \tilde{\mu}_n)'\right) \tag{1.2.2}$$

where the $\tilde{\mu}_i$'s are the posterior means of the μ_i's.

Let $\boldsymbol{P}_n = \boldsymbol{I}_n - n^{-1}\mathbf{1}_n\mathbf{1}_n'$ denote the projection matrix that determines a mapping from $\mathbb{R}^n$ to $\boldsymbol{\Theta}$. We choose the prior of $\boldsymbol{\mu}$, which is restricted to the parameter space $\boldsymbol{\Theta}$, to be $\mathcal{N}\left(\mathbf{0}, \sigma_\mu^2 \boldsymbol{P}_n\right)$. The intuition for choosing this prior is that when $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \sigma_\mu^2 \boldsymbol{I}_n)$, we have $\boldsymbol{P}_n\boldsymbol{\mu} \in \boldsymbol{\Theta}$ and $\boldsymbol{P}_n\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \sigma_\mu^2 \boldsymbol{P}_n)$. For computation, it is equivalent to using the prior $\boldsymbol{\mu} \sim \mathcal{N}(\mathbf{0}, \sigma_\mu^2 \boldsymbol{I}_n)$ and considering the posterior mean of $\boldsymbol{P}_n\boldsymbol{\mu} \equiv \boldsymbol{\mu} - \bar{\boldsymbol{\mu}}$, where $\bar{\boldsymbol{\mu}} = n^{-1}\sum_{i=1}^{n} \mu_i \mathbf{1}_n$. In other words,

$$p_{\pi_1}(\boldsymbol{\mu} \mid \tau_1, \cdots, \tau_m) = p_{\pi_2}(\boldsymbol{\mu} - \bar{\boldsymbol{\mu}} \mid \tau_1, \cdots, \tau_m)$$

where $\pi_1 \sim \mathcal{N}\left(\mathbf{0}, \sigma_\mu^2 \boldsymbol{P}_n\right)$ and $\pi_2 \sim \mathcal{N}(\mathbf{0}, \sigma_\mu^2 \boldsymbol{I}_n)$ denote the prior of $\boldsymbol{\mu}$. More generally,

although we restrict $\boldsymbol{\mu}$ to the parameter space $\boldsymbol{\Theta}$, we only need to specify a prior for unconstrained $\boldsymbol{\mu}$ and make inference based on posterior distribution of $\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}$. Therefore, under such Bayesian model setting, it is extremely flexible to extend the model to incorporate covariate information, as illustrated immediately.

1.2.3 RANKING LISTS WITH COVARIATES

As in both examples, each ranked entity is associated with relevant covariates that are available systematic information determining how a ranker ranks it. To incorporate the covariate information into model (1.2.1), we assume that the score of entity i depends linearly on the p-dimensional covariate vector $\boldsymbol{x}_i$, for $i = 1, \ldots, n$. To avoid being too restrictive, we allow the intercept term for each entity to be different. In sum, we have the following over-parameterized model:

$$\begin{aligned} \mu_i &= \alpha_i + \boldsymbol{x}_i'\boldsymbol{\beta}, && (1 \le i \le n) \\ Z_{ij} &= \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim \mathcal{N}(0,1), && (1 \le i \le n; 1 \le j \le m) \\ \tau_j &= \text{rank}(\boldsymbol{Z}_j), && (1 \le j \le m) \end{aligned} \tag{1.2.3}$$

where the ϵ_{ij}'s are mutually independent.

Model (1.2.3) is over-parameterized because $\boldsymbol{\mu}$ is invariant if we add a constant vector $\boldsymbol{c}$ to $\boldsymbol{\beta}$ and change α_i to $\alpha_i - \boldsymbol{x}_i'\boldsymbol{c}$. As a result, the parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_n)$ and $\boldsymbol{\beta}$ are non-identifiable. However, the structure between $\boldsymbol{\mu}$ and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ help us construct some informative priors on $\boldsymbol{\mu}$, incorporating the covariate information. Intuitively, entities with similar $\boldsymbol{x}_i$'s should be close in the underlying μ_i's. Such intuition is conformed by Model (1.2.3) with suitable priors on $(\boldsymbol{\alpha}, \boldsymbol{\beta})$, because similar entities will have higher correlation among their μ_i's *a priori*. Model (1.2.3) can be helpful when the

ranking information is weak and incomplete, and the covariate information is strongly related to the ranking mechanism.

We further illustrate Model (1.2.3) using the quarterback data in Example 1. The unobserved variable Z_{ij} represents ranker j's evaluation for the performance of quarterback i. The expression $\alpha_i + \boldsymbol{x}_i'\boldsymbol{\beta}$ quantifies a hypothetically universal underlying "quality" of the quarterback, and each ranker evaluates it with a personal variation modeled by ϵ_{ij}. The linear term $\boldsymbol{x}_i'\boldsymbol{\beta}$ can explain the part of their performance, but there are many aspects in a football game that cannot be reflected through a linear combination of these summary statistics. The term α_i can capture the remaining "random effect". Without α_i, Model (1.2.3) reduces to a rank regression model in Johnson (2013), which can be too restrictive in some applications.

We set the prior $p(\boldsymbol{\alpha}, \boldsymbol{\beta}) \equiv p(\boldsymbol{\alpha})p(\boldsymbol{\beta})$, where $p(\boldsymbol{\alpha})$ is simply $\mathcal{N}(0, \sigma_\alpha^2 I)$ and $p(\boldsymbol{\beta})$ is $\mathcal{N}(0, \sigma_\beta^2 I)$. The hyper-parameter σ_α and σ_β can reflect prior belief on the relevance of covariates information to ranking mechanism. Intuitively, the stronger the belief on the role of covariates, the smaller the ratio $\sigma_\alpha^2/\sigma_\beta^2$ will be chosen. We address the choice of hyper-parameters $(\sigma_\alpha^2, \sigma_\beta^2)$ in the simulation studies. With this prior, the posterior mean of $\boldsymbol{\mu} - \bar{\boldsymbol{\mu}} = \boldsymbol{\mu} - (n^{-1}\sum_{i=1}^n \mu_i)\mathbf{1}_n$ is our estimates for $\boldsymbol{\mu} \in \boldsymbol{\Theta}$. Below we name this Bayesian approach based on model (1.2.3) as BARC, standing for Bayesian aggregation of rank data with covariates.

1.2.4 Weighted rank aggregation for varying qualities of rankers

In practice, the rankers in consideration may have different quality or reliability. In these cases, a weighted rank aggregation is often more appropriate, where each ranker j has a weight w_j reflecting the quality of its ranking list. However, it is difficult to design a proper weighting scheme in practice, especially when little or no prior knowl-

edge of the rankers is available. To deal with this difficulty, we incorporate weights into variance parameters in our model, and infer them jointly with other parameters. More precisely, we model the ranker's quality by the precision of the noise, i.e, extending model (1.2.3) to the following weighted version:

$$\begin{aligned} \mu_i &= \alpha_i + \boldsymbol{x}_i'\boldsymbol{\beta}, && (1 \le i \le n) \\ Z_{ij} &= \mu_i + \epsilon_{ij}, \quad \epsilon_{ij} \sim N(0, w_j^{-1}), && (1 \le i \le n; 1 \le j \le m) \\ \tau_j &= \text{rank}(\boldsymbol{Z}_j), && (1 \le j \le m) \end{aligned} \tag{1.2.4}$$

where the ϵ_{ij}'s are mutually independent and $w_j > 0$. Note that the variance of ϵ_{ij}, which is the inverse of the ranker's reliability measure w_j, depends only on ranker j's quality, but does not depend on entity i.

The prior for the w_j's can be any distribution bounded away from zero and infinity such as uniform and truncated chi-square distributions. A more restrictive choice is to let the weights take only on a few discrete values. Our numerical study shows that the more restrictive prior specification for the weights can lead to a much less sticky MCMC sampler without compromising much in the precision of aggregated rank as well as the quality evaluation of rankers. Specifically, we restrict w_j to three different levels for reliable, mediocre and low-quality rankers, separately. The corresponding weights for these rankers are 2, 1 and 0.5, respectively, with equal probabilities *a priori*, i.e.,

$$P(w_j = 0.5) = P(w_j = 1) = P(w_j = 2) = \frac{1}{3}, \quad (1 \le j \le m) \tag{1.2.5}$$

where the w_j's are mutually independent. We call this weighted rank aggregation method as BARCW, standing for Bayesian aggregation of rank data with entities' co-

variates and rankers' (unknown) weights.

1.2.5 RANKER CLUSTERING VIA MIXTURE MODEL

Our previous models assume that the underlying score $\boldsymbol{\mu}$ is universal to all rankers, which can sometimes be too restrictive. Böckenholt (1993) and Gormley & Murphy (2006, 2008a,b) suggested that there are often several categories of voters with very different political opinions in an election, and subsequently a mixture model approach should be applied to cluster voters into subgroups. Differing from BARCW, which studies differences in rankers' reliabilities, this mixture model focuses on the heterogeneity in rankers' opinions while assuming that all rankers are equally reliable.

A common issue in mixture models is to determine the number of mixture components. Here we employ the Dirichlet process mixture model, which overcomes this problem by defining mixture distributions with a countably infinite number of components via a Dirichlet process prior (Antoniak, 1974; Ferguson et al., 1983). We first extend Model (1.2.3) so that the underlying score of entities is ranker-specific:

$$\begin{aligned} \boldsymbol{\mu}^{(j)} &= \boldsymbol{\alpha}^{(j)} + \boldsymbol{X}\boldsymbol{\beta}^{(j)}, && (1 \le j \le m) \\ \boldsymbol{Z}_j &= \boldsymbol{\mu}^{(j)} + \varepsilon_j, \quad \varepsilon_j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n), && (1 \le j \le m) \\ \tau_j &= \text{rank}\left(\boldsymbol{Z}_j\right), && (1 \le j \le m) \end{aligned} \tag{1.2.6}$$

where $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ is the covariate matrix for all ranked entities, $\boldsymbol{\mu}^{(j)}$ represents the underlying true score for ranker j, and ε_j's are jointly independent. We then assume that the distribution of $(\boldsymbol{\alpha}^{(j)}, \boldsymbol{\beta}^{(j)})$ follows a Dirichlet process prior, i.e.

$$(\boldsymbol{\alpha}^{(j)}, \boldsymbol{\beta}^{(j)}) \mid G \overset{iid}{\sim} G, \quad G \sim DP(\gamma, G_0), \tag{1.2.7}$$

where G_0 defines a baseline distribution on $\mathbb{R}^n \times \mathbb{R}^p$ for the Dirichlet process prior, satisfying $E(G) = G_0$, and γ is a concentration parameter. For the ease of understanding, we can equivalently view model (1.2.6)-(1.2.7) as the limit of the following finite mixture model with K components when $K \to \infty$:

$$
\begin{aligned}
(\pi_1, \ldots, \pi_K) &\sim \mathrm{Dir}(\gamma/K, \ldots, \gamma/K), \\
q_j \mid \boldsymbol{\pi} &\overset{iid}{\sim} \mathrm{Multinomial}\left(\pi_1, \ldots, \pi_K\right), && (1 \le j \le m) \\
(\boldsymbol{\alpha}^{\langle k \rangle}, \boldsymbol{\beta}^{\langle k \rangle}) &\overset{iid}{\sim} G_0, && (1 \le k \le K) \\
\boldsymbol{\mu}^{\langle k \rangle} &= \boldsymbol{\alpha}^{\langle k \rangle} + \boldsymbol{X}\boldsymbol{\beta}^{\langle k \rangle}, && (1 \le k \le K) \\
\boldsymbol{Z}_j &= \boldsymbol{\mu}^{\langle q_j \rangle} + \boldsymbol{\varepsilon}_j, \quad \boldsymbol{\varepsilon}_j \sim \mathcal{N}(\mathbf{0}, \boldsymbol{I}_n), && (1 \le j \le m) \\
\tau_j &= \mathrm{rank}\left(\boldsymbol{Z}_j\right), && (1 \le j \le m)
\end{aligned}
\tag{1.2.8}
$$

where the latent variable $q_j \in \{1, 2, \ldots, K\}$ indicates the cluster allocation of ranker j, and $\boldsymbol{\mu}^{\langle k \rangle}$ corresponds to the common underlying score vector for rankers in cluster k.

We choose the baseline distribution G_0 on $\mathbb{R}^n \times \mathbb{R}^p$ using two independent zero-mean Gaussian distributions with covariances $\sigma_\alpha^2 \boldsymbol{I}_n$ and $\sigma_\beta^2 \boldsymbol{I}_p$, i.e., $G_0 \sim \mathcal{N}(\mathbf{0}, \mathrm{diag}(\sigma_\alpha^2 \boldsymbol{I}_n, \sigma_\beta^2 \boldsymbol{I}_p))$. Clearly, G_0 is the same as the prior distribution of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ we use in the previous models, and the conjugacy between G_0 and the distribution of $\boldsymbol{Z}_j$'s leads to a straightforward Gibbs sampler as described in Neal (1992) and MacEachern (1994). Parameter γ represents the degree of concentration of G around G_0 and, thus, is related to the number of distinct clusters. According to the Pólya urn scheme representation of the Dirichlet process in Blackwell & MacQueen (1973), the prior probability that a new ranker belongs to a different cluster with all m existing rankers is $\gamma/(m + \gamma - 1)$. In addition, the expected number of clusters with in total m rankers is $\sum_{j=1}^{m} \gamma/(j + \gamma - 1)$ *a priori*.

We discuss the sensitivity of this hyper-parameter in the simulation studies.

Under this Dirichlet process mixture model, we are interested in rank aggregation within each cluster as well as rank aggregation across all clusters. The aggregated ranking in each cluster k is determined by the order of $\boldsymbol{\mu}^{\langle k \rangle}$, or equivalently $\boldsymbol{\mu}^{(j)}$'s with cluster allocation $q_j = k$. The aggregated ranking list across all clusters depends on the underlying score of all rankers:

$$\rho = \text{rank}\left(m^{-1} \sum_{j=1}^{m} \tilde{\boldsymbol{\mu}}^{(j)} \right), \tag{1.2.9}$$

where $\tilde{\boldsymbol{\mu}}^{(j)}$ is the posterior mean of the $\boldsymbol{\mu}^{(j)}$ for each ranker j. We regard this rank aggregation method as BARCM, standing for Bayesian Aggregation of Rank data with Covariates of entities and Mixture of rankers with different ranking opinions.

1.2.6 Extension to partial ranking lists

Model (1.2.1),(1.2.3),(1.2.4) and (1.2.6) can all be applied when the observations are partial ranking lists. Because we define ranking list as a set of non-contradictory pairwise relations among ranked entities, partial ranking lists appear when any of the pairwise relations is missing. Thus, besides the partial ranking list τ_j $(1 \leq j \leq m)$, we also observe the δ_j's, which indicate which pairwise relationship is missing. Under latent variable models, we denote $\tau_j \simeq \text{rank}(\boldsymbol{Z}_j)$ if the partial ranking list τ_j is consistent with the full ranking list $\text{rank}(\boldsymbol{Z}_j)$. Our models, BARC, BARCW and BARCM, for the observed individual partial ranking lists are the same as in (1.2.3), (1.2.4) and (1.2.6)-(1.2.7), except that $\tau_j = \text{rank}(\boldsymbol{Z}_j)$ is replaced by $\tau_j \simeq \text{rank}(\boldsymbol{Z}_j)$. Let $\boldsymbol{\theta}_\delta$ and $\boldsymbol{\theta}_\tau$ denote the parameters for missing indicators δ_j's and ranking lists τ_j's, respectively. We can

then write the likelihood of (δ_j, τ_j) as

$$p(\delta_j, \tau_j \mid \boldsymbol{\theta}_\delta, \boldsymbol{\theta}_\tau, \boldsymbol{X}) = \sum_{r: r \simeq \tau_j} \int_{\mathbb{R}^n} p(\delta_j \mid r, \boldsymbol{Z}_j, \boldsymbol{\theta}_\delta, \boldsymbol{X}) 1\{r = \text{rank}(\boldsymbol{Z}_j)\} p(\boldsymbol{Z}_j \mid \boldsymbol{\theta}_\tau, \boldsymbol{X}) \mathrm{d}\boldsymbol{Z}_j.$$

If the pairwise relations are missing at random, in the sense that $p(\delta_j \mid r, \boldsymbol{Z}_j, \boldsymbol{\theta}_\delta, \boldsymbol{X}) = p(\delta_j \mid \tilde{r}, \tilde{\boldsymbol{Z}}_j, \boldsymbol{\theta}_\delta, \boldsymbol{X})$ for all possible $(r, \boldsymbol{Z}_j, \tilde{r}, \tilde{\boldsymbol{Z}}_j)$ such that $r = \text{rank}(\boldsymbol{Z}_j) \simeq \tau_j$ and $\tilde{r} = \text{rank}(\tilde{\boldsymbol{Z}}_j) \simeq \tau_j$, then the likelihood of (δ_j, τ_j) can be simplified as

$$p(\delta_j, \tau_j \mid \boldsymbol{\theta}_\delta, \boldsymbol{\theta}_\tau, \boldsymbol{X}) = p(\delta_j \mid \tau_j, \boldsymbol{\theta}_\delta, \boldsymbol{X}) \int_{\mathbb{R}^n} 1\{\tau_j \simeq \text{rank}(\boldsymbol{Z}_j)\} p(\boldsymbol{Z}_j \mid \boldsymbol{\theta}_\tau, \boldsymbol{X}) \mathrm{d}\boldsymbol{Z}_j$$

If the priors for the parameters $\boldsymbol{\theta}_\delta$ and $\boldsymbol{\theta}_\tau$ are mutually independent, we can further ignore the δ_j's when conducting the Bayesian inference for the parameter $\boldsymbol{\theta}_\tau$ of ranking mechanisms.

1.3 MCMC COMPUTATION WITH PARAMETER EXPANSION

We use Gibbs sampling with parameter expansion (Liu & Wu, 1999) in our Bayesian computation for the latent variable models with covariates. We start with model (1.2.3) and then generalize this MCMC strategy to two extended models, (1.2.4) and (1.2.6)-(1.2.7). To simplify the notation, we define $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m) \in \mathbb{R}^{n \times m}$, $\mathcal{T} = \{\tau_j\}_{j=1}^m$, $\boldsymbol{V} = (\boldsymbol{I}_n, \boldsymbol{X}) \in \mathbb{R}^{n \times (n+p)}$, and $\boldsymbol{\Lambda} = \text{diag}(\sigma_\alpha^2 \boldsymbol{I}_n, \sigma_\beta^2 \boldsymbol{I}_p) \in \mathbb{R}^{(n+p) \times (n+p)}$.

1.3.1 PARAMETER-EXPANDED GIBBS SAMPLER

The most computationally expensive part in our model is to sample all the Z_{ij}'s from the truncated Gaussian distributions. Furthermore, because $\boldsymbol{Z}$ and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ are intertwined together due to the posited regression model, they tend to correlate highly,

similar to the difficulty of the data augmentation method introduced by Albert & Chib (1993) for probit regression models.

To speed up the algorithm, we follow Scheme 2 in Liu & Wu (1999) and exploit a parameter-expanded data augmentation (PX-DA) algorithm. In particular, we introduce a group scale transformation of the "missing data" matrix $\boldsymbol{Z}$, the evaluation scores of all rankers for all ranked identities, indexed by a non-negative parameter θ, i.e., $t_\theta(\boldsymbol{Z}) \equiv \boldsymbol{Z}/\theta$. The PX-DA algorithm updates the missing data $\boldsymbol{Z}$ and the expanded parameters $(\theta, \boldsymbol{\alpha}, \boldsymbol{\beta})$ iteratively as follows:

1. For $i = 1, \ldots, n$ and $j = 1, \ldots, m$, draw $[Z_{ij} \mid Z_{[-i],j}, \boldsymbol{Z}_{[-j]}, \boldsymbol{\alpha}, \boldsymbol{\beta}]$ from $\mathcal{N}(\alpha_i + \boldsymbol{x}_i'\boldsymbol{\beta}, 1)$ with truncation points determined by $Z_{[-i],j}$, such that Z_{ij} falls in the correct position according to τ_j, i.e., $\text{rank}(\boldsymbol{Z}_j) \simeq \tau_j$.

2. Draw $\theta \sim p(\theta \mid \boldsymbol{Z}, \mathcal{T}) \propto p(t_\theta(\boldsymbol{Z}))|J_\theta(\boldsymbol{Z})|H(d\theta)$. Here, $|J_\theta(\boldsymbol{Z})| = \theta^{-nm}$ is the Jacobian of scale transformation, $H(d\theta) = \theta^{-1}d\theta$ is the Haar measure on a scale group up to a constant, and

$$p(t_\theta(\boldsymbol{Z})) \propto \int p(t_\theta(\boldsymbol{Z}) \mid \boldsymbol{\alpha}, \boldsymbol{\beta}) p(\boldsymbol{\alpha}) p(\boldsymbol{\beta}) \mathrm{d}\boldsymbol{\alpha} \mathrm{d}\boldsymbol{\beta} \propto \exp\left\{-\frac{S}{2\theta^2}\right\},$$

is the marginal density of latent variables evaluated at $t_\theta(\boldsymbol{Z})$, where

$$S = \sum_{j=1}^{m} \boldsymbol{Z}_j'\boldsymbol{Z}_j - \sum_{j=1}^{m}\sum_{k=1}^{m} \boldsymbol{Z}_j'\boldsymbol{V}(\boldsymbol{\Lambda}^{-1} + m\boldsymbol{V}'\boldsymbol{V})^{-1}\boldsymbol{V}'\boldsymbol{Z}_k.$$

We can derive that $\theta^2 \sim S/\chi^2_{nm}$.

3. Draw $(\boldsymbol{\alpha}, \boldsymbol{\beta}) \sim p(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid t_\theta(\boldsymbol{Z})) \equiv \mathcal{N}\left(\hat{\boldsymbol{\eta}}/\theta, \hat{\boldsymbol{\Sigma}}\right)$, where

$$\hat{\boldsymbol{\eta}} = (\boldsymbol{\Lambda}^{-1} + m\boldsymbol{V}'\boldsymbol{V})^{-1}\boldsymbol{V}' \sum_{j=1}^{m} \boldsymbol{Z}_j \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = (\boldsymbol{\Lambda}^{-1} + m\boldsymbol{V}'\boldsymbol{V})^{-1}.$$

Below we give some intuition on why the PX-DA algorithm improves efficiency. Without Step 2 and with $t_\theta(\boldsymbol{Z})$ in Step 3 replaced by $\boldsymbol{Z}$, the algorithm reduces to the standard Gibbs sampler, which updates the missing data and parameters iteratively.The scale group move of $\boldsymbol{Z}$ under the usual Gibbs sampler is slow due to both the Gibbs update for $\boldsymbol{Z}$ in Step 1 and the high correlation between $\boldsymbol{Z}$ and $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. To overcome such difficulty, the PX-DA algorithm introduces a scale transformation of $\boldsymbol{Z}$ to facilitate its group move and mitigate its correlation with $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. To ensure the validity of the MCMC algorithm, the scale transformation parameter θ has to be drawn from a carefully specified distribution, such that the move is invariant under the target posterior distribution, i.e., $t_\theta(\boldsymbol{Z})$ follows the same distribution as the original $\boldsymbol{Z}$ under stationarity. To aid in understanding, we provide a proof in the Supplementary Material that the specified distribution of θ in Step 2 satisfies this property.

1.3.2 GIBBS SAMPLER FOR BARCW

Under Model (1.2.4) for BARCW, the Gibbs steps for $[\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\alpha} \mid \mathcal{T}, \boldsymbol{W}]$ is very similar to that for $[\boldsymbol{Z}, \boldsymbol{\beta}, \boldsymbol{\alpha} \mid \mathcal{T}]$ in the previous model for BARC, with details relegated to the Supplementary Material. The additional step is to draw w_j given all other variables. For $j = 1, \ldots, m$, we draw discrete random variable w_j from the following conditional

posterior probability mass function:

$$p(w_j \mid \boldsymbol{Z}, \boldsymbol{w}_{[-j]}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathcal{T}) \propto p(w_j) p(\boldsymbol{Z} \mid \boldsymbol{\alpha}, \boldsymbol{\beta}, \boldsymbol{w}) \\ \propto w_j^{\frac{n}{2}} \exp\left(-\frac{w_j}{2} \sum_{i=1}^{n} \left(Z_{ij} - \boldsymbol{x}_i' \boldsymbol{\beta} - \alpha_i \right)^2 \right).$$

1.3.3 Gibbs sampler for BARCM

Under model (1.2.6)-(1.2.7) for BARCM, we first represent the parameters $\{\boldsymbol{\alpha}^{(j)}, \boldsymbol{\beta}^{(j)}\}_{j=1}^{m}$ by cluster allocation vector $\boldsymbol{q} = (q_1, \ldots, q_m)$ and cluster-wise parameters $\{\boldsymbol{\alpha}^{\langle k \rangle}, \boldsymbol{\beta}^{\langle k \rangle} : k \in \{q_1, \ldots, q_m\}\}$, and then use the MCMC algorithm to sample $\boldsymbol{q}$, $(\boldsymbol{\alpha}^{\langle k \rangle}, \boldsymbol{\beta}^{\langle k \rangle})$'s and $\boldsymbol{Z} = (\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m)$.

Let $\mathcal{A}_k(\boldsymbol{q}) = \{j \mid 1 \leq j \leq m, q_j = k\}$ denote the set of rankers that belong to cluster k given cluster allocation $\boldsymbol{q}$. Due to the conjugacy between G_0 and the distribution of $\boldsymbol{Z}_j$'s, we can integrate out $(\boldsymbol{\alpha}^{\langle k \rangle}, \boldsymbol{\beta}^{\langle k \rangle})$'s when sampling $\boldsymbol{q}$, and Gibbs sampling of $\boldsymbol{q}$ given $\boldsymbol{Z}$ follows from Algorithm 3 in Neal (2000). Specifically, the Gibbs steps are as follows:

1. For $j = 1, \ldots, m$, draw q_j from

$$\begin{aligned} & P\left(q_j = k \mid \boldsymbol{Z}, \boldsymbol{q}_{[-j]}, \mathcal{T}\right) \\ \propto \; & P\left(q_j = k \mid \boldsymbol{q}_{[-j]}\right) \int p\left(\boldsymbol{Z}_j \mid \boldsymbol{\alpha}^{\langle k \rangle}, \boldsymbol{\beta}^{\langle k \rangle}\right) p\left(\boldsymbol{\alpha}^{\langle k \rangle}, \boldsymbol{\beta}^{\langle k \rangle} \mid \boldsymbol{Z}_{[-j]}\right) \mathrm{d}\boldsymbol{\alpha}^{\langle k \rangle} \mathrm{d}\boldsymbol{\beta}^{\langle k \rangle} \\ \propto \; & P\left(q_j = k \mid \boldsymbol{q}_{[-j]}\right) \cdot \exp\left(-\frac{1}{2} S_k(\boldsymbol{q}) + \frac{1}{2} S_k(\boldsymbol{q}_{[-j]}) \right), \end{aligned}$$

   where

$$S_k(\boldsymbol{q}) = \sum_{j \in \mathcal{A}_k(\boldsymbol{q})} \boldsymbol{Z}_j' \boldsymbol{Z}_j - \sum_{j \in \mathcal{A}_k(\boldsymbol{q})} \sum_{l \in \mathcal{A}_k(\boldsymbol{q})} \boldsymbol{Z}_j' \boldsymbol{V} \left(\boldsymbol{\Lambda}^{-1} + |\mathcal{A}_k(\boldsymbol{q})| \boldsymbol{V}' \boldsymbol{V} \right)^{-1} \boldsymbol{V}' \boldsymbol{Z}_k,$$

$|\mathcal{A}_k(\boldsymbol{q}_{[-j]})|$ denotes the number of units except j that are in cluster k, and $P\left(q_j \mid \boldsymbol{q}_{[-j]}\right)$ is determined as follows:

$$\begin{aligned} \text{If } k = q_i \text{ for some } i \neq j: \; P\left(q_j = k \mid \boldsymbol{q}_{[-j]}\right) &= \frac{|\mathcal{A}_k(\boldsymbol{q}_{[-j]})|}{(m-1+\gamma)} \\ P\left(q_j \neq q_i \text{ for all } i \neq j \mid \boldsymbol{q}_{[-j]}\right) &= \frac{\gamma}{(m-1+\gamma)}, \end{aligned}$$

2. For each $k \in \{q_1, \ldots, q_m\}$, we sample $[\boldsymbol{Z}_{\mathcal{A}_k(\boldsymbol{q})}, \boldsymbol{\alpha}^{\langle k \rangle}, \boldsymbol{\beta}^{\langle k \rangle} \mid \mathcal{T}, \boldsymbol{q}]$ using very similar Gibbs sampling steps as we sample $[\boldsymbol{Z}, \boldsymbol{\alpha}, \boldsymbol{\beta} \mid \mathcal{T}]$ in the BARC model, with details relegated to the Supplementary Material.

1.4 Rank aggregation via MCMC samples

Following the Bayesian computation in the previous section, we can obtain MCMC samples from the posterior distribution of $(\boldsymbol{\alpha}, \boldsymbol{\beta})$ under BARC or BARCW, and from the posterior distribution of $(\boldsymbol{\alpha}^{(j)}, \boldsymbol{\beta}^{(j)})$'s under BARCM. As described in (1.2.2) and (1.2.9), we use the posterior mean of $\mu_i \equiv \alpha_i + \boldsymbol{x}_i'\boldsymbol{\beta}$'s to generate the aggregated ranking list in BARC and BARCW, and use the posterior mean of $m^{-1}\sum_{j=1}^{m} \mu_i^{(j)} = m^{-1}\sum_k |\mathcal{A}_k(\boldsymbol{q})|(\alpha_i^{\langle k \rangle} + \boldsymbol{x}_i'\boldsymbol{\beta}^{\langle k \rangle})$'s in BARCM. Moreover, we have some byproducts from the Bayesian inference besides the aggregated ranking lists, as illustrated below.

1.4.1 Probability interval for the aggregated ranking list

In existing rank aggregation methods, people usually seek only one aggregated rank, but ignore the uncertainty of the aggregation result. When we observe $i \succ j$ in a single ranking list ρ, we cannot tell whether i is much better than j or they are close. The Bayesian inference provides us a natural uncertainty measure for the ranking result.

Under BARC or BARCW, suppose we have MCMC samples $\{\boldsymbol{\mu}^{[l]}\}_{l=1}^{L}$, from the posterior distribution $p(\boldsymbol{\mu} \mid \tau_1, \cdots, \tau_m)$. For each sample $\boldsymbol{\mu}^{[l]}$, we calculate a ranking list $\rho^{[l]} = \text{rank}(\boldsymbol{\mu}^{[l]})$. We denote $\tau^{[l]}(i)$ as the position of entity i in ranking list $\rho^{[l]}$, and define the $(1-\alpha)$ probability interval of entity i's rank as

$$\left(\tau^{LB}(i), \tau^{UB}(i)\right) = \left(\tau_{(\frac{\alpha}{2})}(i), \tau_{(1-\frac{\alpha}{2})}(i)\right),$$

where $\tau_{(\frac{\alpha}{2})}(i)$ and $\tau_{(1-\frac{\alpha}{2})}(i)$ are the $\frac{\alpha}{2}$th and $(1-\frac{\alpha}{2})$th sample quantiles of $\{\tau^{[l]}(i)\}_{l=1}^{L}$. The construction of credible intervals for entities' ranks under BARCM is very similar, and thus is omitted here.

1.4.2 Measurements of heterogeneous rankers

In BARCW and BARCM, we aim to learn the heterogeneity in rankers and subsequently improve as well as better interpret the rank aggregation results. Both methods deliver meaningful measures to detect heterogeneous rankers.

In BARCW, we assume that all rankers share the same opinion and the samples from $p(\boldsymbol{w} \mid \mathcal{T})$ measure the reliability of the input rankers. In BARCM, we assume that there exist a few groups of rankers with different opinions, despite all being reliable rankers. The MCMC samples from $p(\boldsymbol{q} \mid \mathcal{T})$ estimate ranker clusters with different opinions. The number of clusters is determined by the number of distinct values in cluster allocation $\boldsymbol{q}$. The opinion of rankers in cluster k can be aggregated by the posterior means of $\alpha_i^{\langle k \rangle} + \boldsymbol{x}_i' \boldsymbol{\beta}^{\langle k \rangle}$'s. We compare both methods later in simulation studies and real applications.

1.4.3 Role of covariates in the ranking mechanism

As discussed in Section 1.2.3, the interpretation of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ is difficult due to over-parameterization. However, noting that the α_i's are modeled as i.i.d Gaussian random variables with mean zero *a priori*, the posterior distribution of $\boldsymbol{\beta}$ still provides some meaningful information about the role of covariates in the ranking mechanism. Intuitively, for each ranked entity i, the projection $\boldsymbol{x}_i'\boldsymbol{\beta}$ can be seen as the part of the evaluation score $\boldsymbol{\mu}_i$ linearly explained by the covariates, and α_i as the corresponding residual. The sign and magnitude of the coefficient β_k for the kth covariate indicate the positive or negative role of covariates and its strength in determining the ranking list. In practice, we can incorporate nonlinear transformations of original covariates to allow for more flexible role of covariates in explaining the ranking mechanism.

1.5 Simulation studies

We adopt the normalized *Kendall tau distance* (Kendall, 1938) between ranking lists in evaluation to compare our methods with other rank aggregation methods. Another popular distance measure *Spearman's footrule distance* (Diaconis & Graham, 1977) gives very similar results and is thus omitted here.

1.5.1 Comparison between BARC and other rank aggregation methods

Recall that $\mathcal{U}$ is the set $\{1, \ldots, n\}$ of entities, and entity i has true value μ_i. We generate m full ranking lists $\{\tau_j\}_{j=1}^m$ via the following model:

$$\tau_j = \text{rank}(\boldsymbol{Z}_j), \quad \text{where } \boldsymbol{Z}_j \overset{iid}{\sim} \mathcal{N}(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I}_n).$$

We generate i.i.d. vectors $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})' \in \mathbb{R}^p$ from the multivariate normal distribution with mean 0 and covariance $\mathrm{Cov}(x_{is}, x_{it}) = \rho^{|s-t|}$ for $1 \leq s, t \leq p$, and examine three scenarios. In Scenario 1, the true difference between entities can be linearly explained by covariates. In Scenario 2, a linear combination of covariates can partially explain the ranking. In Scenario 3, the ranking mechanism is barely correlated with the covariates.

1. $\mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta}$, where $\boldsymbol{\beta} = (3, 2, 1, 0.5)'$, $p = 4$, and $\rho = 0.2$.

2. $\mu_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \|\boldsymbol{x}_i\|^2$, where $\boldsymbol{\beta} = (3, 2, 1)'$, $p = 3$, and $\rho = 0.5$.

3. $\mu_i = \|\boldsymbol{x}_i\|^2$, where $p = 4$, and $\rho = 0.5$.

We first examine the impact of the noise level σ on the performance of BARC and other rank aggregation methods. Fixing $n = 50$ and $m = 10$, we tried four different values of σ $(= 5, 10, 20, 40)$. For each configuration, we generated 500 simulated datasets. We applied Borda Count (BC), Markov-Chain based methods (MC1, MC2, MC3), Plackett–Luce based method (PL) and our BARC method. A brief review of the aforementioned methods can be found in the Supplementary Material. When utilizing BARC and its extensions, we input standardized covariates and set hyper-parameters $\sigma_\alpha = 1$ and $\sigma_\beta = 100$ unless otherwise stated. Intuitively, with a small σ_α and a large σ_β, BARC would exploit the role of covariates in rank aggregation.

The Kendall's tau distances between the true rank and the aggregated ranks produced by the six methods, averaged over the 500 simulated datasets, are plotted against the noise level in Figure 1.1. We can observe that BARC uniformly outperformed the competing methods in Scenarios 1 and 2 when the linear combination of covariates is useful, and was competitive in scenario 3. The PL method underperformed all other methods but MC1 due to its misspecified distributional assumption.

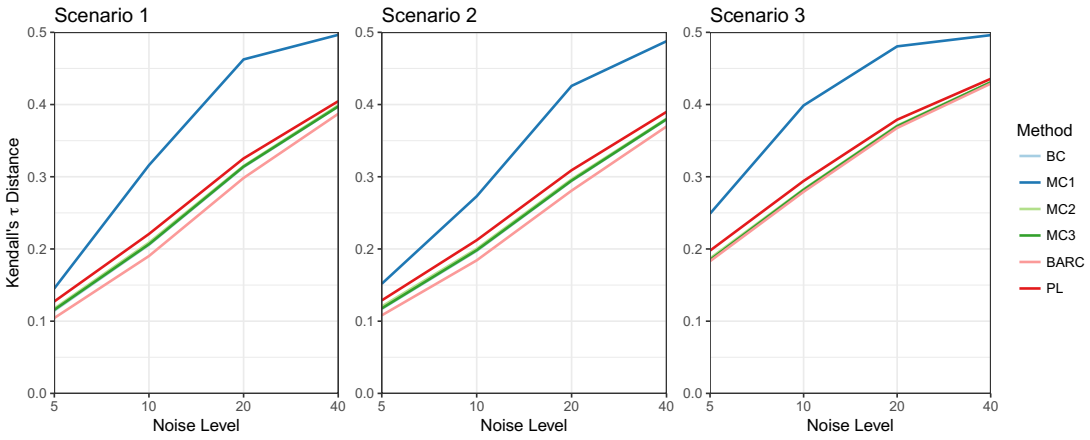


Figure 1.1: Average distance between true rank and aggregated ranks of different methods. As the covariates become increasingly dis-associated with the ranking mechanism from Scenarios 1 to 3, the advantage of BARC over existing methods shrank. Under these scenarios, the lines of MC2, MC3 and Borda Count overlap. In Scenario 3, the results of MC2, MC3, Borda Count and BARC are extremely close as the ranking does not associate with covariates linearly.

1.5.2 Computational advantage of parameter expansion

Before we move to more complicated settings, we would like to use the above simulation to demonstrate the effectiveness of parameter expansion in dealing with rank data. We use Scenario 2 with noise level $\sigma = 5$ as an illustration. Figure 1.2 shows that Gibbs sampler with parameter expansion reduces the autocorrelation in MCMC samples compared to regular Gibbs sampler.

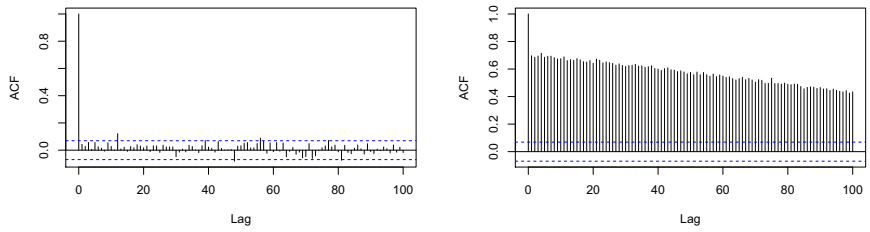


Figure 1.2: The left panel is autocorrelation plot of β_1 in parameter expanded Gibbs sampler; The right panel is the autocorrelation plot of β_1 in regular Gibbs sampler.

1.5.3 BARC WITH PARTIAL RANKING LISTS

We further explore how BARC performs for aggregating partial ranking lists, where subgroups have no overlap with each other. This is a similar situation as we observe in Example 2. We simulated data from Scenario 2 with $n = 80$, $m = 10$ and $p = 3$. We randomly divide these 80 entities into $k\,(= 1, 2, 4, 8, 10, 16)$ subgroups, each with size n/k. As k increases, the pairwise comparison information decreases. For example, when $k = 16$, we have only 5.06% of all the pairwise comparisons in a partial ranking list. Figure 1.3 displays the Kendall's tau distances between the true rank and the aggregated ranking lists inferred by BARC in different cases. BARC is quite robust with respect to partial ranking lists when unobserved pairwise comparisons are missing completely at random and the input ranking lists have moderate dependence on the available covariates. In contrast, denoted by BAR in Figure 1.3, the BARC method without using covariates is susceptible to partial lists.

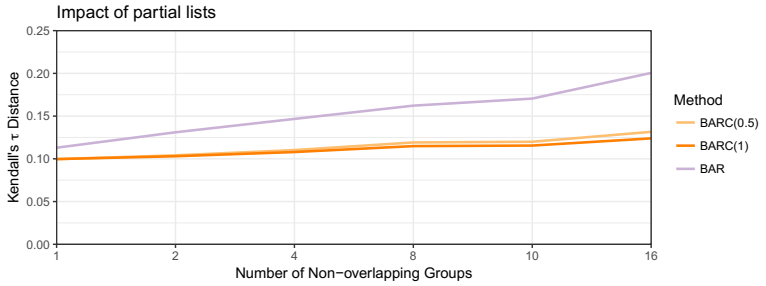


Figure 1.3: We applied BARC with two different settings for hyper-parameter σ_α to aggregate partial ranking lists. For comparison, we also applied our method without using covariates, denoted as BAR. With the help of covariates, BARC's performance were relatively unaffected by the increase of the incompleteness of the ranking lists. BARC is also robust with hyper-parameter choices—the BARC lines with different values of σ_α (i.e., 0.5 and 1) were very close to each other.

1.5.4 BARCM FOR HETEROGENEOUS OPINIONS IN RANKING LISTS

In real world, there can be a few groups of rankers with different opinions, despite all being reliable rankers. Dirichlet process mixture model (1.2.6)-(1.2.7) clusters the rankers and can automatically determine the total number of clusters. Here, we use simulation to study the sensitivity of BARCM to hyper-parameter γ in the Dirichlet process prior. In addition, we explore the performance of BARCW under this misspecified model setting.

We simulated under the BARC model with three mixture components. Mimicking the dataset in Example 2, we have $m = 69$ rankers, $p = 11$ covariates each entity, and $n = 108$ entities divided into 9 non-overlapping groups of equal size. The categories of rankers are generated with probability $\boldsymbol{\pi} = (0.5, 0.3, 0.2)$. The covariates $\boldsymbol{x}_i$'s are generated from multivariate normal distribution with mean 0 and covariance $Cov(x_{is}, x_{it}) = (0.2)^{|s-t|}$, and the coefficients are generated from $\boldsymbol{\beta}^{\langle k \rangle} \overset{iid}{\sim} \mathcal{N}(0, \boldsymbol{I}_p)$ and $\alpha_i^{\langle k \rangle} \overset{iid}{\sim} \mathcal{N}(0, 2^2)$ for $k = 1, 2, 3$ and $i = 1, \ldots, n$. The noise level is fixed at $\sigma = 1$. Table 1.5 shows the average clustering accuracy under different hyper-parameters. The clustering accuracy here is measured by Rand Index, which is the percentage of pairwise clustering decisions that are correct (Rand, 1971). The hyper-parameters clearly impacts the number of clusters in the mixture model, but the results are quite robust in terms of the clustering error.

Table 1.5: Clustering analysis under heterogeneous setting: average clustering accuracy and number of clusters given by BARCM over 100 simulations under each γ value.

	$\gamma = m^{-1}$	$\gamma = m^{-1/2}$	$\gamma = 1$	$\gamma = m^{1/2}$
Clustering accuracy	0.994	0.990	0.987	0.979
Expected # of clusters *a posteriori*	3.629	4.162	4.671	5.746
Expected # of clusters *a priori*	1.069	1.557	4.819	18.986

We also applied BARCW to this simulated data set. Figure 1.4 shows that the minority opinions are down-weighted by BARCW, which assumes that all rankers share the same opinion. As a result, BARCW reinforces the majority's opinion in rank aggregation. In practice, we recommend to apply BARCM to check if there are several sizable ranker subgroups. By studying rankers' heterogeneity, we can better understand our ranking data even if we seek only one aggregated ranking list.

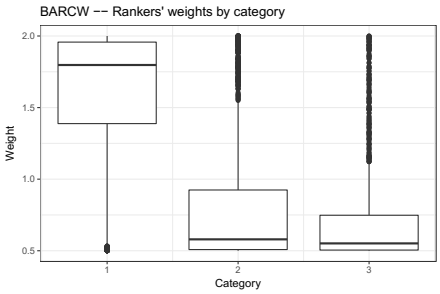


Figure 1.4: Box plot of weights by ranker categories given by BARCW in 100 simulations. Ranker category with the largest proportion is weighted higher than the other two.

1.5.5 Robustness of BARCM and BARCW under homogeneous setting

In contrast to the simulation with heterogeneous ranker qualities or opinions, we also simulated the BARC model under the homogeneous setting to verify the robustness of BARCM and BARCW. The simulation is the same as 1.5.4 except that all rankers are from one component with equal qualities. Table 1.6 shows the average clustering accuracy under different hyper-parameters. Figure 1.5 shows the histogram of the rankers' weights given by BARCW. Under this homogeneous setting, BARCM clustered the rankers into one group, and BARCW assigned the rankers' weights mostly near the maximum.

Table 1.6: Clustering analysis under homogeneous setting: average clustering accuracy and number of clusters given by BARCM over 100 simulations under each γ value.

	$\gamma = m^{-1}$	$\gamma = m^{-1/2}$	$\gamma = 1$	$\gamma = m^{1/2}$
Clustering accuracy	1.000	0.999	0.998	0.993
Expected # of clusters *a posteriori*	1.007	1.033	1.057	1.215

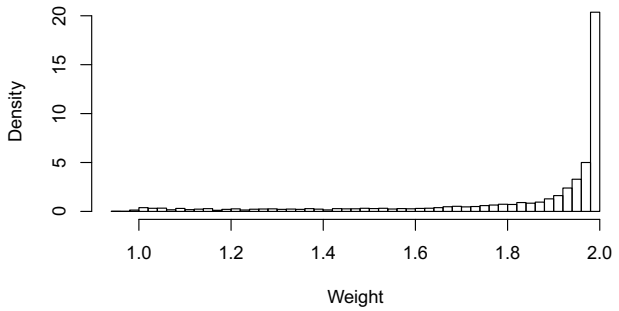


Figure 1.5: Histogram of rankers' weights given by BARCW in 100 simulations under homogeneous setting. Almost all the rankers are learned to be reliable.

1.6 Analyses of the two real data sets

1.6.1 Aggregating NFL Quarterback rankings

Ranking NFL quarterbacks is a classic case where experts' ranking schemes are clearly related to some performance statistics of the players in their games. Information in Tables 1.1 and 1.2 enables us to generate aggregated lists using both rank data and the covariates information, as shown in Table 1.7. For quarterbacks at the top and bottom of the list, these methods mostly agree with each other. Among all compared rank aggregation results, the PL method has the largest discrepancy with other methods, especially in the bottom half where the ranking uncertainty is large. Some diagnostics plots for MCMC convergence are provided in the Supplementary Material.

Table 1.7: Rank aggregation results of NFL quarterbacks listed in the order of BARCW posterior means. The rankings are listed to the right of the underlying values given by rank aggregation methods.

Player	BARCW μ	Rank	BARC μ	Rank	BC	Rank	PL.γ	Rank	MC3.π	Rank
Andrew Luck	6.518	1	6.069	1	1.286	1	0.361	1	0.207	1
Aaron Rodgers	4.76	2	4.635	2	2.571	2	0.195	2	0.137	2
Peyton Manning	4.466	3	4.39	3	3	3	0.171	3	0.12	3
Tom Brady	2.937	4	2.942	4	5.071	4	0.072	4	0.071	4
Tony Romo	2.805	5	2.744	5	5.214	5	0.062	5	0.07	5
Drew Brees	2.469	6	2.448	6	5.857	6	0.043	6	0.063	6
Ben Roethlisberger	2.149	7	2.094	7	6.571	7	0.035	7	0.052	7
Ryan Tannehill	1.435	8	1.342	8	8	8	0.020	8	0.04	8
Matthew Stafford	0.965	9	0.72	9	8.857	9	0.015	9	0.034	9
Mark Sanchez	-0.005	10	-0.098	10	11.5	10	0.005	11	0.023	10
Russell Wilson	-0.716	11	-0.496	11	12.214	11	0.005	10	0.021	11
Philip Rivers	-0.93	12	-0.602	12	13.214	12	0.003	12	0.02	12
Cam Newton	-1.197	13	-1.136	13	14.5	13	0.002	13	0.017	13
Matt Ryan	-1.413	14	-1.474	14	16.357	15	0.001	15	0.014	15
Eli Manning	-1.474	15	-1.497	15	16.071	14	0.002	14	0.014	14
Alex Smith	-1.793	16	-1.717	17	16.714	16	0.001	17	0.013	16
Colin Kaepernick	-1.813	17	-1.601	16	16.786	17	0.001	18	0.013	17
Joe Flacco	-1.815	18	-1.778	19	16.929	18	0.001	20	0.013	18
Jay Culter	-1.884	19	-1.726	18	17.143	19	0.001	19	0.013	19
Andy Dalton	-1.987	20	-2.052	20	17.357	20	0.001	16	0.012	20
Josh McCown	-2.733	21	-2.645	21	19.5	21	0.001	21	0.01	21
Drew Stanton	-2.812	22	-2.823	22	20.286	22	0.001	22	0.009	22
Teddy Bridgewater	-3.476	23	-3.378	23	21.714	23	0.000	23	0.008	23
Brian Hoyer	-4.462	24	-4.361	24	23.286	24	0.000	24	0.007	24

Figure 1.6 shows the 95% probability interval for each quarterback's rank under both BARC and BARCW. We can see that the interval width is large for mediocre quarterbacks, and that is exactly where a majority of discrepancies occurred among different rankers and different rank aggregation methods. The interval estimates of aggregated ranks can separate several elite quarterbacks from the others.

All methods except BARCW assume equal reliability for all input lists. Figure 1.7 shows the posterior boxplots and posterior means of the weights. Out of the 13 rankers, seven are inferred to have significantly higher quality than the other six rankers. We further validated our weights estimation using the prediction accuracy of each expert at the end of the season. Table 1.8 shows the means and standard deviations of two well separated groups of rankers.

Figure 1.8 gives us intuition about the role of covariates in our rank aggregation. TD and Int, which stand for percentage of touchdowns and interceptions thrown when

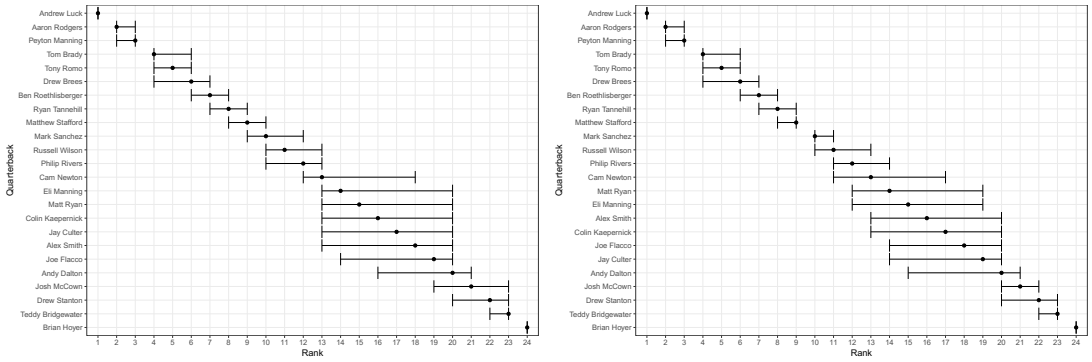


Figure 1.6: Interval estimates of aggregated ranks of NFL quarterbacks, as of week 12 in the 2014 NFL season. The plot on the left is given by BARC, and the right one is from BARCW. The differences between these two results are very small. For example, BARCW separates the interval estimate of Matthew Stafford and Mark Sanchez after down-weighting a few rankers.

Table 1.8: Summary of 13 experts' prediction accuracy evaluated after the 2014 NFL season. Throughout the season, *FantasyPros.com* compare each expert's player preference to the actual outcomes. The prediction accuracy is calculated based on the incremental fantasy points implied by ranking lists.

	"reliable" rankers	"unreliable" rankers
mean (accuracy)	0.589	0.550
sd (accuracy)	0.013	0.027

attempting to pass, are the most significant covariates; touchdowns have a positive effect, while interceptions have a negative one. Based on our football common sense, touchdowns and interceptions can directly impact the result of a game.

1.6.2 Aggregating orthodontics data

As mentioned in Section 1, the orthodontics data set contains 69 ranking lists for each of the 9 groups of the orthodontic cases. With ranking lists produced by a group of high-profile specialists, the rank aggregation problem emerges because the average perception of experienced orthodontists is considered the cornerstone of systems for the evaluation of orthodontic treatment outcome (Liu et al., 2012; Song et al., 2014,

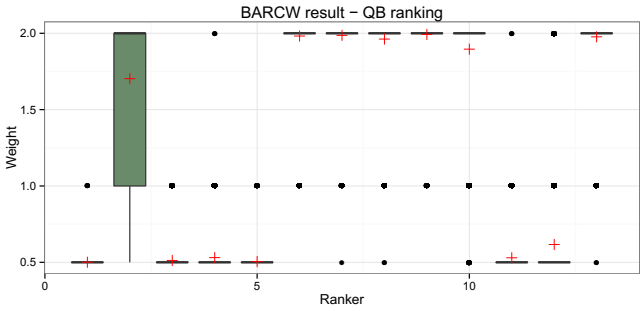


Figure 1.7: Boxplots of posterior samples of weights given by BARCW. Red cross marks the posterior means of weights. Black points are samples outside of the range between first and third quartile of the posterior samples, and black lines are collapsed boxes when interquartile range is 0. Seven experts are learned to be reliable rankers.

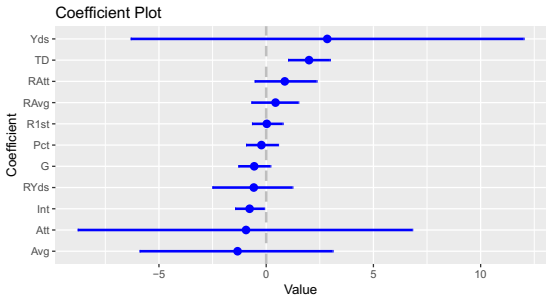


Figure 1.8: Posterior mean and 95% probability interval of β given by BARCW in aggregating quarterback rankings. Please refer Table 1.2 for the covariates information, and each column of covariates are standardized when applied in BARCW.

2015). The covariates for these cases are objective assessments on their teeth. It is quite difficult to aggregate ranking lists of many non-overlapping subgroups, as covariates are the only source of information available in bridging different groups. In addition, Table 1.3 shows that the rankers did not have very similar opinions.

Previously, Liu et al. (2012) and Song et al. (2014) assessed the reliability and the overall consistency of these experienced orthodontists through simple statistics including Spearman's correlation among these highly incomplete ranking lists within each subgroup of cases. To gain deeper understanding of these ranking lists, we first study

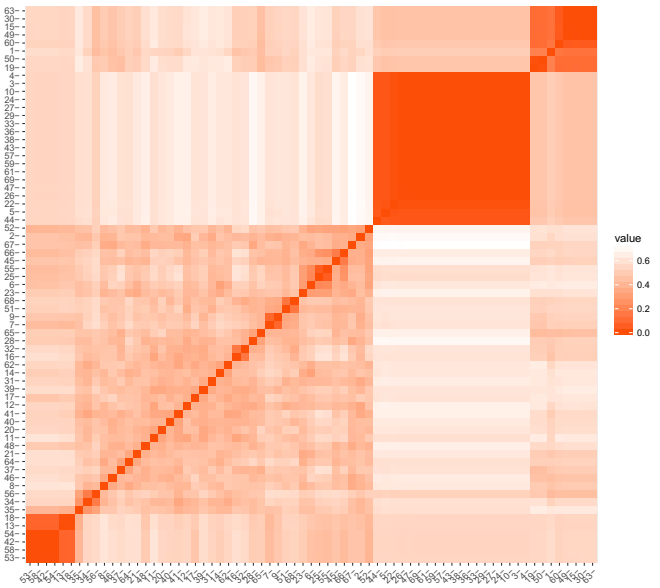


Figure 1.9: Kendall tau distance of posterior mean of $\boldsymbol{\mu}^{(j)}$s based on BARCM result. The ranking discrepancy increases as the color shifts from dark to light.

the heterogeneity among rankers using BARCM. We applied Dirichlet process mixture model with $\gamma = 1$. The 69 experts are clustered into 24 subgroups. The sizes of the leading three clusters are $(19, 9, 6)$. Other clusters have fewer than 5 rankers each. Figure 1.9 shows the Kendall tau distances among the posterior means of the $\boldsymbol{\mu}^{(j)}$s, which are the underlying ranking criteria of all rankers. We see that almost half of the rankers cannot be grouped into sizable clusters, indicating that their opinions are closer to the baseline distribution in Dirichlet process than to other rankers. Because a ranking lists drawn from the baseline distribution is just noise, the rankers in the small groups either were unreliable rankers, or used information other than the available covariates in their ranking systems.

We subsequently applied BARCW to this data set. Figure 1.10 shows the box plot of rankers' weights by their estimated clusters from BARCM. The rankers in the largest

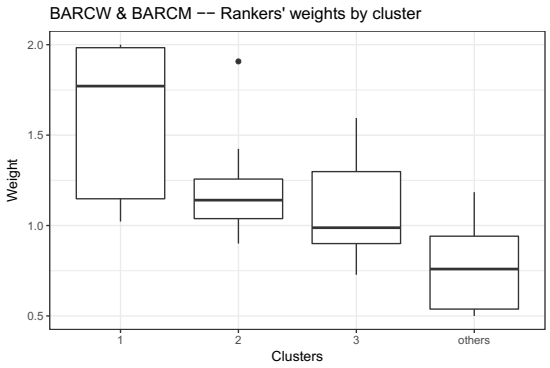


Figure 1.10: Boxplots of rankers' weights by clusters. The weight of ranker j is given by the posterior mean of w_j in BARCW. The clusters are estimated from BARCM ordered by their size. The sizes of clusters 1-3 are $(19, 9, 6)$, while "other" combines all rankers from the remaining 21 fragmented clusters.

cluster are mostly considered reliable rankers, while the rankers in the small clusters are labeled as low-quality rankers. Implied by lower weights, the noisier ranking evaluation explains why the small-size clusters are not combined into the big ones. The weights of rankers in cluster 2 and cluster 3 are around the middle. This result is similar to our demonstration using simulation in Section 1.5.4. Among clusters 1-3, BARCW tends to down-weight the minority opinions when heterogeneous opinions exist. Based on the results from BARCM and BARCW, we conclude that there are three ranking opinions among half of the experts, while the others have considerable discrepancy that can be attributed to low individual reliability.

Finally, we use both BARCW and BARCM for rank aggregation. The key to aggregate these nine non-overlapping groups of patients is to figure out the rank of patients' orthodontics conditions using, but not overly relying on, the covariates. Tables 1.9 and 1.10 show the top and bottom cases in aggregated ranking lists. Recall that BARCM aggregates opinions of the whole sample by averaging over all clusters with their corresponding proportions. The results from BARCW and BARCM are quite consistent

with each other although they employed different assumptions. The Kendall distance between these two aggregated lists is 0.047. It supports our conclusion that rankers' discrepancy can be mostly explained by their heterogeneous reliability.

Table 1.9: The five cases that are considered to have the best conditions based on rank aggregation.

	BARC	BARCW	BARCM	Cl. 1	Cl. 2	Cl. 3
1	H2	G7	G7	E2	A1	G7
2	E2	E2	E2	H3	G7	A1
3	G7	H2	H2	G7	H2	E2
4	H3	H3	H3	H2	E2	A4
5	H4	H4	A1	F8	H4	H4

Table 1.10: The five cases that are considered to have the worst conditions based on rank aggregation.

	BARC	BARCW	BARCM	Cl. 1	Cl. 2	Cl. 3
108	F4	F4	F4	H5	F4	F4
107	H5	H5	F10	F10	F10	F10
106	F10	F10	H5	F4	H5	E6
105	E6	E6	E6	D11	D11	H5
104	D11	D11	D11	E6	E6	E8

Figure 1.11 shows coefficient plot of β in BARCW. It illustrates the role of covariates in our rank aggregation, especially in positioning those non-overlapping groups. Among the covariates, overjet, overbite and centerline all measure certain types of overall displacement, and thus are generally considered to have stronger negative effect compared to the other local displacements in this study.

1.7 Discussion

We described three model-based Bayesian rank aggregation methods (BARC, BARCW, BARCM) for combining different ranking lists when some covariates for the entities in

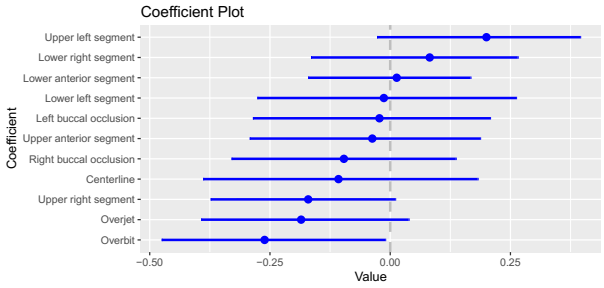


Figure 1.11: Posterior mean and 95% probability interval of β given by BARCW in aggregating orthodontics data. Please refer Table 1.4 for the covariates information, and each column of covariates are standardized when applied in BARCW.

consideration are also observed. With the help of covariates, these methods can accommodate various types of input ranking lists, including highly incomplete ones. Under the assumption of homogeneous ranking opinion, BARCW learns the qualities of rankers from data, and over-weighs high quality ones in rank aggregation. BARCM, on the other hand, studies the possibility of having heterogeneous opinion groups among rankers under the same framework. All three methods generate uncertainty measures for the aggregated ranks. Our simulation studies and real-data applications validate the importance of covariate information and our estimation of rankers' qualities and their heterogeneous opinions.

We note that our methods consider only the covariate information of the ranked entities. It is of interest to further incorporate available covariate information of rankers, which can be helpful for detecting rankers' qualities and clustering rankers into subgroups with different opinions. We leave this extension of BARC for a further work.

The foundation of our rank data modeling is the Thursthone-Mosteller-Daniels model, which can be extended in many ways. Although these models have been around for a long time, the standard MCMC procedure for their Bayesian inference does not mix well for our real-data applications due to the entangled latent structure of the

models. We took advantage of the conjugacy of Gaussian distributions and exploited parameter-expanded Gibbs sampler to improve the computation efficiency. We can also speed up the MCMC algorithm through parallelization when there are many rankers, i.e., when m is large. For all the three models, BARC, BARCW and BARCM, we can parallelize the Gibbs steps for updating $\{\boldsymbol{Z}_j\}_{j=1}^m$ given $\boldsymbol{\mu}$, or equivalently $(\boldsymbol{\alpha}, \boldsymbol{\beta})$. However, this full Bayesian inference still has its limitation in computational scalability when dealing with very large datasets such as those arisen from voting. An interesting future work is to develop approximate likelihood and Bayesian priors for the BARC model family under "big-data" settings that can enable us to do both efficient point estimation and approximate Bayesian inference.

2

Forecasting unemployment using Internet search data

2.1 Introduction

Driven by the growth and wide availability of Internet and online platforms, big data are generated with an unprecedented speed nowadays. They offer the potential to

inform and transform decision making in industry, business, social policy and public health (McKinsey Global Institute, 2011; McAfee & Brynjolfsson, 2012; Chen et al., 2012; Khoury & Ioannidis, 2014; Kim et al., 2014; Murdoch & Detsky, 2013). Big data can be used for developing predictive models for systems that would have been challenging to predict with traditional small-sample-based approaches (Einav & Levin, 2014; Siegel, 2016). For instance, numerous studies have demonstrated the potential of using Internet search data in tracking influenza outbreaks (Ginsberg et al., 2009; Yang et al., 2015), dengue fever outbreaks (Yang et al., 2017), financial market returns (Preis et al., 2013), consumer behaviors (Goel et al., 2010), unemployment (Ettredge et al., 2005; Choi & Varian, 2012) and housing prices (Wu & Brynjolfsson, 2015).

In this article, we focus on using Internet users' Google search to forecast US unemployment initial claims weeks into the future. Unemployment initial claims measure the number of jobless claims filed by individuals seeking to receive state jobless benefits. It is closely watched by government and the financial market, as it provides timely insight into the direction of the economy. A sustained increase of initial claims would indicate rising unemployment and a challenging economic environment.

Weekly unemployment initial claim is the (unadjusted) total number of actual initial claims filed under the Federal-State Unemployment Insurance Program in each week ending on Saturday. The Employment and Training Administration (ETA) of the U.S. Department of Labor collects the weekly unemployment insurance claims reported by each state's unemployment insurance program offices, and releases the data to the public at 8:30 A.M. (eastern time) on the following Thursday. Thus, the weekly unemployment initial claim data are reported with a one-week delay in the sense that the number reported on Thursday of a given week is actually the unemployment initial claim number of the preceding week. For accessing the general economic trend, it is,

therefore, highly desirable for government agencies and financial institutions to predict the unemployment situation of the current week, which is known as nowcasting (Giannone et al., 2008), as well as weeks into the future. In this article, we use the general phrase *forecasting* to cover both nowcasting (the current week) and predicting into future weeks.

In contrast to the unemployment reports by the Department of Labor, which is one week behind real time, Internet users' online search of unemployment related query terms provides highly informative and *real-time* information for the current unemployment situation. For instance, a surge of people's Internet search of "unemployment office", "unemployment benefits", "unemployment extension", etc. in a given week could indicate an increase of unemployment of *that* week compared to the week before, as presumably more people unemployed are searching for information of getting unemployment aid. Therefore, Internet search data, offering a real-time "peek" of the current week, augments the delayed official time-series unemployment data. In this article, we study how to effectively combine the real-time Internet search information and the traditional time series information to forecast unemployment initial claims. The Internet search data that we use are publicly available from Google Trends, which will be detailed in Section 2.1.1.

In developing an effective way to forecast weekly unemployment initial claims with Internet search data, there are several main challenges. First, the volatile seasonality pattern accounts for most of the variation of the targeted time series. Figure 2.1 plots the weekly unemployment initial claims from 2000 to 2016; the seasonal spikes are particularly noteworthy. A prediction method should address and utilize the strong seasonality in order to achieve good prediction performance. Second, the method needs to effectively incorporate the most up-to-date Internet search data into the modeling

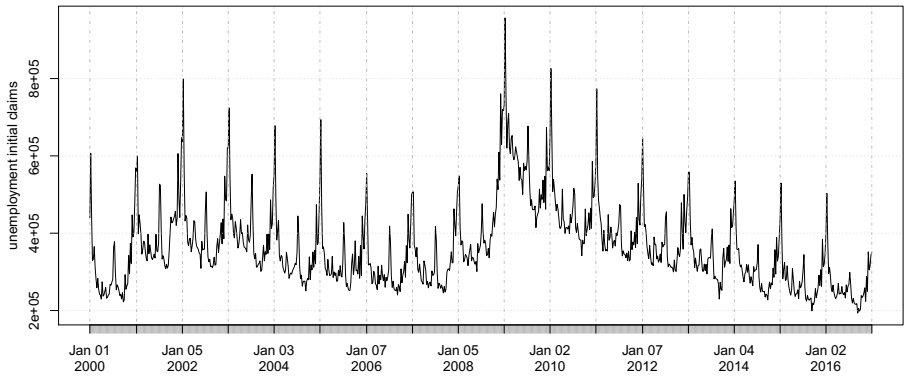


Figure 2.1: Weekly unemployment initial claims in 2000-2016.

of targeted time series. Third, as people's search pattern and the search engine both evolve over time, the method should be able to incorporate this dynamic change.

When dealing with seasonality, most time series models rely on state space models, where the latent components capture the trend and seasonality (Aoki, 1987; Harvey, 1989). Among the time series models, structural time series models and innovation state space models are two main frameworks (Harvey & Koopman, 1993; Durbin & Koopman, 2012; Hyndman et al., 2008). Both frameworks come with various extensions of seasonal pattern modeling and often incorporate exogenous signals as regression components (see Section 2.2 for more detailed discussion).

For nowcasting time series with seasonal pattern, Scott & Varian (2013, 2014) recently developed a Bayesian method based on the structural time series model, and applied the method to nowcast unemployment initial claims with Google search data. Their method took the search data as regressors, and used a spike-and-slab prior for variable selection. Alternative to this regression formulation, Banbura et al. (2013) proposed a nowcasting method using a factor model, in which targeted time series and related exogenous time series are driven by common factors.

In this article, we introduce a novel prediction method PRISM, which stands for Penalized Regression with Inferred Seasonality Module, for forecasting times series with seasonality, and use it to forecast unemployment initial claims. Our method is semi-parametric in nature; it takes advantage of both the state space formulation for time series forecasting and penalized regression that is effective and computationally efficient for large datasets. In particular, we formulate a novel state space model that contains a variety of widely used time series models as special cases, including structural time series models and additive innovation state space models. We then derive a universal predictive model for forecasting initial claim data that is coherent with all possible models under our state space formulation, and develop a two-stage estimation procedure PRISM using nonparametric seasonal decomposition and L_1 penalized regression.

With the semi-parametric method PRISM, we significantly expand the range of time series models for forecasting, going beyond the traditional approaches, which are often tailored for individual parametric models. From a methodological standpoint, PRISM offers a robust and more accurate forecasting alternative to traditional parametric approaches (owing to its robustness against model misspecification). PRISM effectively addresses the three aforementioned challenges in forecasting time series with strong seasonality. First, our method accommodates various nonparametric and model-based seasonal decomposition tools, and effectively incorporates the estimated seasonal components into predictive modeling. It thus can robustly handle complex seasonal patterns. Second, different from the traditional regression formulation, our joint modeling of the targeted time series and the exogenous variables accommodates the potential causal relationship between them — people do online Google search in response of being unemployed. Third, PRISM uses dynamic forecasting — training its predictive

equation each weak for the forecasting — and utilizes rolling window and exponential weighting to account for the time-varying relationship between the targeted time series and the exogenous variables. From an applied standpoint, we applied PRISM to the forecasting of unemployment initial claim data; our method delivers superior performance over all existing forecasting methods for the entire time period of $2007-2016$, and is exceptionally robust to the ups and downs of the general economic environment, including the huge volatility caused by the 2008 financial crisis. While we concentrate the discussion on unemployment initial claims in this article, PRISM can be applied to forecasting other time series with seasonal pattern.

2.1.1 Initial Claims Data and Internet Search Data from Google

The weekly (non-seasonally adjusted) initial claims are our targeted time series. The initial claims for the preceding week are released every Thursday. The time series of the initial claims from 1967 to present is available at https://fred.stlouisfed.org/series/ICNSA. Figure 2.1 shows the weekly initial claims data in 2000-2016.

The real-time Internet search data we use were obtained from Google Trends (www.google.com/trends). The Google Trends website, which is publicly available, provides weekly (relative) search volume of search query terms specified by a user. Specifically, for a user-specified query term, Google Trends provides integer-valued weekly times series (after 2004); each number in the time series, ranging from 0 to 100, represents the search volume of that search query term in a given week divided by the total online search volume of that week; and the number is normalized to integer values from 0 to 100, where 100 corresponds to the maximum weekly search within the time period (also specified by the user). Figure 2.2, the upper panel, illustrates the Google Trend time series of several search query terms in a 5-year span. Comparing these time

series to the lower panel of the unemployment initial claims of the same time period, they evidently provide noisy signal about the latter.

The search query terms that we use in our study are also identified from the Google Trends tool. One feature of Google Trends is that, in addition to the time series of a specific term (or a general topic), it also returns the top query terms that are most highly correlated with the specific term. In our study, we use a list of top 25 Google search terms that are the most highly correlated with the term "unemployment". Table 2.1 lists these 25 terms, which are generated by Google Trends on January 11, 2018; they include 12 general unemployment related query terms, such as *unemployment office*, *unemployment benefits* and *unemployment extension*, as well as 13 query terms that are combinations of state names and "unemployment", such as *California unemployment* and *unemployment Florida*.

Table 2.1: Top 25 nationwide search query terms associated with the term "unemployment" generated by Google Trends as of January 11, 2018.

unemployment	unemployment benefits	unemployment rate
unemployment office	pa unemployment	claim unemployment
ny unemployment	nys unemployment	ohio unemployment
unemployment florida	unemployment extension	texas unemployment
nj unemployment	unemployment number	file unemployment
unemployment insurance	california unemployment	unemployed
unemployment oregon	new york unemployment	indiana unemployment
unemployment washington	unemployment wisconsin	unemployment online
unemployment login		

There is a notable restriction on the length of historical weekly data at the Google Trends website. The weekly data is available for at most a 5-year span in a query, and it would be automatically transformed to monthly data if one asks for more than 5 years. As we are modeling and forecasting the unemployment claims for the entire period of 2007-2016, we downloaded separate weekly data sets from Google Trends, covering 2004-2008, 2006-2010, 2008-2012, 2010-2014 and 2012-2016, respectively.

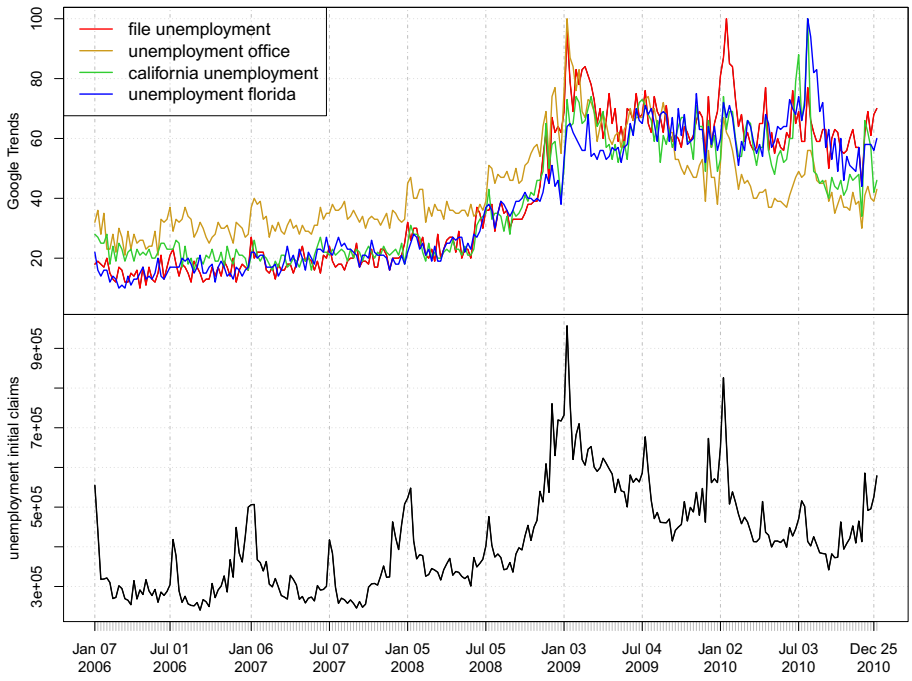


Figure 2.2: The upper panel shows the Google Trends data of four unemployment related search queries in 2006-2010. The lower panel shows the weekly unemployment initial claims data in the same period.

2.1.2 ORGANIZATION OF THE ARTICLE

The rest of the article is organized as follows. Section 2.2 presents our general state space formulation for modeling time series data and highlights a few widely used special cases. The general state space formulation serves to motivate PRISM. Section 2.3 introduces a joint model of time series of interest (initial claims) and contemporaneous exogenous variables (Internet search data). Section 2.4 describes our two-step estimation procedure PRISM for forecasting (including nowcasting) the targeted time series with (and without) Internet search data. Section 2.5 evaluates the performance of our proposed method on forecasting the unemployment initial claims and compares it to the results of other existing time series forecasting methods. Section 2.6 concludes the article with a summary.

2.2 STATE SPACE FORMULATION FOR TIME SERIES WITH SEASONAL PATTERN

In this section, we introduce a state space formulation for univariate time series with seasonal pattern, which contains several widely used time series models as special cases. Let $\{y_t\}$ be the univariate time series of interest, and let $\{\gamma_t\}$ be the unobserved seasonal component of y_t. Define $z_t \stackrel{\triangle}{=} y_t - \gamma_t$ as the seasonally adjusted time series, which is also unobservable. We postulate that $\{z_t\}$ and $\{\gamma_t\}$ each follow a linear state space model with state vectors $\{\boldsymbol{h}_t\}$ and $\{\boldsymbol{s}_t\}$ respectively. The state space formulation is

$$y_t = z_t + \gamma_t \tag{2.2.1a}$$

$$\begin{cases} z_t = \boldsymbol{w}'\boldsymbol{h}_t + \epsilon_t & \text{(2.2.1b)} \\ \boldsymbol{h}_t = \boldsymbol{F}\boldsymbol{h}_{t-1} + \boldsymbol{\eta}_t & \text{(2.2.1c)} \end{cases} \qquad \begin{cases} \gamma_t = \boldsymbol{v}'\boldsymbol{s}_t + \zeta_t & \text{(2.2.1d)} \\ \boldsymbol{s}_t = \boldsymbol{P}\boldsymbol{s}_{t-1} + \boldsymbol{\omega}_t & \text{(2.2.1e)} \end{cases}$$

where $(\epsilon_t, \zeta_t, \boldsymbol{\eta}_t', \boldsymbol{\omega}_t')' \overset{iid}{\sim} \mathcal{N}(\mathbf{0}, \boldsymbol{H})$. The parameters are $\boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{F}, \boldsymbol{v}, \boldsymbol{P}, \boldsymbol{H})$.

Our state space model contains a variety of widely used time series models, including structural time series models and additive innovation state space models. Under the general formulation (2.2.1), a specific parametric model can be obtained by specifying the state space models for z_t and γ_t along with their dependence structure $\boldsymbol{H}$.

2.2.1 Special Cases

We highlight a few special cases of model (2.2.1) in this subsection.

Seasonal AR model

The state space formulation (2.2.1) contains the following AR model with seasonal pattern: modeling z_t as an autoregressive process with lag N and assuming a dummy variable formulation with period S for the seasonal component γ_t:

$$\begin{aligned} y_t &= z_t + \gamma_t, \\ z_t &= \mu_z + \sum_{j=1}^{N} \alpha_j z_{t-j} + \eta_t, \quad \eta_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\eta^2) \\ \gamma_t &= -\sum_{j=1}^{S-1} \gamma_{t-j} + \omega_t, \quad \omega_t \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\omega^2) \end{aligned} \tag{2.2.2}$$

The dummy variable model for the seasonal component implies that sum of the seasonal components over the S periods, $\sum_{j=0}^{S-1} \gamma_{t-j}$, has mean zero and variance σ_ω^2.

The seasonal AR model (2.2.2) tells us that each time series block of $\{z_{(t-N+1):t}\}_{t \geq N}$ and $\{\gamma_{(t-S+2):t}\}_{t \geq (S-1)}$ evolves as a Markov Chain. Under our general state-space model (2.2.1), if we set $\boldsymbol{h}_t = (1, z_t, z_{t-1}, \ldots, z_{t-N+1})$ and $\boldsymbol{s}_t = (\gamma_t, \gamma_{t-1}, \ldots, \gamma_{t-S+2})$, then it reduces to the seasonal AR model (2.2.2).

Structural Time Series Models

The basic structural model assumes that a univariate time series is the sum of trend, seasonal and irregular components, each of which follows an independent stochastic process (Harvey, 1989). The model is

$$y_t = \mu_t + \gamma_t + \epsilon_t, \qquad \epsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\epsilon^2), \tag{2.2.3}$$

where μ_t is the trend component, and γ_t and ϵ_t are the seasonal and irregular components, respectively.

The trend is often specified by a local level model

$$\mu_t = \mu_{t-1} + \delta_t + \eta_t, \qquad \eta_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\eta^2), \tag{2.2.4a}$$

$$\delta_t = \delta_{t-1} + \zeta_t, \qquad \zeta_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\zeta^2), \tag{2.2.4b}$$

where μ_t is the level and δ_t is the slope. η_t and ζ_t are assumed mutually independent.

For time series with S periods, the seasonal component can be specified through the seasonal dummy variable model

$$\gamma_t = -\sum_{j=1}^{S-1} \gamma_{t-j} + \omega_t, \qquad \omega_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_\omega^2). \tag{2.2.5}$$

which is the same as the seasonal component in the seasonal AR model (2.2.2). Alternatively, the seasonal pattern can be modeled by a set of trigonometric terms at seasonal

frequencies $\lambda_j = 2\pi j/S$ (Harvey, 1989):

$$\gamma_t = \sum_{j=1}^{[S/2]} \gamma_{j,t}, \tag{2.2.6a}$$

$$\begin{pmatrix} \gamma_{j,t} \\ \gamma^*_{j,t} \end{pmatrix} = \begin{pmatrix} \cos\lambda_j & \sin\lambda_j \\ -\sin\lambda_j & \cos\lambda_j \end{pmatrix} \begin{pmatrix} \gamma_{j,t-1} \\ \gamma^*_{j,t-1} \end{pmatrix} + \begin{pmatrix} \omega_{j,t} \\ \omega^*_{j,t} \end{pmatrix}, \tag{2.2.6b}$$

where $\omega_{j,t}$ and $\omega^*_{j,t}$, $j = 1, \ldots, [S/2]$, are independent and normally distributed with common variance σ^2_ω.

Under our general state-space model (2.2.1), if we take $z_t = \mu_t + \epsilon_t$ and $\boldsymbol{h}_t = (\mu_t, \delta_t)$, then it specializes to structural time series models. In particular, for the dummy variable seasonality of equation (2.2.5), $\boldsymbol{s}_t$ in model (2.2.1) corresponds to $\boldsymbol{s}_t = (\gamma_t, \gamma_{t-1}, \ldots, \gamma_{t-S+2})$; and for the trigonometric seasonality of equation (2.2.6), $\boldsymbol{s}_t$ in model (2.2.1) corresponds to $\boldsymbol{s}_t = (\gamma_{1,t}, \ldots, \gamma_{[S/2],t}, \gamma^*_{1,t}, \ldots, \gamma^*_{[S/2],t})$.

Additive innovations State Space Models

An alternative to structural time series models, which have multiple sources of error, innovation state space model (Aoki, 1987), where the same error term appears in each equation, is also popular. These innovation state space models underlie exponential smoothing methods, which are widely used in time series forecasting and have been proven optimal under many specifications of the innovation state space model (Ord et al., 1997; Hyndman et al., 2008). Among exponential smoothing methods, Holt-Winters' method (Holt, 1957; Winters, 1960) is developed to capture both trend and seasonality, and it postulates a model specification similar to the basic structural model

(2.2.3)- (2.2.5). In particular, Holt-Winters' additive method is

$$
\begin{aligned}
y_t &= \mu_{t-1} + \delta_{t-1} + \gamma_{t-S} + \epsilon_t, && \text{(2.2.7a)}\\
\mu_t &= \mu_{t-1} + \delta_{t-1} + \alpha\epsilon_t, && \text{(2.2.7b)}\\
\delta_t &= \delta_{t-1} + \beta\epsilon_t, && \text{(2.2.7c)}\\
\gamma_t &= \gamma_{t-S} + \omega\epsilon_t, && \text{(2.2.7d)}
\end{aligned}
$$

where the components μ_t, δ_t and γ_t represent level, slope and seasonal components of time series, and $\epsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma^2)$ is the only source of error.

Since equation (2.2.7a) can be rewritten as

$$y_t = \mu_t + \gamma_t + (1 - \alpha - \omega)\epsilon_t,$$

we observe that model (2.2.7) is special case of our general model (2.2.1) with $z_t = \mu_t + (1 - \alpha - \omega)\epsilon_t$, $\boldsymbol{h}_t = (\mu_t, \delta_t)$ and $\boldsymbol{s}_t = (\gamma_t, \gamma_{t-1}, \ldots, \gamma_{t-S+1})$.

The Holt-Winters model is among a collection of innovation state space models that Hyndman et al. (2008) summarizes using the triplet (E,T,S), representing model specification for the three components: error, trend and seasonality. For instance, equation (2.2.7) is also referred to as local additive seasonal model or ETS(A,A,A), where A stands for additive. Our general state space formulation (2.2.1) also incorporates many useful model extensions as special cases, including the damped trend (Gardner Jr & McKenzie, 1985) and multiple seasonal patterns (Gould et al., 2008; De Livera et al., 2011). For example, the damped trend double seasonal model extends model (2.2.7) to

include a factor $\phi \in [0, 1)$ and a second seasonal component as follows:

$$\begin{aligned} y_t &= \mu_{t-1} + \phi\delta_{t-1} + \gamma^{(1)}_{t-S_1} + \gamma^{(2)}_{t-S_2} + \epsilon_t, \\ \mu_t &= \mu_{t-1} + \phi\delta_{t-1} + \alpha\epsilon_t, \\ \delta_t &= \phi\delta_{t-1} + \beta\epsilon_t, \\ \gamma^{(1)}_t &= \gamma^{(1)}_{t-S_1} + \omega_1\epsilon_t, \\ \gamma^{(2)}_t &= \gamma^{(1)}_{t-S_2} + \omega_2\epsilon_t. \end{aligned} \tag{2.2.8}$$

Our general model (2.2.1) contains this extended model as well, where $z_t = \mu_t + (1 - \alpha - \omega_1 - \omega_2)\epsilon_t$, $\gamma_t = \gamma^{(1)}_t + \gamma^{(2)}_t$, $\boldsymbol{h}_t = (\mu_t, \delta_t)$ and $\boldsymbol{s}_t = (\gamma^{(1)}_t, \ldots, \gamma^{(1)}_{t-S_1+1}, \gamma^{(2)}_t, \ldots, \gamma^{(2)}_{t-S_2+1})$.

2.2.2 The general formulation

The motivation of our general state space formulation (2.2.1) is to collectively consider all possible models under it and to *semi-parametrically* obtain the prediction under this large class of models. In comparison, traditional time series studies often rely on parameter estimation of specified models such as those highlighted in the previous subsection. For instance, exponential smoothing is tailored for computing the likelihood and obtaining maximum likelihood estimates of the innovation state space models. For other parametric models with multiple sources of error, their likelihood might be evaluated by the Kalman filter, but the parameter estimation can be difficult in many cases. In the traditional parametric times series model setting, model selections are often applied by optimizing certain selection criteria (e.g. AIC or BIC), but when the class of models under consideration become really large such as (2.2.1), traditional model selection methods encounter serious challenges (as they lack scalability) to operate on such a wide range of models. As a consequence, traditional parametric time series models

often consider a much smaller collection of models compared to (2.2.1). The cost of focusing on a small class of models is that the forecasting accuracy can substantially suffer as the risk of model misspecification is high.

To relieve these challenges and improve the performance of forecasting, we will use our general state space formulation (2.2.1) as a motivation to introduce a semi-parametric method for forecasting time series. We will derive and study a linear predictive model that is coherent with all possible models under (2.2.1). With forecasting as our main goal, we essentially transfer the question from the inference of a complicated class of state space models into penalized regression and forecasting based on a linear prediction formulation.

2.3 JOINT MODEL WITH EXOGENOUS TIME SERIES

2.3.1 JOINT MODELING

In this section, we consider modeling of univariate time series y_t with contribution from contemporaneous exogenous variables $\boldsymbol{x}_t$. In the particular case of forecasting unemployment initial claims, the exogenous variables are the weekly Internet search data from Google Trends. Let $\boldsymbol{x}_t = (x_{1,t}, x_{2,t}, \ldots, x_{p,t})'$ be the vector of the (normalized) search volumes of p search terms at week t. We postulate a state space model for the Google Trends variables $\boldsymbol{x}_t$ on top of y_t, instead of adding them as regressors as in traditional models. In particular, at each time t, we assume a multivariate normal distribution for $\boldsymbol{x}_t$ conditional on the level of unemployment initial claims y_t,

$$\boldsymbol{x}_t | y_t \sim \mathcal{N}_p(\boldsymbol{\mu}_x + y_t \boldsymbol{\beta}, \boldsymbol{Q}) \tag{2.3.1}$$

where $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$, $\boldsymbol{\mu}_x = (\mu_{x_1}, \mu_{x_2}, \ldots, \mu_{x_p})'$ and $\boldsymbol{Q}$ is the covariance matrix. $\boldsymbol{x}_t$ is assumed to be independent of $\{y_l, \boldsymbol{x}_l : l < t\}$ conditional on y_t. For $\{y_t\}$ following the general state space model (2.2.1), our joint model for initial claims and Google Trends search data can be diagrammed as follows:

$$\begin{array}{ccccccc} \cdots & \rightarrow & (\boldsymbol{s}_t, \boldsymbol{h}_t) & \rightarrow & (\boldsymbol{s}_{t+1}, \boldsymbol{h}_{t+1}) & \rightarrow & \cdots \\ & & \downarrow & & \downarrow & & \\ & & y_t & & y_{t+1} & & \\ & & \downarrow & & \downarrow & & \\ & & \boldsymbol{x}_t & & \boldsymbol{x}_{t+1} & & \end{array} \tag{2.3.2}$$

Our joint model (2.3.2) of $\boldsymbol{x}_t$ and y_t is related to the factor models of multivariate time series (Forni et al., 2000; Stock & Watson, 2002; Harvey & Koopman, 1997). It can be interpreted as that the multivariate time series $\{\boldsymbol{x}_t\}$ is driven by a common factor $\{y_t\}$, which is observed with lags.

2.3.2 Comparing to traditional modeling of exogenous time series

In contrast to our joint model, traditional methods usually treat the contemporaneous exogenous variables as regressors. For example, Harvey & Shephard (1993) regards the univariate structural time series models as regression models in which the explanatory variables are functions of time and the parameters are time-varying. Thus, the addition of exogenous variables became an extension to observation equation (2.2.3) in traditional methods, leading to

$$y_t = \mu_t + \gamma_t + \boldsymbol{x}_t'\boldsymbol{\beta} + \epsilon_t, \ t = 1, \ldots, T, \tag{2.3.3}$$

where $\{\mu_t\}$ and $\{\gamma_t\}$ follows the same transition equations as in (2.2.4) and (2.2.5) or (2.2.6).

Scott & Varian (2014) developed a full Bayesian inference of model (2.3.3) with variable selection, termed Bayesian Structural Time Series (BSTS). With $\boldsymbol{x}_t$ being contemporaneous Google search data, BSTS assumes spike-and-slab prior for $\boldsymbol{\beta}$, through which a high degree of sparsity of the regression coefficients is achieved (Scott & Varian, 2013).

The model (2.3.3) can be depicted by diagram (2.3.4) below. Adding contemporaneous Google search data as regression components to state space model have been useful for nowcasting economic time series (Scott & Varian, 2014). However, this model structure (2.3.3) is unlikely to correspond to the data generating process in our case in the sense that it is quite feasible that people search the Internet for information about unemployment aid in response to their being unemployed, rather than the other way around.

$$\begin{array}{ccccccccc} \cdots & \rightarrow & \left(\mu_t, \delta_t, \gamma_{(t-S+2):t}, \boldsymbol{x}_t\right) & \rightarrow & \left(\mu_{t+1}, \delta_{t+1}, \gamma_{(t-S+3):(t+1)}, \boldsymbol{x}_{t+1}\right) & \rightarrow & \cdots \\ & & \downarrow & & \downarrow & & \\ & & y_t & & y_{t+1} & & \end{array} \tag{2.3.4}$$

2.4 Forecasting with PRISM

Based on our joint model of $\{y_t\}$ and $\{\boldsymbol{x}_t\}$, in this section we construct a structural predictive model and propose a two-step estimation procedure PRISM, which stands for Penalized Regression with Inferred Seasonality Module, for forecasting time series $\{y_t\}$ using its lagged values and the available exogenous variables $\{\boldsymbol{x}_t\}$ as input. The

structural predictive model accommodates a collection of state space model of $\{y_t\}$, including structural time series models and additive innovation state space models, as we have seen in Section 2.2.

2.4.1 Overview of our two-stage methodology for nowcast and forecast

We first provide an overview of the proposed methodology for nowcasting y_t and forecasting y_{t+l} ($l \geq 1$), i.e., weeks into the future, using all available information at time t. The derivation and rationale of each step will be described in the following subsections.

Input: Target time series $\{y_{1:(t-1)}\}$ and exogenous time series $\{\boldsymbol{x}_{t_0:t}\}$. In our forecasting of unemployment initial claims, $\{y_{1:(t-1)}\}$ is the weekly unemployment initial claim data reported with one-week lag, and $\{\boldsymbol{x}_{t_0:t}\}$ is the multivariate Google Trends data starting from 2004.

Stage 1 of PRISM. Seasonal decomposition: For a fixed rolling window length M, nonparametrically decompose $\{y_{(t-M):(t-1)}\}$ into estimated seasonal component $\{\hat{\gamma}_{i,t}\}_{i=(t-M),\ldots,(t-1)}$ and estimated seasonally adjusted component $\{\hat{z}_{i,t}\}_{i=(t-M),\ldots,(t-1)}$, where $\hat{\gamma}_{i,t}$ and $\hat{z}_{i,t}$ are estimates of γ_i and z_i using data available at time t.

Stage 2 of PRISM. Penalized regression: Forecast (and nowcast) target time series using the following predictive equation:

$$\hat{y}_{t+l} = \mu + \sum_{j=1}^{K} \alpha_j \hat{z}_{t-j,t} + \sum_{j=1}^{K} \delta_j \hat{\gamma}_{t-j,t} + \sum_{i=1}^{p} \beta_i x_{i,t}$$

where the coefficients above are estimated by a rolling-window L_1 penalized linear regression using historical data for each forecasting horizon l.

2.4.2 PREDICTIVE MODEL FOR NOWCASTING

Under our general state space model (2.2.1) and (2.3.1), given the historical data of $\{y_{1:(t-1)}\}$ and contemporaneous exogenous time series $\{\boldsymbol{x}_{t_0:t}\}$, the predictive distribution for nowcasting y_t at time t would be $p(y_t \mid y_{1:(t-1)}, \boldsymbol{x}_{t_0:t})$. In PRISM, we consider, instead, the predictive distribution of y_t by further conditioning on the latent seasonal component $\{\gamma_t\}$:

$$p(y_t \mid y_{1:(t-1)}, \gamma_{1:(t-1)}, \boldsymbol{x}_{t_0:t}) \propto p(\boldsymbol{x}_t \mid y_t) p(y_t \mid y_{1:(t-1)}, \gamma_{1:(t-1)}). \tag{2.4.1}$$

Note that since $z_t = y_t - \gamma_t$ for all t, $z_{1:(t-1)}$ is known given $y_{1:(t-1)}$ and $\gamma_{1:(t-1)}$. The advantage of working on (2.4.1) is that we can establish a universal representation of this predictive distribution as given by the next proposition.

Proposition 1 *Under model (2.2.1) and (2.3.1), y_t conditioning on $\{z_{1:(t-1)}, \gamma_{1:(t-1)}, \boldsymbol{x}_{t_0:t}\}$ follows a normal distribution with the conditional mean $E(y_t \mid z_{1:(t-1)}, \gamma_{1:(t-1)}, \boldsymbol{x}_{t_0:t})$ linear in $z_{1:(t-1)}$, $\gamma_{1:(t-1)}$ and $\boldsymbol{x}_t$.*

Remark 1 *As a partial result that lead to Proposition 1, $y_t \mid z_{1:(t-1)}, \gamma_{1:(t-1)}$ follows normal distribution with mean linear in $z_{1:(t-1)}$ and $\gamma_{1:(t-1)}$ under model (2.2.1).*

Based on Proposition 1, the predictive distribution $p(y_t \mid y_{1:(t-1)}, \gamma_{1:(t-1)}, \boldsymbol{x}_{t_0:t})$ is given by

$$y_t = \mu_t + \sum_{j=1}^{t-1} \alpha_{j,t} z_{t-j} + \sum_{j=1}^{t-1} \delta_{j,t} \gamma_{t-j} + \sum_{i=1}^{p} \beta_{i,t} x_{i,t} + \epsilon_t, \quad \epsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_t^2) \tag{2.4.2}$$

where μ_t, $\alpha_{j,t}$, $\delta_{j,t}$, $\beta_{i,t}$ and σ_t^2 are fixed but unknown constants that are determined by original parameters $\boldsymbol{\theta}$ and the initial values of the state vectors.

2.4.3 Predictive model for forecasting into future weeks

As the exogenous variables $\boldsymbol{x}_t$ carry information about time t, they also help forecast future y_{t+l} ($l \geq 1$). Under the same framework as for nowcasting, we can calculate the predictive distribution of y_{t+l} conditioning on $z_{1:(t-1)}$, $\gamma_{1:(t-1)}$ and $\boldsymbol{x}_{t_0:t}$.

Proposition 2 *Under model (2.2.1) and (2.3.1), the predictive distribution* $p(y_{t+l} \mid z_{1:(t-1)}, \gamma_{1:(t-1)}, \boldsymbol{x}_{t_0:t})$ *for* $l \geq 1$ *is normal with the conditional mean* $E(y_{t+l} \mid z_{1:(t-1)}, \gamma_{1:(t-1)}, \boldsymbol{x}_{t_0:t})$ *being a linear combination of* $z_{1:(t-1)}$, $\gamma_{1:(t-1)}$ *and* $\boldsymbol{x}_t$.

Remark 2 *Similar to Remark 1, under model (2.2.1),* $y_{t+l} \mid z_{1:(t-1)}, \gamma_{1:(t-1)}$ *follows a normal distribution with mean* $E(y_{t+l} \mid z_{1:(t-1)}, \gamma_{1:(t-1)})$ *linear in* $z_{1:(t-1)}$ *and* $\gamma_{1:(t-1)}$.

Based on Proposition 2, we can represent $p\left(y_{t+l} \mid z_{1:(t-1)}, \gamma_{1:(t-1)}, \boldsymbol{x}_{t_0:t}\right)$ as

$$y_{t+l} = \mu_t^{(l)} + \sum_{j=1}^{t-1} \alpha_{j,t}^{(l)} z_{t-j} + \sum_{j=1}^{t-1} \delta_{j,t}^{(l)} \gamma_{t-j} + \sum_{i=1}^{p} \beta_{i,t}^{(l)} x_{i,t} + \epsilon_t, \quad \epsilon_t \overset{iid}{\sim} \mathcal{N}(0, \sigma_{t,l}^2). \tag{2.4.3}$$

2.4.4 The two stages of PRISM

We now describe our semi-parametric estimation procedure PRISM for nowcasting y_t and forecasting y_{t+l} ($l \geq 1$) using all available information at time t.

Stage 1: seasonal decomposition

In Propositions 1 and 2, the predictive distribution for y_t is universal for all possible models under our general state space model (2.2.1) and (2.3.1) once conditioned on

the historical seasonal components $\gamma_{1:(t-1)}$. Since $\gamma_{1:(t-1)}$ is unobserved, we estimate these seasonal components in the first stage of our semi-parametric estimation procedure. For this purpose, various seasonal decomposition methods can be applied here, including nonparametric methods such as the classical additive seasonal decomposition (Kendall & Stuart, 1983) and parametric methods based on innovation state space models.

We used the method of Seasonal and Trend decomposition using Loess (STL) by Cleveland et al. (1990) as the default choice. The STL method is nonparametric; it is widely used and robust for decomposing time series with few assumptions owing to its nonparametric nature. Unlike the classic additive seasonal decomposition, STL allows the seasonal components to change over time, and the rate of change can be controlled by the user; the smoothness of the trend-cycle can also be controlled by the user. We describe the procedure of STL in Appendix B.2, where we also used PRISM with the classic additive seasonal decomposition. We found that PRISM is robust to the choice of seasonal decomposition methods with STL performing slightly better. We therefore take STL as the default choice due to its simplicity and ease of use.

Under the default setting of STL, at every time t for forecasting, we apply STL to historical initial claims observations $y_{(t-M):(t-1)}$ with M being a large number. For each rolling window from $t-M$ to $t-1$, STL decomposes the univariate time series $y_{(t-M):(t-1)}$ into three components: seasonal, trend and the remainder. Denote $\hat{\gamma}_{i,t}$ and $\hat{z}_{i,t}$ as the estimates of γ_i and z_i using data available at time t. Then, at each t the STL decomposition generates estimated seasonal component time series $\{\hat{\gamma}_{i,t}\}_{i=(t-M),\ldots,(t-1)}$ and seasonally adjusted time series $\{\hat{z}_{i,t}\}_{i=(t-M),\ldots,(t-1)}$; the latter is the sum of trend component and remainder component from STL. In our forecasting of unemployment initial claims, we take $M = 700$.

STAGE 2: PENALIZED LINEAR REGRESSION

In the second stage, we use the predictive equations (2.4.2) and (2.4.3) with the estimated seasonal components and estimated seasonally adjusted components from the previous stage for prediction. Specifically, for each fixed l (≥ 0), we estimate y_{t+l} by the following linear predictive equation:

$$\hat{y}_{t+l} = \mu_y + \sum_{j=1}^{K} \alpha_j \hat{z}_{t-j,t} + \sum_{j=1}^{K} \delta_j \hat{\gamma}_{t-j,t} + \sum_{i=1}^{p} \beta_i x_{i,t}, \tag{2.4.4}$$

where for notational ease we have used the generic notations μ_y, α_j, δ_j etc. with the understanding that there is a separate set of $\{\mu_y, \boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K), \boldsymbol{\delta} = (\delta_1, \ldots, \delta_K), \boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)\}$ for each l. At each time t and for each horizon l, the regression coefficients μ_y, $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ are obtained by minimizing

$$\begin{aligned} \frac{1}{N} \sum_{\tau=t-l-N}^{t-l-1} w^{t-\tau} \left(y_{\tau+l} - \mu_y - \sum_{j=1}^{K} \alpha_j \hat{z}_{\tau-j,\tau} - \sum_{j=1}^{K} \delta_j \hat{\gamma}_{\tau-j,\tau} - \sum_{i=1}^{p} \beta_i x_{i,\tau} \right)^2 \\ + \lambda_1 \left(\|\boldsymbol{\alpha}\|_1 + \|\boldsymbol{\delta}\|_1 \right) + \lambda_2 \|\boldsymbol{\beta}\|_1, \end{aligned} \tag{2.4.5}$$

where N is the length of a rolling window, w is a discount factor, and λ_1 and λ_2 are nonnegative regularization parameters.

There are several distinct features of our estimation procedure at Stage 2. First, a rolling window of length N is employed. This is to address the fact that the parameters in the predictive equation (2.4.4) can vary with time t. In our case, people's search pattern and the search engine tend to evolve over time, and it is quite likely that the same search phrases would contribute in different ways over time to the response variable. Correspondingly, the coefficients in (2.4.4) need to be estimated dynamically each

week, and the more recent observations should be considered more relevant than the distant historical observations for inferring the predictive equations of current time. The rolling window of observations and the exponentially decreasing weights are utilized for such purpose. Our use of exponential weighting is related to the weighted least square formulation that is usually referred to as discounted weighted regression in the econometrics literature (Ameen & Harrison, 1984; Taylor, 2010).

Second, since the number of unknown coefficient in (2.4.4) tends to be quite large compared to the number of observations within the rolling window, we apply L_1 regularization in our rolling-window estimation (Tibshirani, 1996), which gives robust and sparse estimate of the coefficients. Up to two L_1 penalties are applied: on $(\boldsymbol{\alpha}, \boldsymbol{\delta})$ and on $\boldsymbol{\beta}$, as they represent two sources of information — information from time series components $\{\hat{z}_t\}$ and $\{\hat{\gamma}_t\}$, and information from the exogenous variables $\{\boldsymbol{x}_t\}$.

Third, PRISM is a semi-parametric method. The predictive equation (2.4.4) is motivated and derived from our state space formulation (2.2.1). However, the estimation is not parametric in that (i) the seasonal and seasonally adjusted components are learned non-parametrically from Stage 1, and (ii) the coefficients in equation (2.4.4) are dynamically estimated each week in Stage 2. Combined together, the two stages of PRISM gives us a simple and robust estimation procedure. This approach is novel and different from the typical approaches for linear state space models, which often estimate unknown parameters using specific parametrizations and select a model based on information criteria (Hyndman et al., 2008).

In the case when exogenous time series $\{\boldsymbol{x}_t\}$ do not exist, PRISM uses Remarks 1

and 2 and estimates y_{t+l} according to the following linear predictive equation:

$$\hat{y}_{t+l} = \mu_y + \sum_{j=1}^{K} \alpha_j \hat{z}_{t-j,t} + \sum_{j=1}^{K} \delta_j \hat{\gamma}_{t-j,t}, \tag{2.4.6}$$

which is a degenerated special case of the predictive equation (2.4.4). Under the same estimation procedure as in (2.4.5) except that $\boldsymbol{\beta}$ and $\boldsymbol{x}_t$ are dropped, predictive equation (2.4.6) can be used to forecast univariate time series with seasonal patterns without exogenous time series. We will later use this predictive equation to quantify the contribution from Internet search data in forecasting unemployment initial claims.

2.4.5 PREDICTIVE INTERVALS

The semi-parametric nature of PRISM makes it more difficult to construct predictive intervals on PRISM forecasts, as we cannot rely on parametric specifications, such as posterior distributions, for predictive interval construction. However, the fact that we are forecasting time series suggests a (non-parametric) method for us to construct predictive intervals based on the historical performances of PRISM.

For nowcasting at time t, given the historical data available up to time $t-1$, we can evaluate the root mean square error of nowcasting for the last L time periods as

$$\mathrm{se}_t = \sqrt{\frac{1}{52} \sum_{\tau=t-L}^{t-1} (\hat{y}_\tau - y_\tau)^2},$$

where $\hat{y}_\tau$ is the pseudo real time PRISM estimate for y_τ generated at time τ. Under the assumption of stationarity, se_t would be an estimate for the standard error of $\hat{y}_t$. We can thus use it to construct predictive interval for the current PRISM estimate. An $1-\alpha$ predictive interval is given by $(\hat{y}_t - z_{\alpha/2}\ \mathrm{se}_t,\ \hat{y}_t + z_{\alpha/2}\ \mathrm{se}_t)$, where $z_{\alpha/2}$ is the $1-\alpha/2$

quantile of the standard normal distribution. The predictive intervals for forecasting into future weeks can be constructed similarly. We will study in Section 2.5.3 the performance of our predictive intervals.

2.5 Application to unemployment initial claim data

In this section, we apply PRISM to forecasting weekly unemployment initial claims. The unemployment initial claims data $\{y_t\}$ is available from 1967 onward, but Google Trends data $\boldsymbol{x}_t$ is available only since 2004. We thus take $2007 - 2016$ as the testing period. In the test, we let the forecasting horizon $l = 0, 1, 2, 3$ to predict y_t up to 3 weeks ahead.

We compare PRISM to four alternative methods: (a) Bayesian Structural Time Series (BSTS) (Scott & Varian, 2014), (b) and (c), two forecasting methods using exponential smoothing: BATS and TBATS (De Livera et al., 2011), and (d) the naive method, which without any modeling effort simply uses y_{t-1} to predict y_t, y_{t+1}, y_{t+2} and y_{t+3} at time t. The naive method serves as a baseline. Both BATS and TBATS are based on innovation state space framework that contains model (2.2.7) and (2.2.8) as special cases. The name BATS is an acronym for key features of the model: Box-Cox transform, ARMA errors, Trend, and Seasonal components. TBATS extends BATS to handle complex seasonal patterns with trigonometric representations, and the initial T connotes "trigonometric".

2.5.1 Specification of PRISM for forecasting unemployment initial claims

In our forecasting of unemployment initial claims, we apply a 3-year rolling window of historical data to estimate the parameters in (2.4.5), i.e. $N = 156$ (weeks). We take $K = 52$ (weeks) to employ the most recent 1-year estimated seasonal and seasonally adjusted components, and $p = 25$ (Google search terms) according to the list of top 25 nationwide query terms related to "unemployment" in Table 2.1. In addition, we take $w = 0.99$ as the default choice of the discount factor following Lindoff (1997), which suggest that setting the discount factor between 0.95 and 0.995 works in most applications. We tested the choice of w in Appendix B.3, and indeed found that the performance of PRISM is quite robust for $w \in [0.95, 0.995]$.

For the regularization parameters λ_1 and λ_2 in (2.4.5), we use cross-validation. We find empirically that the extra flexibility of having two separate λ_1 and λ_2 does not give improvement over fixing $\lambda_1 = \lambda_2$. In particular, we found that for every forecasting horizon $l = 0, 1, 2, 3$, in the cross-validation process of setting (λ_1, λ_2) for separate L_1 penalty, over 80% of the weeks showed that the smallest cross-validation mean error when restricting $\lambda_1 = \lambda_2$ is within 1 standard error of the global smallest cross-validation mean error. For model simplicity, we thus choose to further restrict $\lambda_1 = \lambda_2$ when forecasting unemployment initial claims.

2.5.2 Forecasting Results

In generating retrospective estimates of initial claims, we rerun all methods each week using only the information available up to that week, i.e., we use the same information in the retrospective estimation as if we relived the testing period of $2007 - 2016$.

The two exponential smoothing methods BATS and TBATS do not offer the option of including exogenous variable in their forecasting, while BSTS allows the inclusion of exogenous variables. Therefore, to forecast unemployment initial claims at time t, both PRISM and BSTS take the initial claim data $\{y_{1:(t-1)}\}$ and Google Trends data $\{\boldsymbol{x}_{t_0:t}\}$ as input, whereas BATS and TBATS are trained using available historical initial claim data $\{y_{1:(t-1)}\}$ at each week t. For fair comparison, both PRISM and BSTS are fitted with a 3-year rolling window of exogenous variables at each week. The results of BSTS, BATS and TBATS are produced by their respective R packages under their default settings.

To quantify the contribution of the exogenous variables (i.e., the contemporaneous Google Trends data) as compared to using time series alone, we also use the degenerated predictive equation (2.4.6) for the retrospective forecasting. The predictive equation (2.4.6) is estimated without the exogenous variables $\boldsymbol{x}_t$ — it is fitted under the same procedure as PRISM except that $\boldsymbol{\beta}$ and $\boldsymbol{x}_t$ are dropped in Stage 2. For simplicity, we denote this method as "PRISM w/o $\boldsymbol{x}_t$" in the coming exhibitions.

We use root-mean-square error (RMSE) and mean absolute error (MAE) as accuracy metrics in the evaluation of the performance of different methods. For an estimator $\{\hat{y}_t\}$ and horizon l, the RMSE and MAE are defined, respectively, as $\text{RMSE}_l(\hat{y}_t, y_t) = [\frac{1}{(n_2-n_1-l+1)}\sum_{t=n_1+l}^{n_2}(\hat{y}_t - y_t)^2]^{1/2}$ and $\text{MAE}_l(\hat{y}_t, y_t) = \frac{1}{(n_2-n_1-l+1)}\sum_{t=n_1+l}^{n_2}|\hat{y}_t - y_t|$, where we denote $n_1 + l$ and n_2 respectively as the start and end of the forecasting for each l.

Table 2.2 presents the overall performance of forecasting (including nowcasting) unemployment initial claims over the entire period of $2007 - 2016$. The RMSE and MAE numbers reported here are relative to the naive method, which uses y_{t-1} to predict y_{t+l} ($l \geq 0$) at time t. BSTS does not produce numbers for forecasting y_{t+l} for $l \geq 1$, as its R package gives prediction of the target time series only as far as exogenous variables are inputed.

Table 2.2 reveals the following. First, PRISM uniformly outperforms all the other methods for the entire period of $2007-2016$ under all forecasting horizons. Second, contemporaneous Google Trends data is very helpful for real-time nowcasting, as PRISM and BSTS have better nowcasting results than the other methods that only utilize historical initial claim data. Third, the contribution of contemporaneous Google Trends data becomes less significant in forecasting future weeks, as evidenced by the performance gap between PRISM and "PRISM w/o $\boldsymbol{x}_t$" shrinking from nowcasting to forecasting. Fourth, among the three methods that only use historical initial claim data, the predictive method based on PRISM without $\boldsymbol{x}_t$ outperforms the exponential smoothing methods BATS and TBATS.

Table 2.2: The performance of different methods over the time period of $2007-2016$. RMSE and MAE here are relative to the error of naive method; that is, the number reported is the ratio of the error of a given method to that of the naive method. The naive method use y_{t-1} to predict y_{t+l}, and the absolute RMSE and MAE of the naive method are reported in the parentheses. The boldface indicates the best performer for each forecasting horizon and each accuracy metric.

	real-time	forecast 1 wk	forecast 2 wk	forecast 3 wk
RMSE				
PRISM	**0.498**	**0.492**	**0.453**	**0.467**
PRISM w/o $\boldsymbol{x}_t$	0.659	0.534	0.501	0.527
BSTS	0.588	-	-	-
BATS	1.002	0.897	0.848	0.832
TBATS	0.711	0.559	0.544	0.528
naive	1 (50551.3)	1 (62226.6)	1 (69746.5)	1 (73528.9)
MAE				
PRISM	**0.542**	**0.534**	**0.479**	**0.465**
PRISM w/o $\boldsymbol{x}_t$	0.670	0.561	0.507	0.502
BSTS	0.612	-	-	-
BATS	0.992	0.898	0.825	0.781
TBATS	0.750	0.599	0.570	0.525
naive	1 (33636.6)	1 (41120.8)	1 (47902.3)	1 (52793.7)

Figure 2.3 shows the RMSE of the yearly nowcasting results of the different methods. Here, RMSE are measured relative to the error of the naive method. It is seen that

PRISM gives consistent relative RMSE throughout the $2007-2016$ period. It is noteworthy that PRISM outperforms all other methods in 2008 and 2009 when the financial crisis caused significant instability in the US economy.

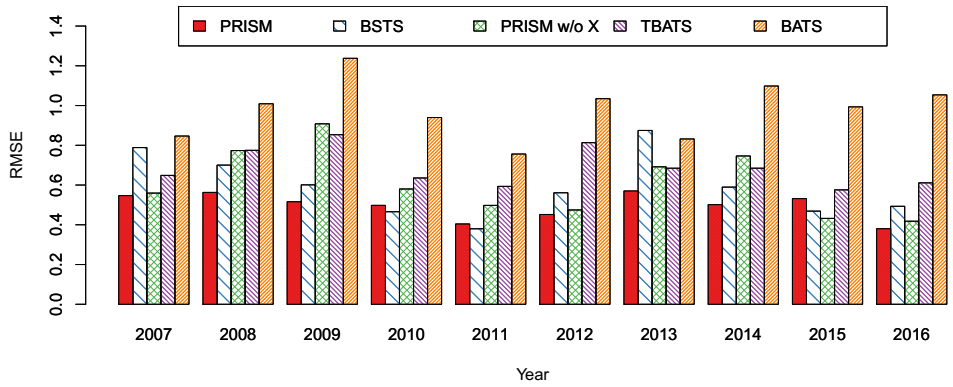


Figure 2.3: Yearly nowcasting performance. RMSE are measured relative to the error of the naive method, i.e., the numbers are the ratio of the error of a given method to that of the naive method.

For a closer look of the performance of different methods, Figure 2.4 shows how the absolute errors of nowcasting accumulate through $2007-2016$. Compared to the other methods, the cumulated absolute error of PRISM rises at the slowest rate. As shown in the lower panel, the 2008 financial crisis caused significantly more unemployment initial claims. PRISM handles the financial crisis period well, as the accumulation of errors is rather smooth for the financial crisis period. Other methods all accumulate loss in a considerably higher rate during the financial crisis. Furthermore, PRISM handles the strong seasonality of initial claim data well, since the accumulation of error is smooth within each year. Among all the methods considered, BATS is visibly bumpy in handling seasonality, as the accumulation jumps when the initial claim data spikes.

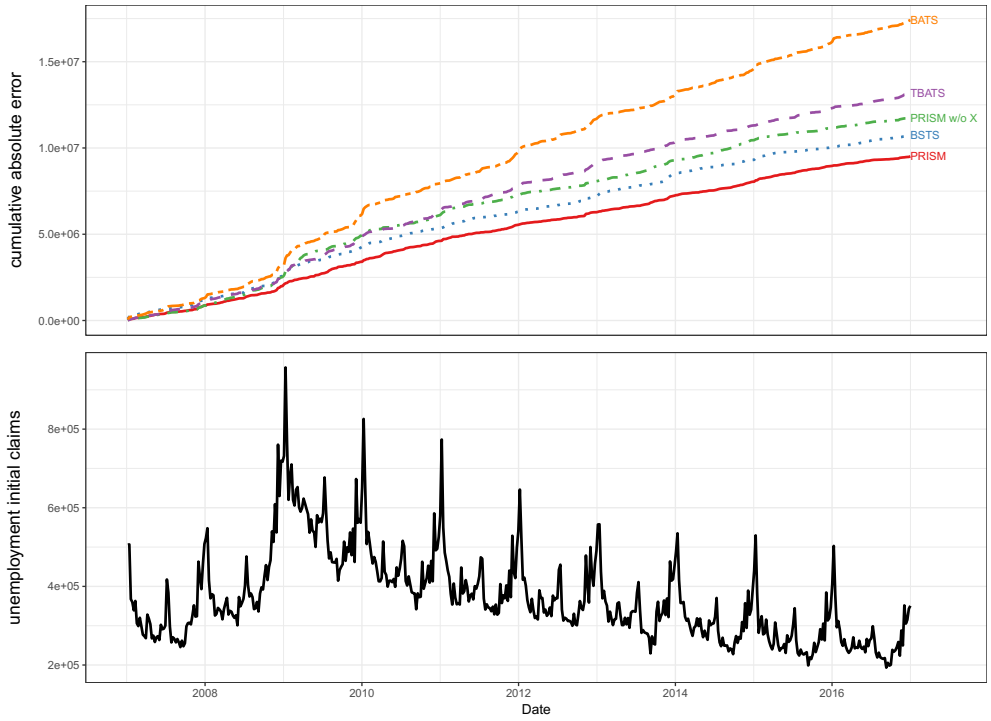


Figure 2.4: The top panel displays the cumulative absolute error of nowcasting given by the different methods. The bottom panel shows the unemployment initial claims y_t for the same period of $2007 - 2016$.

2.5.3 Assessing the predictive intervals

Using the method of Section 2.4.5, we can construct the predictive intervals for PRISM estimates. In particular, we take $L = 52$ and obtain $\text{se}_t = (\frac{1}{52}\sum_{\tau=t-52}^{t-1}(\hat{y}_\tau - y_\tau)^2)^{1/2}$ based on the estimation result of the past year (52 weeks). Figure 2.5 shows for nowcasting the point estimates and 95% predictive intervals of PRISM based on $(\hat{y}_t - 1.96\ \text{se}_t, \hat{y}_t + 1.96\ \text{se}_t)$ for $2008 - 2016$, comparing to the true unemployment initial claims officially revealed a week later (in blue). Note that since we need one year of historical performance to compute se_t, we can evaluate the predictive intervals starting from 2008. For $2008 - 2016$, the actual coverage of the predictive interval is 95.3%, which is very close to the nominal 95%.

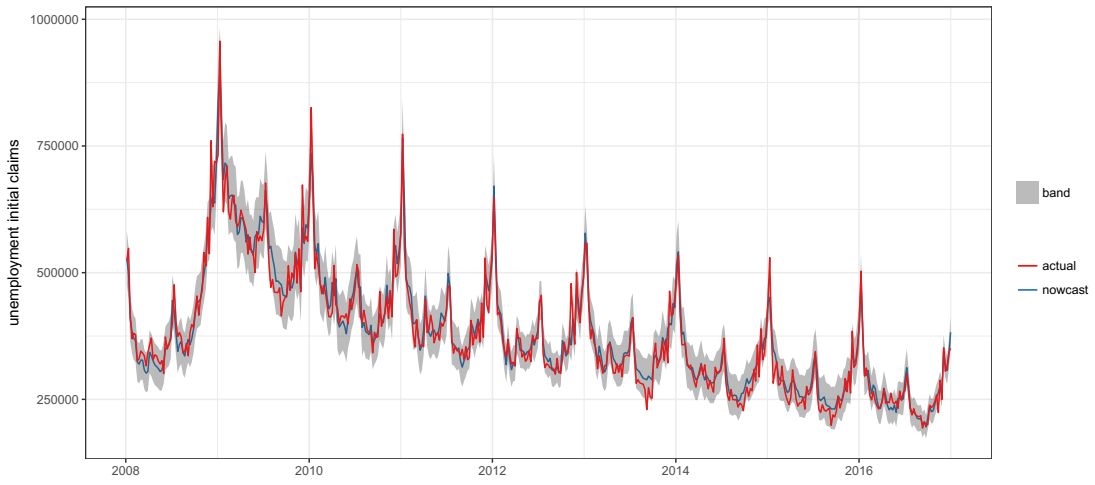


Figure 2.5: Predictive Interval of PRISIM. The shaded area corresponds to the 95% predictive interval of PRISM. The red curve is the point estimate of PRISM nowcasting. The blue curve is the true unemployment initial claims. The actual coverage of the PRISM predictive interval is 95.3% in $2008 - 2016$.

2.6 SUMMARY

The wide availability of data generated from the Internet offers great potential for predictive analysis and decision making. In this article we focus on using Internet search data to forecast unemployment initial claims weeks into the future. We introduced a novel statistical method PRISM for forecasting time series with strong seasonality. PRISM is semi-parametric and can be generally applied with or without exogenous variables. PRISM is motivated from a general state-space formulation that contains a variety of widely used time series models as special cases. The two stages of PRISM are easy to implement. The numerical evaluation shows that PRISM outperforms all alternatives in forecasting unemployment initial claim data for the time period of $2008-2016$. It is noteworthy that PRISM's performance is robust throughout the whole testing period, particularly during the $2008-2009$ financial crisis.

Although this article focuses on forecasting unemployment initial claims, PRISM can be generally used to forecast time series with complex seasonal patterns, owing to its nonparametric seasonal decomposition at the first stage. The semi-parametric approach of PRISM covers a wider range of time series models than traditional methods, as PRISM transfers the inference of a complicated class of state space models into penalized regression of linear predictive models. In addition, our joint modeling with contemporaneous exogenous variables accommodates the data generation process; for example, intuitively, people search online for unemployment benefits related information in response of being unemployed. Furthermore, dynamically fitting the predictive equations of PRISM addresses the time-varying relationship between the exogenous variables and the underlying time series. The R package that implements the PRISM method is available at https://github.com/ryanddyi/prism.

Data derived from the Internet has presented many opportunities and interesting problems for statisticians. Our study on using Google search data to forecast unemployment initial claims illustrates one such example. The arrival of new data (sometimes in new forms) requires new methodology to analyze and utilize them. PRISM is an example where traditional statistical modeling are brought together with more recent statistical tools, such as L_1 regularization and dynamic training. We hope our study will serve to generate further interest in developing statistical methodology to big data problems.

3

Bayesian Factor Model with Multiple Change-points

3.1 Introduction

In various scientific and industrial applications, it is crucial to obtain an accurate estimation of the covariance among a large number of measurements that varies over time.

In the financial industry, the estimated covariances among asset returns are key inputs for portfolio construction and risk management (Markowitz, 1952; Fan et al., 2012). In cognitive science, estimated correlations from fMRI time series help explain the interactive functions of the human brain (Barnett & Onnela, 2016). In both applications, as the time series unfolds the covariance structure evolves as well, so it is desirable to determine the points in time when a structural change takes place, in order to model the full evolution process of the covariance.

To accurately estimate a changing covariance matrix, we are faced with two major challenges. One concerns the variable dimension of the time series, p, relative to the length of time series, T. This is otherwise known as the "$p > T$" scenario in high-dimensional statistics, in which the sample covariance matrix estimator is singular since the number of unknown parameters that remain to be estimated clearly grows fast with p. The second challenge concerns the principled modeling of the underlying covariance structure, which changes over time.

For tackling the high-dimensionality issue, well-established methodologies mostly rely on factor model, which assume individual time series are driven by a small number of common latent factors and a set of sparsely correlated idiosyncratic errors. With a long history and wide range of applications, many factor models were proposed for time series analysis and frequently used in economic and financial studies (Chamberlain & Rothschild, 1983; Stock & Watson, 2002; Bai & Ng, 2002). In particular, they were studied as high-dimensional covariance estimation methods for stationary time series in statistical literature (Fan et al., 2013).

As for the second challenge, change-point model is arguably the simplest model to accommodate the time-varying nature of the covariance. Change-point model partitions the observational time frame into multiple segments, and the time series is

modeled as stationary within each segment. Although multiple change-point detection problems have been studied under general frameworks from both Bayesian and frequentist perspective (Fearnhead, 2006; Killick et al., 2012), only a very few studies concerned change-point models for covariance matrices of high-dimensional time series. Barigozzi et al. (2016) studies high-dimensional time series factor models with multiple change-points in their second-order structure, while Ma & Su (2016) and Sun et al. (2017) consider a similar problem for a smaller number of time series. They exploit multiple stage methodologies, which combine principle component analysis (PCA)-type factor model estimation and change-point detection methods like binary segmentation.

The goal of our work is to provide a Bayesian analysis of high-dimensional time series factor model with multiple change-points. In contrast to PCA-type estimation of factor models, Bayesian factor analysis can generate interpretable factors by exploiting patterns of sparsity in factor loading matrix (Carvalho et al., 2008; Knowles & Ghahramani, 2011; Ročková & George, 2016a). Without giving up interpretability of latent factors, we built our model assuming that the changes in covariance are driven by the variances of underlying factors and thus are low-dimensional.

Our work is also related to a collection of multivariate GARCH and stochastic volatility (SV) models. Reviewed by Bauwens et al. (2006) and Asai et al. (2006), multivariate GARCH and SV models have been much more thoroughly studied than change-point models for the time-varying covariance estimation, especially for financial time series data. A leading example in this category is dynamic conditional correlation (DCC) model which considers the time-varying feature of variance and correlation separately (Engle, 2002; Tse & Tsui, 2002). However, when the multivariate time series is in hundreds of dimensions, most existing methods cannot provide satisfactory

solutions to the problem of time-varying covariance estimation, as highlighted in Engle et al. (2008). In particular, the computational challenges emerge when inverting large p-dimensional covariance matrices, which is required as part of likelihood estimation by many methods. Bollerslev (1990) and Engle (2007) proposed methodologies in line with traditional multivariate GARCH setting but with less computation burdens. Simple heuristic methods such as sample covariance matrix of rolling windows or RiskMetrics exponential smoother (Longerstaey & Spencer, 1996) are widely used in industry for this type of problems. Another line of work impose factor model structure on multivariate GARCH and SV models (Harvey et al., 1994; Diebold & Nerlove, 1989; Aguilar & West, 2000; Chib et al., 2006).

In practice, although GARCH and SV models are generally preferred over change-point models for financial time series due to the ever-changing volatility, our change-point model can generate explainable underlying factors and can detect meaningful structural break when applied on top of a separate volatility model as shown in real data example later. In addition, we apply our factor model with change-points to fMRI time series to study the connectivity of human brain and potential structure changes during a sequence of experiments.

3.2 Bayesian change-point model for covariance matrix

Before considering high-dimensional time-varying covariance, we study a Bayesian change-point model for covariance matrix under low-dimensional setting. Let $\{\boldsymbol{y}_t\}_{1:T}$ be a d-dimensional time series and $1 \leq \tau_1 < \cdots < \tau_N \leq T-1$ be N change-points for the underlying covariance $\boldsymbol{\Sigma}_t$, then the time span is divided into $N+1$ non-overlap segments. We let $\tau_0 = 0$ and $\tau_{N+1} = T$, then the jth segment contains time series

data $\boldsymbol{y}_{(\tau_{j-1}+1):\tau_j}$. We assume the mean of $\boldsymbol{y}_t$ is $\mathbf{0}$, which is the same assumption as in multivariate GARCH and SV models. Then, for all time t in the jth segment, the model is

$$\boldsymbol{y}_t \overset{iid}{\sim} \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_j\right), \ \tau_{j-1} < t \leq \tau_j. \tag{3.2.1}$$

Similar to Fearnhead (2006), we consider a Bayesian analysis for this specific change-point detection problem. From the Bayesian perspective, we treat the number and the positions of change-points as random variables instead of fix but unknown constants. The prior of change points is a point process specified by the probability mass function $g(t)$ for the time between two successive points. We utilize negative binomial distribution here,

$$g(t) = \binom{t-1}{k-1} p^k (1-p)^{t-k}, \tag{3.2.2}$$

where $t = k, k+1, k+2, \ldots$ This negative binomial formulation counts the number of trials given k success. The point process is determined by parameter k and p, and the expected length between two successive change-points is k/p. Larger k implies that larger distance between two successive change-points. When $k = 1$, the point process is Markov.

Note that $\tau_0 = 0$ is not a true change-point, so the gap between τ_0 and τ_1 is likely to be shorter than the other gaps $\tau_j - \tau_{j-1}$ on average. Thus, the prior probability mass function for τ_1 is specified as

$$g_0(t) = \sum_{i=1}^{k} \binom{t-1}{i-1} p^i (1-p)^{t-i} / k, \tag{3.2.3}$$

where $t = 1, 2, 3, \ldots$ With $G(t) = \sum_{s=1}^{t} g(s)$ and $G_0(t) = \sum_{s=1}^{t} g_0(s)$, the probability

mass function for $\tau_{1:T}$ is

$$p(\tau_{1:N}) = g_0(\tau_1) \prod_{j=2}^{N} g(\tau_j - \tau_{j-1}) \left(1 - G(T - \tau_N - 1)\right) \tag{3.2.4}$$

for $N \geq 1$, and $p(N = 0) = 1 - G_0(T - 1)$.

In addition, we assume independent priors for the parameters associated with each segment. For computation simplicity, we let the prior for $\mathbf{\Sigma}_j$ be inverse Wishart distribution

$$p(\mathbf{\Sigma}_j) \propto |\mathbf{\Sigma}_j|^{-\frac{\nu+d+1}{2}} \exp\left(-\frac{1}{2} tr(\boldsymbol{S}\mathbf{\Sigma}_j^{-1})\right) \tag{3.2.5}$$

where $\boldsymbol{S}$ is positive definite and $\nu > d + 1$.

3.2.1 Exact Bayesian inference

With the priors specified above, we can carry out a exact and efficient Bayesian inference for this change-point in covariance problem. Denote

$$Q(t) = P(\boldsymbol{y}_{t:T} \mid \text{changepoint at } t - 1)$$

with $Q(1) = P(\boldsymbol{y}_{1:T})$ and $P(t, s) = P(\boldsymbol{y}_{t:s} \mid t, s \text{ in the same segment})$. With inverse Wishart prior as (3.2.5), we have

$$P(t, s) = \int \prod_{i=t}^{s} p(\boldsymbol{y}_i \mid \mathbf{\Sigma}) p(\mathbf{\Sigma}) d\mathbf{\Sigma} = \frac{|\boldsymbol{S}|^{\frac{\nu}{2}} \Gamma_d\left(\frac{\nu+s-t+1}{2}\right)}{\pi^{\frac{(s-t+1)d}{2}} |\boldsymbol{S} + \boldsymbol{A}|^{\frac{\nu+s-t+1}{2}} \Gamma_d\left(\frac{\nu}{2}\right)} \tag{3.2.6}$$

where $\boldsymbol{A} = \sum_{i=t}^{s} \boldsymbol{y}_i \boldsymbol{y}_i'$ and Γ_d is the multivariate gamma function.

With $P(\cdot)$ and $Q(\cdot)$ defined above, previous literatures have developed the recursion formula to calculate $Q(t)$ for $t = 1, \ldots, T$ with computation complexity $O(T^2)$ (Yao,

1984; Barry & Hartigan, 1993; Fearnhead, 2006). The recursion is given by

$$Q(t) = \sum_{s=t}^{T-1} P(t,s)Q(s+1)g(s+1-t) + P(t,T)(1-G(T-t)), \tag{3.2.7}$$

for $t = 2, \ldots, T$, and

$$Q(1) = \sum_{s=1}^{T-1} P(1,s)Q(s+1)g_0(s) + P(1,T)(1-G_0(T-1)), \tag{3.2.8}$$

where $G_0(t) = \sum_{s=1}^{t} g_0(s)$.

Given the values of $Q(t)$ for $t = 1, \ldots, T$, we can sample change-points from the posterior distribution as following. The posterior of the first change-point is

$$\begin{aligned} P(\tau_1 \mid \boldsymbol{y}_{1:T}) &= P(\tau_1)P(\boldsymbol{y}_{1:\tau_1} \mid \tau_1)P(\boldsymbol{y}_{(\tau_1+1):T} \mid \tau_1)/Q(1) \\ &= P(1,\tau_1)Q(\tau_1+1)g(\tau_1)/Q(1) \end{aligned}$$

for $\tau_1 = 1, ..., T-1$. The probability of no further change-point being $P(1,n)(1-G(T-1))/Q(1)$. The posterior distribution of the τ_j given τ_{j-1} is

$$P(\tau_j \mid \tau_{j-1}, \boldsymbol{y}_{1:T}) = P(\tau_{j-1}+1, \tau_j)Q(\tau_j+1)g(\tau_j - \tau_{j-1})/Q(\tau_{j-1}+1),$$

for $\tau_j = \tau_{j-1}+1, ..., T-1$, and the probability of no further breakpoint is $P(\tau_{j-1}+1, T)(1-G(n-\tau_{j-1}-1))/Q(\tau_{j-1}+1)$.

Due to the matrix determinant computation in (3.2.6), the computation complexity of this Bayesian analysis is more than quadratic with respect to d. Thus, it is too expensive when d, the dimension of time series is large.

3.3 BAYESIAN FACTOR MODEL WITH MULTIPLE CHANGE-POINTS

In this section, we study factor model with multiple change-points from a Bayesian perspective. Let $\tau_{1:N}$ be the change-point positions and N be the number of change-points. Assuming that the multivariate time series $\boldsymbol{y}_{1:T}$ is driven by K common factors, we study the following factor model

$$\boldsymbol{y}_t = \boldsymbol{B}_{d\times K}\boldsymbol{f}_t + \boldsymbol{\epsilon}_t,\ \boldsymbol{\epsilon}_t \sim \mathcal{N}_d\left(\boldsymbol{0}, \boldsymbol{\Sigma}\right),\ \boldsymbol{f}_t \sim \mathcal{N}_K\left(\boldsymbol{0}, \boldsymbol{\Lambda}_j\right) \tag{3.3.1}$$

for $\tau_{j-1} < t \leq \tau_j$, where $\boldsymbol{B}$ is the matrix of factor loadings, $\boldsymbol{f}_t$ is the vector of latent factors at time t, and $\boldsymbol{\epsilon}_t$ is the vector of idiosyncratic errors that have covariance $\boldsymbol{\Sigma} = \text{diag}\left\{\sigma_1^2, \ldots, \sigma_d^2\right\}$. Thus, the distribution of $\boldsymbol{y}_t$ unconditional on latent factors $\boldsymbol{f}_t$ is

$$\boldsymbol{y}_t \mid \boldsymbol{B}_{d\times K}, \boldsymbol{\Sigma}, \tau_{1:N}, \boldsymbol{\Lambda}_{1:N+1} \overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{B}\boldsymbol{\Lambda}_j\boldsymbol{B}' + \boldsymbol{\Sigma}) \tag{3.3.2}$$

for $\tau_{j-1} < t \leq \tau_j\ (j = 1, \ldots, N+1)$.

Due to identifiability issue, traditional factor model usually assumes that the K latent factors $\boldsymbol{f}_t$ are independent and have unit variance, i.e. $\boldsymbol{f}_t \sim \mathcal{N}\left(\boldsymbol{0}, \boldsymbol{I}_K\right)$. In contrast, we assume the K common factors independent, but we allow the variance of factors to be time-varying. In other words, we impose change-point model to the covariance of the K independent common factors, and the covariance is $\boldsymbol{\Lambda}_j = \text{diag}\left\{\lambda_{j1}^2, \ldots, \lambda_{jK}^2\right\}$ for the jth segment.

Under this factor model, the change in covariance of a high-dimensional time series is driven by the variance change of a few major common factors. The similar intuition can be found in many classic papers about factor multivariate SV models (Jacquier et al., 1995; Aguilar & West, 2000). In our model, when there is a change point at time t,

the difference bewteen $\text{cov}(\boldsymbol{y}_t)$ and $\text{cov}(\boldsymbol{y}_{t+1})$ is at most of rank K. Hence, the change of a high-dimensional covariance matrix lies in a low-dimensional space.

We first consider the prior distribution for model (3.3.2) with K, the dimension of latent factor space fixed. We assume the following prior for change-points $\tau_{1:N}$

$$p(\tau_{1:N}) \propto \exp\left(-\frac{1}{2}NK\log T\right) \tag{3.3.3}$$

where N, K and T are the number of change-points, factors and observations correspondingly. We chose prior (3.3.3) over prior (3.2.4) mainly due to computational convenience, as fast change-point detection method can be applied if $\log p(\tau_{1:N})$ is linear in N. In the context of traditional multiple change-point detection setting, $\log p(\tau_{1:N})$ serves the role as a penalty to log-likelihood, and $K\log T$ is Bayesian information criterion (BIC; Schwarz et al. (1978)), because K is the number of additional parameters introduced by adding a change-point.

For the diagonal elements in $\boldsymbol{\Sigma}$, we assume independent scaled inverse chi-square priors

$$\sigma_1^2, \ldots, \sigma_d^2 \overset{iid}{\sim} \text{Scale-Inv-}\chi^2\left(\xi, s_\sigma^2\right) \tag{3.3.4}$$

with the relatively noninfluential choice $\xi = 1$. For the diagonal elements in $\boldsymbol{\Lambda}_j$ in each segment, we assume scaled inverse chi-square priors

$$\lambda_{j1}^2, \ldots, \lambda_{jK}^2 \overset{iid}{\sim} \text{Scale-Inv-}\chi^2\left(\eta, s_\lambda^2\right) \tag{3.3.5}$$

with $\eta = 1$.

To automatically identify interpretable factor orientations, we exploit a spike-and-slab prior on the individual coefficients in factor loading $\boldsymbol{B}$ so that the coefficients

have high probability to be zero in their posterior distribution. The spike-and-slab prior have been used in many Bayesian sparse factor analysis to naturally separate important coefficients from the coefficients that are ignorable (Carvalho et al., 2008; Knowles & Ghahramani, 2011; Ročková & George, 2016a). In particular, we assume that each factor loading β_{ik} follows spike-and-slab LASSO (SSL) prior that is a mixture distribution of two Laplace components: a slab component with a penalty λ_1 and a spike component with a penalty λ_{0k} (Ročková & George, 2016b). The prior is

$$\beta_{ik} \mid \gamma_{ik}, \delta_0, \delta_1 \sim (1 - \gamma_{ik})\, \phi\left(\beta_{ik} \mid \delta_0\right) + \gamma_{ik} \phi\left(\beta_{ik} \mid \delta_1\right), \tag{3.3.6}$$

where $\phi\left(\beta \mid \delta\right) = \frac{\delta}{2} \exp\left\{-\delta|\beta|\right\}$ and $\delta_1 << \delta_0$ $(i = 1, \ldots, d; k = 1, \ldots, K)$. This SSL prior pull the unselected $(\gamma_{ik} = 0)$ coefficient β_{ik} sharply towards zero with δ_0 substantially larger than λ_1, and thus lead to a more sparse factor loading $\boldsymbol{B}$ compared to traditional factor analysis and spike-and-slab priors with continuous Gaussian spike distributions (George & McCulloch, 1993). As a result, each learned latent factor can be linked to a subset of individual time series of $\{\boldsymbol{y}_t\}$, and many of the factors can be labeled based on background information. In other words, this technique provides enhanced interpretability when applied to time series of hundreds of dimensions.

Binary matrix $\boldsymbol{\Gamma} = (\gamma_{ik})_{d \times K}$ is regarded as feature allocation matrix, since γ_{ik} indicates whether the kth common factor contributes to the ith individual time series. When the number of factor K is fixed, the prior for γ_{ik} can simply be

$$\gamma_{ik} \overset{iid}{\sim} \text{Bernoulli}(\theta), \quad k = 1, \ldots, K, \tag{3.3.7}$$

with $\theta = \frac{1}{2}$, implying equal prior probability for the spike component and the slab

component.

3.3.1 Determination of the number of factors

The remaining issue is how to determine the dimensionality of latent factor space, as it is generally unrealistic to assume the number of latent factors is known. To determine the number of factors in Bayesian factor analysis, Ročková & George (2016a) considered a truncated finite dimensional approximation of Indian Buffett Process in the prior construction for binary matrix $\mathbf{\Gamma} = (\gamma_{ik})_{d\times K^*}$, with K^* being a predetermined maximum possible number of factors and the columns of $\mathbf{\Gamma}$ having decreasing probability to be nonzero from left to right. The number of nonzero columns essentially determines the number of factors.

In a similar fashion, we let K^* being a predetermined maximum possible number of factors. Then, we introduce random variable K such that

$$\gamma_{ik} \overset{iid}{\sim} \text{Bernoulli}(\theta_1), \quad k = 1, \ldots, K, \tag{3.3.8a}$$

$$\gamma_{ik} \overset{iid}{\sim} \text{Bernoulli}(\theta_0), \quad k = K+1, \ldots, K^*, \tag{3.3.8b}$$

where $\theta_1 >> \theta_0$, which is close to 0. Under this model, the last $K^* - K$ columns of $\mathbf{\Gamma}_{d\times K^*}$ are likely to be zero, implying that last $K^* - K$ columns of factor loading matrix $\boldsymbol{B}_{d\times K^*}$ are likely to be zero. We regard the first K factors as active factors and last $K^* - K$ factors as inactive ones.

With K^* being a large number, it is likely that $\boldsymbol{f}_t$ contains many inactive factors, which have little influence to $\boldsymbol{y}_t$. We further assume that the inactive factors follow a

stationary distribution through time. In other words, for $\tau_{j-1} < t \leq \tau_j \, (j = 1, \ldots, N)$

$$\boldsymbol{f}_t \sim \mathcal{N}_{K^*} \left(\mathbf{0}, \boldsymbol{\Lambda}_j \right) \tag{3.3.9}$$

where $\boldsymbol{\Lambda}_j = \text{diag} \left\{ \lambda_{j1}^2, \ldots, \lambda_{jK}^2, \lambda_0^2, \ldots, \lambda_0^2 \right\}$ and λ_0^2 is a universal variance for inactive factors. Under this model, change-point occurs only in the K-dimensional subspace of active factors. Hence, the number of active factors determines the additional effective number of parameters when adding change-points. Following the same representation as (3.3.3), we assume that the joint distribution of change-points $\tau_{1:N}$ and the number of active factors K is

$$p \left(\tau_{1:N}, K \right) \propto \exp \left(-\frac{1}{2} NK \log T \right). \tag{3.3.10}$$

With model setting (3.3.9) and prior (3.3.10), we have the change-points $\tau_{1:N}$, the covariances of factors $\boldsymbol{\Lambda}_{1:(N+1)}$ and binary matrix $\boldsymbol{\Gamma}$ all dependent on the number of active factors K. We carry out an estimation procedure of our model with unknown number of factors in the next section.

3.4 EM APPROACH TO FACTOR ANALYSIS WITH CHANGE-POINTS

We estimate the Bayesian factor model with multiple change-points through an EM approach, which takes advantage of existing implementations for static factor model (Ročková & George, 2016a) and algorithms for multiple change-point detection (Killick et al., 2012). For simplicity, we denote $\boldsymbol{Y}_{d \times T} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)$ and $\boldsymbol{F}_{K^* \times T} = (\boldsymbol{f}_1, \ldots, \boldsymbol{f}_T)$.

The marginal distribution of y_t is determined by $\tau_{1:N}$, $\boldsymbol{\Lambda}_{1:(N+1)}$, $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ as in (3.3.2). Besides marginal distribution of y_t, we are also interested in the number of active factors K. Thus, we consider EM algorithm (Dempster et al., 1977) by treating binary ma-

trix $\mathbf{\Gamma}$ and factor matrix $\boldsymbol{F}$ as missing data, while treating N, $\tau_{1:N}$, $\mathbf{\Lambda}_{1:(N+1)}$, K, $\boldsymbol{B}$ and $\mathbf{\Sigma}$ as parameters of interest. With the prior we assumed in the previous section, our latent factor model with multiple change-points can be depicted by diagram (3.4.1).

$$\begin{array}{ccccccc} \tau_{1:N} & \rightarrow & \mathbf{\Lambda}_{1:(N+1)} & \rightarrow & \boldsymbol{F} & \rightarrow & \boldsymbol{Y} \\ \uparrow & \nearrow & & & & \nearrow & \uparrow \\ K & \rightarrow & \mathbf{\Gamma} & \rightarrow & \boldsymbol{B} & & \mathbf{\Sigma} \end{array} \tag{3.4.1}$$

Let $\mathbf{\Omega} = \left(\tau_{1:N}, \mathbf{\Lambda}_{1:(N+1)}, K, \boldsymbol{B}, \mathbf{\Sigma}\right)$. We propose EM algorithm to find the parameter $\widehat{\mathbf{\Omega}}$ to maximize the posterior distribution given observed data $\boldsymbol{Y}$, i.e. $\widehat{\mathbf{\Omega}} = \arg\max_{\mathbf{\Omega}} \log p\left(\mathbf{\Omega} \mid \boldsymbol{Y}\right)$. Given an initialization $\mathbf{\Omega}^{(0)}$, EM algorithm seeks to find $\widehat{\mathbf{\Omega}}$ by iteratively applying these following two steps:

Expectation step (E-step): Calculate the expected logarithm of the augmented posterior with respect to the conditional distribution of unobserved latent data $(\mathbf{\Gamma}, \boldsymbol{F})$ given $\boldsymbol{Y}$ and $\mathbf{\Omega}^{(m)}$ at the mth iteration:

$$Q\left(\mathbf{\Omega} \mid \mathbf{\Omega}^{(m)}\right) = \mathrm{E}_{\mathbf{\Gamma}, \boldsymbol{F} \mid \boldsymbol{Y}, \mathbf{\Omega}^{(m)}}\left[\log p\left(\mathbf{\Omega}, \mathbf{\Gamma}, \boldsymbol{F}, \boldsymbol{Y}\right)\right] \tag{3.4.2}$$

Maximization step (M-step): Update parameter with $\mathbf{\Omega}^{(m+1)} = \arg\max_{\mathbf{\Omega}} Q\left(\mathbf{\Omega} \mid \mathbf{\Omega}^{(m)}\right)$.

We now simplify the calculation of (3.4.2) by decompose it into separate pieces. For notation convenience, let $\langle X \rangle$ denote the conditional expectation $\mathrm{E}_{\mathbf{\Gamma}, \boldsymbol{F} \mid \boldsymbol{Y}, \mathbf{\Omega}^{(m)}}\left(X\right)$. According the relationship of all parameters and latent variables as in (3.4.1), the parameters are separated into two disconnected parts by $\mathbf{\Gamma}$, $\boldsymbol{F}$ and $\boldsymbol{Y}$. $\left(\tau_{1:N}, \mathbf{\Lambda}_{1:(N+1)}, K\right)$ and $(\boldsymbol{B}, \mathbf{\Sigma})$ are independent given $\mathbf{\Gamma}$, $\boldsymbol{F}$ and $\boldsymbol{Y}$. Thus, $Q\left(\mathbf{\Omega} \mid \mathbf{\Omega}^{(m)}\right)$ can be decomposed as

$$Q\left(\mathbf{\Omega} \mid \mathbf{\Omega}^{(m)}\right) = C^{(m)} + Q_1^{(m)}\left(\tau_{1:N}, \mathbf{\Lambda}_{1:(N+1)}, K\right) + Q_2^{(m)}\left(\boldsymbol{B}, \mathbf{\Sigma}\right) \tag{3.4.3}$$

where $Q_1^{(m)}(\cdot)$ and $Q_2^{(m)}(\cdot)$ are the conditional expectation $\left\langle \log p\left(\tau_{1:N}, \boldsymbol{\Lambda}_{1:(N+1)}, K, \boldsymbol{\Gamma}, \boldsymbol{F}\right)\right\rangle$ and $\langle \log p\left(\boldsymbol{B}, \boldsymbol{\Sigma}, \boldsymbol{\Gamma}, \boldsymbol{F}, \boldsymbol{Y}\right)\rangle$ correspondingly, and $C^{(m)}$ is a constant not involving $\boldsymbol{\Omega}$. We lay out the E-step and M-step implementation in the coming sections.

3.4.1 THE E-STEP

We calculate the three components of $Q\left(\boldsymbol{\Omega} \mid \boldsymbol{\Omega}^{(m)}\right)$ separately. The details are included in Appendix C.1. First, under our Bayesian factor model with multiple change-points, we have

$$
\begin{aligned}
& Q_1^{(m)}\left(\tau_{1:N}, \boldsymbol{\Lambda}_{1:(N+1)}, K\right) \\
= \; & C_1^{(m)} - \frac{1}{2}\sum_{k=1}^{K}\sum_{j=1}^{N+1}\left[\frac{1}{\lambda_{jk}^2}\left(\eta s_\lambda^2 + \sum_{t=\tau_{j-1}+1}^{\tau_j}\left\langle f_{tk}^2\right\rangle\right) + \left(\tau_j - \tau_{j-1} + \eta + 2\right)\log\lambda_{jk}^2\right] \\
& - \frac{1}{2\lambda_0^2}\left(\eta s_\lambda^2 + \sum_{k=K+1}^{K^*}\sum_{t=1}^{T}\left\langle f_{tk}^2\right\rangle\right) - \frac{1}{2}\left(\left(K^* - K\right)T + \eta + 2\right)\log\lambda_0^2 \\
& - \frac{1}{2}NK\log T + \sum_{i=1}^{d}\sum_{k=1}^{K}\left(\langle\gamma_{ik}\rangle\log\theta_1 + \left(1 - \langle\gamma_{ik}\rangle\right)\log(1-\theta_1)\right) \\
& + \sum_{i=1}^{d}\sum_{k=K+1}^{K^*}\left(\langle\gamma_{ik}\rangle\log\theta_0 + \left(1 - \langle\gamma_{ik}\rangle\right)\log(1-\theta_0)\right). \qquad (3.4.4)
\end{aligned}
$$

For the second part, $Q_2^{(m)}\left(\boldsymbol{B}, \boldsymbol{\Sigma}\right)$ has a similar form to the log-likelihood of multivariate regression. Explicitly, we have the following equation,

$$
\begin{aligned}
& Q_2^{(m)}\left(\boldsymbol{B}, \boldsymbol{\Sigma}\right) \\
= \; & C_2^{(m)} - \frac{1}{2}\sum_{t=1}^{T}\left\{\left(\boldsymbol{y}_t - \boldsymbol{B}\left\langle\boldsymbol{f}_t\right\rangle\right)'\boldsymbol{\Sigma}^{-1}\left(\boldsymbol{y}_t - \boldsymbol{B}\left\langle\boldsymbol{f}_t\right\rangle\right) + \operatorname{tr}\left[\boldsymbol{B}'\boldsymbol{\Sigma}^{-1}\boldsymbol{B}\left(\left\langle\boldsymbol{f}_t\boldsymbol{f}_t'\right\rangle - \left\langle\boldsymbol{f}_t\right\rangle\left\langle\boldsymbol{f}_t\right\rangle'\right)\right]\right\} \\
& - \sum_{i=1}^{d}\sum_{k=1}^{K}\left|\beta_{ik}\right|\left(\delta_1\left\langle\gamma_{ik}\right\rangle + \delta_0\left(1 - \left\langle\gamma_{ik}\right\rangle\right)\right) - \frac{T + \xi + 2}{2}\sum_{i=1}^{d}\log\sigma_i^2 - \sum_{i=1}^{d}\frac{\xi s_\sigma^2}{2\sigma_i^2}, \qquad (3.4.5)
\end{aligned}
$$

which involves the conditional expectation of $\mathbf{\Gamma}$, $\boldsymbol{f}_t$ and the quadratic terms of $\boldsymbol{f}_t$.

In each E-step, we update the conditional expectation of the sufficient statistics in $\langle\cdot\rangle$ in (3.4.4) and (3.4.5) given $\mathbf{\Omega}^{(m)}$. Conditional on $\mathbf{\Omega}^{(m)}$, latent factors $\boldsymbol{f}_t$ follows multivariate normal distribution: for $\tau_{j-1} < t \leq \tau_j \, (j = 1, \ldots, N)$,

$$\boldsymbol{f}_t \mid \mathbf{\Omega}^{(m)}, \boldsymbol{Y} \sim \mathcal{N}\left(\boldsymbol{\mu}_t, \boldsymbol{M}_j\right),$$

where $\boldsymbol{\mu}_t = \boldsymbol{M}_j \boldsymbol{B}^{(m)\prime} \mathbf{\Sigma}^{(m)^{-1}} \boldsymbol{y}_t$ and $\boldsymbol{M}_j = \left(\mathbf{\Lambda}_j^{(m)^{-1}} + \boldsymbol{B}^{(m)\prime} \mathbf{\Sigma}^{(m)^{-1}} \boldsymbol{B}^{(m)}\right)^{-1}$. Thus, the conditional mean vector of $\boldsymbol{f}_t$ is $\boldsymbol{\mu}_t$, while the conditional second moment of $\boldsymbol{f}_t$ is obtained from $\langle \boldsymbol{f}_t \boldsymbol{f}_t' \rangle = \boldsymbol{M}_j + \langle \boldsymbol{f}_t \rangle \langle \boldsymbol{f}_t \rangle'$. In addition, conditional on $\mathbf{\Omega}^{(m)}$, each entry of binary matrix $\mathbf{\Gamma}$ follows Bernoulli distribution independently, and thus we have

$$\langle \gamma_{ik} \rangle \equiv P\left(\gamma_{ik} = 1 \mid \mathbf{\Omega}^{(m)}\right) = \frac{\theta_{\mathbb{I}\{k \leq K^{(m)}\}} \phi\left(\beta_{ik}^{(m)} \mid \delta_1\right)}{\theta_{\mathbb{I}\{k \leq K^{(m)}\}} \phi\left(\beta_{ik}^{(m)} \mid \delta_1\right) + \left(1 - \theta_{\mathbb{I}\{k \leq K^{(m)}\}}\right) \phi\left(\beta_{ik}^{(m)} \mid \delta_0\right)}. \tag{3.4.6}$$

3.4.2 The M-step

Once the latent sufficient statistics have been updated, the M-step consists of maximizing (3.4.3) with respect to the unknown parameters. The M-step for regular EM algorithm would seek $\left(\tau_{1:N}^{(m+1)}, \mathbf{\Lambda}_{1:(N+1)}^{(m+1)}, K^{(m+1)}\right)$ and $\left(\boldsymbol{B}^{(m+1)}, \mathbf{\Sigma}^{(m+1)}\right)$ which optimize $Q_1^{(m)}$ and $Q_2^{(m)}$ correspondingly. However, due to the complicated form in (3.4.4) and (3.4.5), we proceed with the following Conditional Maximization (CM) steps.

Conditional on $K^{(m)}$, the maximization of $Q_1^{(m)}$ can be transformed to a change-point detection problem such that $\tau_{1:N}$ minimizes $\sum_{j=1}^{N+1} \mathcal{C}\left(\tau_{j-1} + 1, \tau_j\right) + cN$. The cost func-

tion $\mathcal{C}(\cdot)$ is the negative maximum expected log-likelihood, i.e.

$$\mathcal{C}\left(\tau_{j-1}+1, \tau_j\right) = -\max_{\boldsymbol{\Lambda}_j}\left[\log p\left(\boldsymbol{\Lambda}_j \mid K^{(m)}\right) + \sum_{t=\tau_{j-1}+1}^{\tau_j}\left\langle \log p\left(\boldsymbol{f}_t \mid \boldsymbol{\Lambda}_j, K^{(m)}\right)\right\rangle\right]$$

where $\langle X \rangle$ denotes the conditional expectation $\mathrm{E}_{\boldsymbol{\Gamma},\boldsymbol{F}|\boldsymbol{Y},\boldsymbol{\Omega}^{(m)}}(X)$, and we have linear penalty $c = \frac{1}{2}K\log T$ according to (3.4.4). Killick et al. (2012) proposed an efficient algorithm PELT method to solve the above change-point detection problem with linear computational cost. We applied the PELT method here to find

$$\left(\tau_{1:N}^{(m+1)}, \boldsymbol{\Lambda}_{1:(N+1)}^{(m+1/2)}\right) = \arg\max_{(\tau,\boldsymbol{\Lambda})} Q_1^{(m)}\left(\tau_{1:N}, \boldsymbol{\Lambda}_{1:(N+1)}, K^{(m)}\right)$$

with details deferred to Appendix C.2.1. Then, conditional on $\tau_{1:N}^{(m+1)}$, we update

$$\left(\boldsymbol{\Lambda}_{1:(N+1)}^{(m+1)}, K^{(m+1)}\right) = \arg\max_{(\boldsymbol{\Lambda},K)} Q_1^{(m)}\left(\tau_{1:N}^{(m+1)}, \boldsymbol{\Lambda}_{1:(N+1)}, K\right).$$

Similarly, we exploit the following conditional maximization steps to find $\left(\boldsymbol{B}^{(m+1)}, \boldsymbol{\Sigma}^{(m+1)}\right)$ such that

$$\boldsymbol{B}^{(m+1)} = \arg\max_{\boldsymbol{B}} Q_2^{(m)}\left(\boldsymbol{B}, \boldsymbol{\Sigma}^{(m)}\right), \boldsymbol{\Sigma}^{(m+1)} = \arg\max_{\boldsymbol{\Sigma}} Q_2^{(m)}\left(\boldsymbol{B}^{(m+1)}, \boldsymbol{\Sigma}\right)$$

with details deferred to Appendix C.2.2. With the above conditional updates in the M-step, monotone convergence of EM algorithm is still guaranteed (Meng & Rubin, 1993).

3.4.3 Parameter Expansion

A common problem in estimating factor model is that the strong ties between factor loading and latent factor can cause slowdown the convergence of an EM algorithm. Our factor model with change-points is vulnerable in the same way. In practice, some sub-optimal solutions might be obtained due to the rotational invariance of the likelihood of a factor model although our prior should guide the estimation towards sparse loadings. Liu et al. (1998) proposed parameter expanded EM algorithm (PX-EM) to accelerate the convergence by embedding the complete data model within a larger model with extra parameters. Ročková & George (2016a) introduced a variant of a PX-EM algorithm, namely PXL-EM, to rotate the factor loading toward orientations which best match the prior assumptions of independent latent components and sparse loadings. A key to PXL-EM approach is to employ the parameter expansion only on the likelihood portion of the posterior, while using the SSL prior to guide the algorithm toward sparse factor orientations. Due to the fact that PXL-EM changes the target function of the original optimization, we only apply PXL-EM to help us find better initial points.

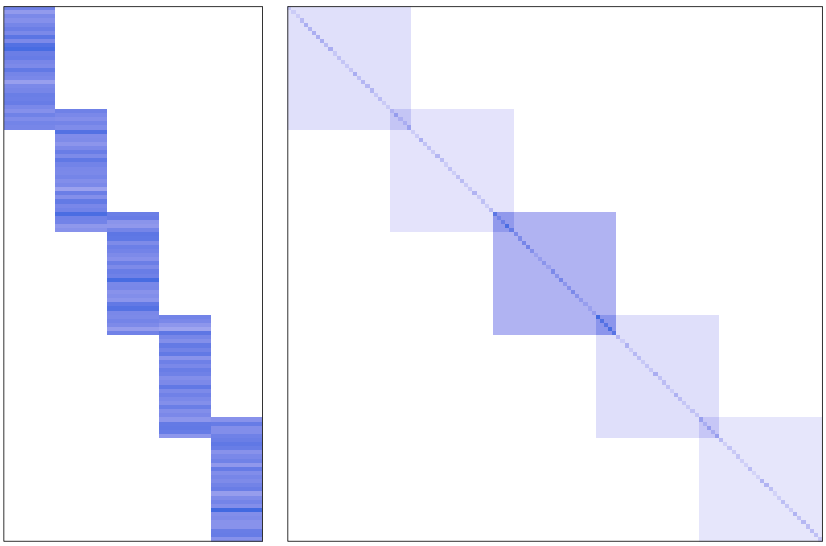
3.5 Simulation studies

In this section, we simulate datasets from factor model (3.3.1) with sparse factor loading, and study the effectiveness of our proposed EM approach in detecting change-points and recovering true factor loadings.

3.5.1 Change-point detection

We generate a dataset with $T = 300$ observations and three change-points $\tau_{1:3} = \{80, 160, 200\}$, and thus the observations can be divided into four segments. We let

the number of variables $d = 130$ and the true number of factors $K = 5$. The true factor loading $\boldsymbol{B}$ is a matrix which has a block-diagonal pattern as shown in Figure 3.1, as the factor loadings are either 0, or follow distribution $\mathcal{N}\left(1, 0.1^2\right)$. In jth segment, the covariance of true factors is determined by diagonal matrix $\boldsymbol{\Lambda}_j = \text{diag}\left\{\lambda_{j1}^2, \ldots, \lambda_{jK}^2\right\}$ We simulate $\lambda_{j1}^2, \ldots, \lambda_{jK}^2 \overset{iid}{\sim} \text{Lognormal}\left(0, \log(v)/2\right)$, and the parameters of the distribution are chosen so that 95% of the simulated variances are within the range $(1/v, v)$. The covariance matrix of idiosyncratic errors is $\boldsymbol{\Sigma} = \text{diag}\left\{\sigma_1^2, \ldots, \sigma_d^2\right\}$, and it doesn't change through time. We simulate $\sigma_1^2, \ldots, \sigma_d^2 \overset{iid}{\sim} s^2 \cdot \text{Uniform}\left(0.5, 1.5\right)$. Based on (3.3.2), the distribution of $\boldsymbol{y}_t$ is $\mathcal{N}(\boldsymbol{0}, \boldsymbol{B}\boldsymbol{\Lambda}_j\boldsymbol{B}' + \boldsymbol{\Sigma})$ for $\tau_{j-1} < t \leq \tau_j$. We vary $v \in \{2, 5, 10\}$ and $s^2 \in \{1/4, 1/2, 1, 2\}$ in this simulation.



(a) Factor loadings: $\boldsymbol{B}$ **(b)** Covariance in a segment: $\boldsymbol{B}\boldsymbol{\Lambda}_1\boldsymbol{B}' + \boldsymbol{\Sigma}$

Figure 3.1: A realization of factor model simulation with multiple change-points

We exploit our proposed EM algorithm to the simulated dataset. The input for our method is standardized such each variable has mean 0 and unit variance. We let the

hyper-parameters $K^* = 10$, $s_\lambda^2 = s_\sigma^2 = 1$ and $\delta_0 = 20$. The sensitivity to δ_0 is discussed in next subsection. In the computation, we follow the guidance in Ročková & George (2016a) by keeping the slab variance steady and gradually increasing the spike variance δ_0 over a ladder of values $\{1, 5, 10, 20\}$. In addition, since the scale of $\boldsymbol{B}$ and $\boldsymbol{\Lambda}_j$s is unidentifiable in likelihood, we output rescaled $\widehat{\boldsymbol{\Lambda}}_j$s and $\widehat{\boldsymbol{B}}$ such that each factor have unit variance on average throughout the whole periods for better interpratablility, i.e. $\frac{1}{T}\sum_{j=1}^{\hat{N}+1}\left(\hat{\tau}_j - \hat{\tau}_{j-1}\right)\hat{\lambda}_{jk}^2 = 1$.

Figure 3.2 shows the change-point detection results. The change-point detection gets more accurate as the variation between segments increases from top to bottom, while the results are robust to the variance of idiosyncratic error.

3.5.2 Factor loading recovery

In the above simulation, the EM algorithm correctly estimate the number of factors for more than 99% simulations. Following the simulation described in 3.5.1, we now fix $\boldsymbol{\Lambda}_1 = 4\boldsymbol{I}$, $\boldsymbol{\Lambda}_2 = \boldsymbol{I}$, $\boldsymbol{\Lambda}_3 = 3\boldsymbol{I}$ and $\boldsymbol{\Lambda}_4 = \boldsymbol{I}$ to take a closer look at the factor loading recovery and the sensitivity to hyper-parameter δ_0. The recovery of sparse factor loading is very important to the explainability of factor analysis. Figure 3.3 shows the estimated factor loadings under different hyper-parameter δ_0. We also compare our estimated factor loading to the results given by PCA and Sparse PCA (Zou et al., 2006) in Figure 3.4.

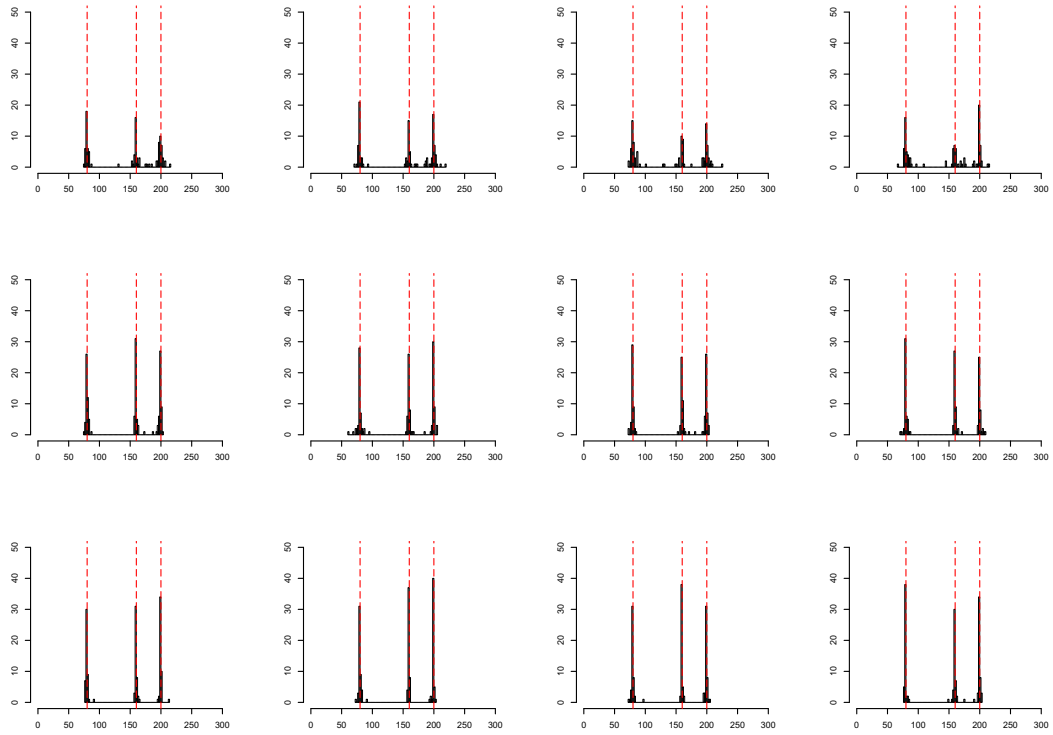
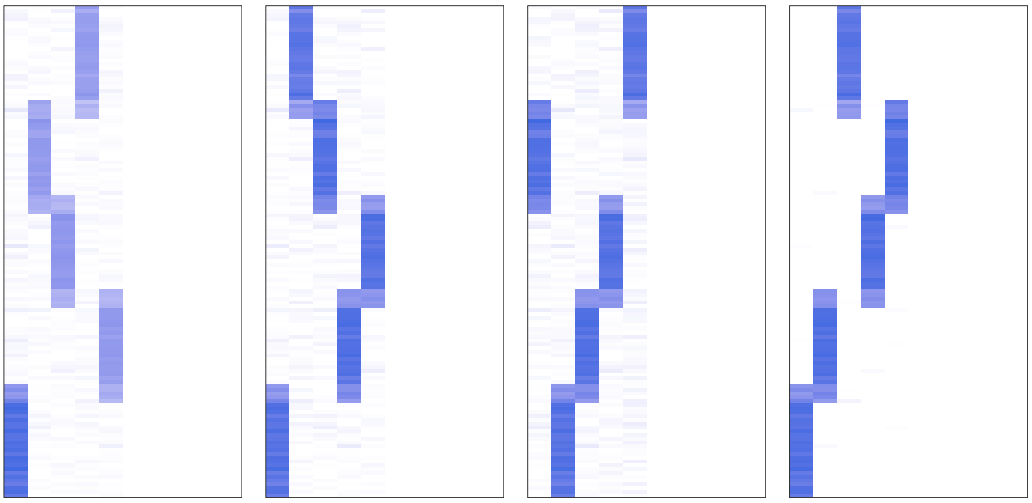


Figure 3.2: Change-point detection over 12 different settings of (ν, s^2). From left to right, $s^2 = 1/4, 1/2, 1, 2$. From top to bottom, $\nu = 2, 5, 10$.

3.6 REAL DATA EXAMPLES

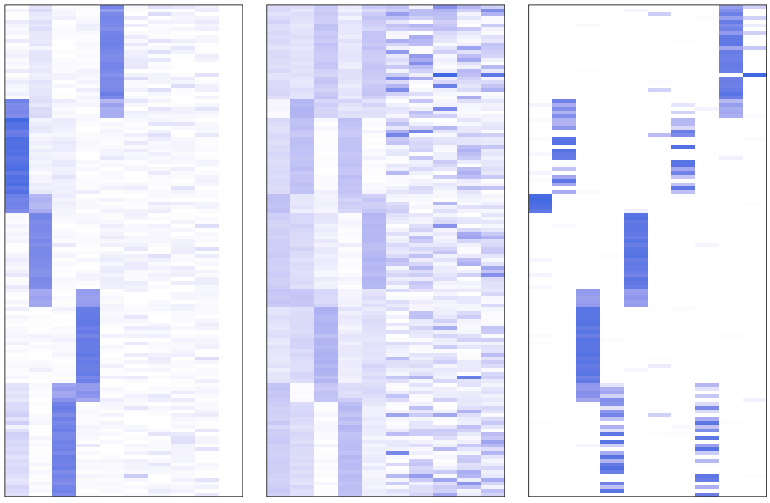
3.6.1 S&P 100 STOCK RETURN

This example studies the S&P 100 stock daily return in time period 2007-2016. Since the components of S&P 100 change over time, we use the collection all S&P 100 lists from 2009 to 2016 as our estimation universe, and eliminate the names which was not listed in the stock market until after 2007 or is no longer listed by the end of 2016. The number of stocks is $d = 88$ in this study. We denote the time series of daily return as $\{\boldsymbol{r}_t\}_{1:T}$ where $T = 2518$. Since the daily returns have rapid volatility change, the original data



(a) $\delta_0 = 5$ **(b)** $\delta_0 = 10$ **(c)** $\delta_0 = 20$ **(d)** $\delta_0 = 50$

Figure 3.3: Estimated factor loadings for different values of δ_0.



(a) $\delta_0 = 1$ **(b)** PCA **(c)** SPCA

Figure 3.4: Estimated factor loadings for $\delta_0 = 1$, PCA and SPCA.

should not be modeled by our factor model with change-points. Therefore, we first

Table 3.1: For each factor, we list the stocks that have factor loadings greater than 0.2 in absolute value.

factor 2	AEP, CL, COST, CPB, ETR, EXC, FCX, JNJ, KO, MCD, MO, PEP, PG, SO, T, VZ, WMT
factor 3	AA, BHI, COP, CVX, DVN, FCX, HAL, NOV, OXY, SLB, WMB, XOM
factor 4	ALL, AXP, BAC, BK, C, COF, GS, JPM, MET, MS, RF, USB, WFC
factor 5	ABT, AMGN, BAX, BMY, GILD, JNJ, MDT, MRK, PFE, UNH
factor 6	AAPL, AMZN, CSCO, GOOG, HPQ, IBM, INTC, MSFT, ORCL, TXN
factor 7	AEP, ETR, EXC, SO
factor 8	COST, HD, LOW, TGT
factor 9	BA, GD, LMT, RTN, UTX
factor 10	BHI, HAL, NOV, SLB

preprocess the return of individual stocks using GARCH(1,1):

$$\begin{aligned} r_{i,t} &= \sigma_{i,t} y_{i,t} \\ \sigma_{i,t}^2 &= \alpha_0 + \alpha_1 r_{i,t-1}^2 + \beta_1 \sigma_{i,t-1}^2. \end{aligned}$$

where $i = 1, \ldots, d$ and $t = 1, \ldots, T$. After GARCH estimation, we generate stabilized daily return $y_{i,t} = r_{i,t}/\hat{\sigma}_{i,t}$. We denote $\boldsymbol{y}_t = (y_{1,t}, \ldots, y_{d,t})'$ and exploit our method on the time series $\{\boldsymbol{y}_t\}_{1:T}$. For the hyper-parameter settings, we let $s_\lambda^2 = s_\sigma^2 = 1$, $\delta_1 = 0.001$, $\delta_0 = 50$ and $K^* = 20$ in this study. To relieve the burden of computation, we assume that the change-points can exist only at $t \in \{5, 10, 15, \ldots, 2515\}$.

The estimated number of factors $\hat{K} = 10$, and Figure 3.5 shows the estimated factor loadings $\widehat{\boldsymbol{B}}'$. The first row of $\widehat{\boldsymbol{B}}'$ is nonzero for all stocks, and it represents the main index component of stock market. To get a better understanding of this factor model estimation, we highlight the major stock tickers in the other factors. Table 3.1 list the stocks that have factor loadings greater than 0.2 in absolute value. Based on the Table 3.1, the main component stocks in a same factor are closely related fundamentally, and thus we can easily interpret the factors. For example, factor 3-6 can represent energy, financial, healthcare and technology sector correspondingly.

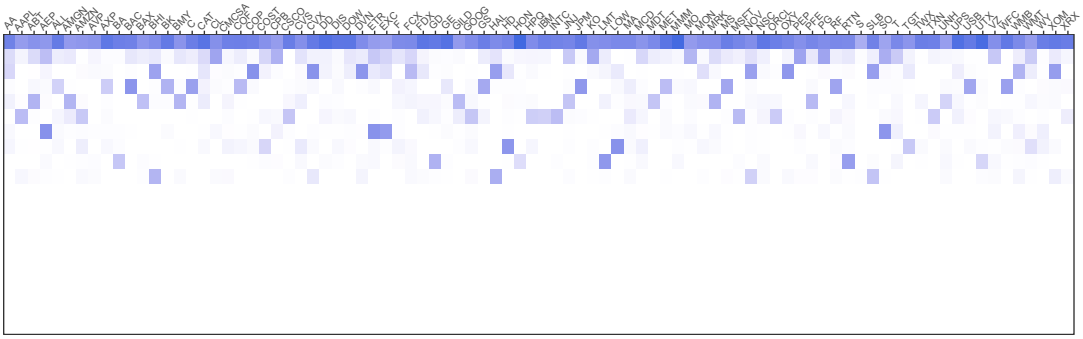


Figure 3.5: Estimated factor loadings (transposed): $\widehat{B}'$. For each entry, darker color implies larger absolute value of factor loading.

The detected change-points are at: 2007/12/28, 2008/09/23, 2010/05/04, 2010/10/08, 2011/07/26, 2011/12/30, 2014/10/07, 2016/02/22. Many of these change-points are in a neighborhood of critical events that influence the financial market. For example, 2008/09/23 is right after the bankruptcy of Lehman Brothers during the 2008 financial crisis; 2011/07/26 is around critical moments during the Greek government-debt crisis.

3.6.2 fMRI ACTIVITY

This example concerns the multivariate time series of fMRI activity in human brain. The data is collected by the Center for Cognitive Brain Imaging at Carnegie Mellon University as part of the star/plus experiment for six individuals (Keller et al., 2001). Each experiment consists of a sequence of 40 trials and resting periods in between. The typical timing of each such trial can be summarized as: each subject was shown a picture (or sentence) for four seconds, a blank screen for four seconds, a related sentence (or picture) for four seconds and a resting period for 15 seconds. When shown a sentence, the subjects are instructed to press a button to indicate whether the sentence

correctly described the picture. For the first half of the trials, the picture was presented, and the sentence was presented first on the other half of the trials.

During the experiment of about 1400 seconds, the subjects were positioned inside an MRI scanner, and magnetic resonance images are collected every 0.5 second. Each image was partitioned into 4698 voxels of width 3 mm, making the raw data a high-dimensional multivariate time series. Due to the high-dimensionality of the original data, existing work usually group the voxels into distinct regions of interest (ROIs) in the brain, and then average the signals over all voxels within the same ROI. Barnett & Onnela (2016) used this dataset to illustrate a method detecting change-point in correlation networks. They analyze this dataset by combining the analysis on the eight trials where the picture is presented first and the sentence agrees with the picture for all six individuals.

Here, we apply Bayesian factor model with multiple change-points to the whole experiment for each individual, assuming that structural change may occur only between trials, or equivalently 51 possible positions. We first apply our method to the multivariate time series of 25 ROIs with length 2800. Let $\delta_1 = 0.001$, $\delta_0 = 50$, $K^* = 5$ as hyper-parameter setting, then the estimated change-point set is $\{1375\}$, which is at the middle of the experiment when the showing order of pictures and sentences was switched. Based on our factor model estimation, Figure 3.6 displays the estimated correlation network of Brain region of interest (ROI) before and after the detected change-point, where the edge indicates that the absolute value of correlation is greater than 0.5.

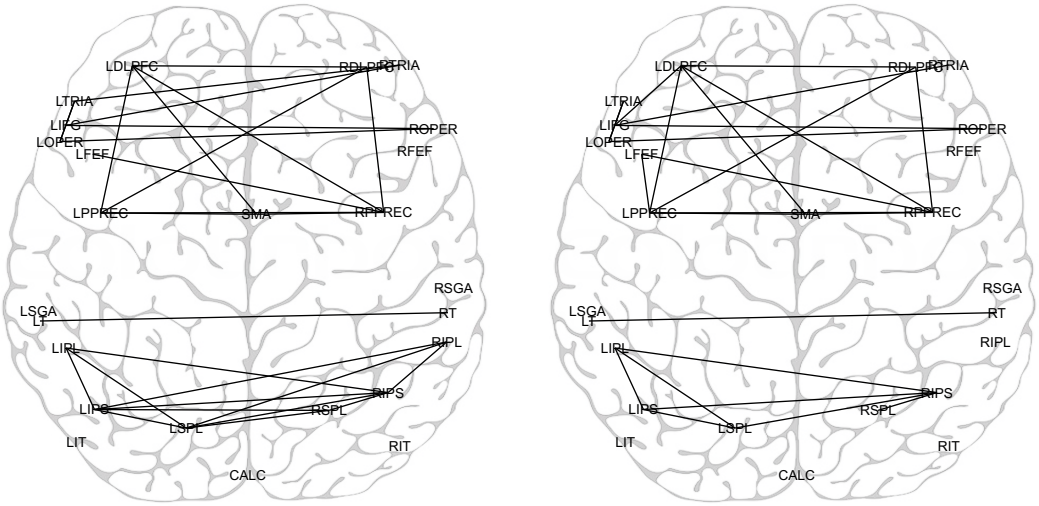


Figure 3.6: Correlation network of Brain region of interest (ROI) before and after the detected change-point. We construct an edge between two ROIs if the absolute value of correlation between them is greater than 0.5.

3.7 Discussion

In this article, we introduced a Bayesian framework on factor model with multiple change-points in the quest for solving time-varying covariance estimation of high-dimensional time series. Different from PCA-type methods, our study focuses on both the interpretability of latent factors and the time-varying nature of observed data. Under the high-dimensional setting, we exploit spike-and-slab LASSO prior on factor loadings such that the estimated factor loading matrix is sparse and interpretable. On top of factor model, we consider the existence of multiple change-points to accommodate the change over time. Furthermore, the number of factors and the number of change-points are both consider unknown in our Bayesian analysis. We then proposed an efficient EM algorithm to estimate the Bayesian factor model with multiple change-points. Our EM algorithm takes advantage of existing implementations for

sparse factor model and efficient algorithms for multiple change-point detection. In the application to real data examples, our method delivers highly interpretable latent factor and meaningful change-points.

In dealing with time-varying covariance of high-dimensional times series data, all statistical inference would rely on a few assumptions. With change-point model being arguably the simplest model to accommodate time-varying nature, the Bayesian change-point model for covariance matrix faces limitations in handling high-dimensional data in the sense that change-points are hardly detectable. To better exploit the latent structure under high-dimensional setting, our analysis relies on the factor model and assumes that the change is driven by the main factors but not the idiosyncratic errors. It would be interesting to further relax our assumptions under Bayesian frameworks.

A

Appendix to Chapter 1

A.1 Proof for the consistency of MLE

We need the following two lemmas to prove Theorem 1.

Lemma 1 *Given* $\boldsymbol{\mu} \in \boldsymbol{\Theta}$*, for every* $\epsilon > 0$*, there exists* R *such that for all* $\boldsymbol{v} \in \boldsymbol{\Theta}$ *with* $\|\boldsymbol{\mu} - \boldsymbol{v}\| < 1$*,*

$$\int_{\mathcal{B}(R)^c} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{v}\|^2} d\boldsymbol{Z} < \epsilon$$

where $\mathcal{B}(R) = \{\boldsymbol{Z} \in \mathbb{R}^n : \|\boldsymbol{Z}\| \leq R\}$.

Proof: According to the triangle inequality and the fact that $\|\boldsymbol{v} - \boldsymbol{\mu}\| < 1$, we have

$$\|\boldsymbol{Z} - \boldsymbol{v}\| \geq \|\boldsymbol{Z}\| - \|\boldsymbol{u}\| - 1 = \|\boldsymbol{Z}\|/2 + (\|\boldsymbol{Z}\|/2 - 1 - \|\boldsymbol{u}\|).$$

Therefore, when $R \geq 2(1 + \|\boldsymbol{\mu}\|)$, $\|\boldsymbol{Z} - \boldsymbol{v}\| \geq \|\boldsymbol{Z}\|/2$ for any $\boldsymbol{Z} \in \mathcal{B}(R)^c$, and thus

$$\int_{\mathcal{B}(R)^c} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{v}\|^2} \mathrm{d}\boldsymbol{Z} \leq \int_{\mathcal{B}(R)^c} e^{-\frac{1}{8}\|\boldsymbol{Z}\|^2} \mathrm{d}\boldsymbol{Z}. \tag{A.1.1}$$

Note that the right hand side of inequality (A.1.1) converges to zero as R goes to infinity. For any $\epsilon > 0$, there exists R such that $R \geq 2(1 + \|\boldsymbol{\mu}\|)$ and $\int_{\mathcal{B}(R)^c} e^{-\frac{1}{8}\|\boldsymbol{Z}\|^2} \mathrm{d}\boldsymbol{Z} < \epsilon$, which further implies $\int_{\mathcal{B}(R)^c} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{v}\|^2} \mathrm{d}\boldsymbol{Z} < \epsilon$ for any $\|\boldsymbol{v} - \boldsymbol{\mu}\| < 1$. □

Lemma 2 *(Identifiability) The true distribution of* τ *is identifiable on parameter space* $\boldsymbol{\Theta}$*:* $P_{\boldsymbol{\mu}} \neq P_{\boldsymbol{\mu}'}$ *for every* $\boldsymbol{\mu} \neq \boldsymbol{\mu}'$.

Proof: Suppose $\boldsymbol{\mu}, \boldsymbol{\mu}' \in \boldsymbol{\Theta}$ and $P_{\boldsymbol{\mu}}(\tau) = P_{\boldsymbol{\mu}'}(\tau)$ for all possible full ranking list τ on U. Since

$$P_{\boldsymbol{\mu}}(\tau(i_1) < \tau(i_2)) = \sum_{\tau(i_1)<\tau(i_2)} P_{\boldsymbol{\mu}}(\tau),$$

we have $P_{\boldsymbol{\mu}}(\tau(i_1) < \tau(i_2)) = P_{\boldsymbol{\mu}'}(\tau(i_1) < \tau(i_2))$ for any $i_1, i_2 \in U$. On the other hand, $P_{\boldsymbol{\mu}}(\tau(i_1) < \tau(i_2)) = \Phi(\frac{\mu_{i_1} - \mu_{i_2}}{\sqrt{2}})$, where $\Phi(\cdot)$ is the Normal CDF. Hence, $\mu_{i_1} - \mu_{i_2} = \mu'_{i_1} - \mu'_{i_2}$ for $1 \leq i_1 < i_2 \leq n$, and $\mathbf{1}'\boldsymbol{\mu} = \mathbf{1}'\boldsymbol{\mu}' = 0$. Hence, we have $\boldsymbol{\mu} = \boldsymbol{\mu}'$. □

Proof of Theorem 1: We apply Wald's consistency proof (Van der Vaart, 2000) by verifying the following three conditions.

First, we denote $m_{\boldsymbol{\mu}}(\tau) \overset{def}{=} \log P_{\boldsymbol{\mu}}(\tau)$ and show that the map $\boldsymbol{\mu} \mapsto m_{\boldsymbol{\mu}}(\tau)$ is continuous for all τ. According to Lemma 1, given $\boldsymbol{\mu} \in \boldsymbol{\Theta}$, for every $\epsilon > 0$, there exists R such

that for all $\boldsymbol{v} \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\mu} - \boldsymbol{v}\| < 1$, $\int_{\mathcal{B}(R)^c} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{v}\|^2} d\boldsymbol{Z} < \epsilon/4$. Thus, for any $\mathcal{A} \subset \mathbb{R}^n$ and $\|\boldsymbol{v} - \boldsymbol{\mu}\| < 1$,

$$\int_{\mathcal{A}\cap\mathcal{B}(R)^c} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{\mu}\|^2} \mathrm{d}\boldsymbol{Z} + \int_{\mathcal{A}\cap\mathcal{B}(R)^c} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{v}\|^2} \mathrm{d}\boldsymbol{Z} < \epsilon/2.$$

Since $e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{\mu}\|^2}$ is bound by 1 and $\mathcal{A} \cap \mathcal{B}(R)$ is a bounded area, by the Bounded Convergence Theorem, we have for any sequence $\{\boldsymbol{\mu}_k\}_{k=1}^{\infty}$ converging to $\boldsymbol{\mu}$,

$$\lim_{k\to\infty} \int_{\mathcal{A}\cap\mathcal{B}(R)} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{\mu}_k\|^2} \mathrm{d}\boldsymbol{Z} = \int_{\mathcal{A}\cap\mathcal{B}(R)} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{\mu}\|^2} \mathrm{d}\boldsymbol{Z}.$$

Hence, $\int_{\mathcal{A}\cap\mathcal{B}(R)} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{\mu}\|^2} \mathrm{d}\boldsymbol{Z}$ is a continuous function of $\boldsymbol{\mu}$. Thus, for every $\epsilon > 0$, there exist δ such that for all $\boldsymbol{v} \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\mu} - \boldsymbol{v}\| < \delta$,

$$\left| \int_{\mathcal{A}\cap\mathcal{B}(R)} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{\mu}\|^2} \mathrm{d}\boldsymbol{Z} - \int_{\mathcal{A}\cap\mathcal{B}(R)} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{v}\|^2} \mathrm{d}\boldsymbol{Z} \right| \leq \epsilon/2$$

Therefore, given $\boldsymbol{\mu} \in \boldsymbol{\Theta}$ and $\mathcal{A} \subset \mathbb{R}^n$, for every $\epsilon > 0$, there exists R such that for all $\boldsymbol{v} \in \boldsymbol{\Theta}$ with $\|\boldsymbol{\mu} - \boldsymbol{v}\| < \min\{1, \delta\}$,

$$\begin{aligned}
& \left| \int_{\mathcal{A}} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{\mu}\|^2} \mathrm{d}\boldsymbol{Z} - \int_{\mathcal{A}} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{v}\|^2} \mathrm{d}\boldsymbol{Z} \right| \\
\leq & \left| \int_{\mathcal{A}\cap\mathcal{B}(R)^c} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{\mu}\|^2} \mathrm{d}\boldsymbol{Z} - \int_{\mathcal{A}\cap\mathcal{B}(R)^c} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{v}\|^2} \mathrm{d}\boldsymbol{Z} \right| \\
& + \left| \int_{\mathcal{A}\cap\mathcal{B}(R)} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{\mu}\|^2} \mathrm{d}\boldsymbol{Z} - \int_{\mathcal{A}\cap\mathcal{B}(R)} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{v}\|^2} \mathrm{d}\boldsymbol{Z} \right| \\
\leq & \int_{\mathcal{A}\cap\mathcal{B}(R)^c} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{\mu}\|^2} \mathrm{d}\boldsymbol{Z} + \int_{\mathcal{A}\cap\mathcal{B}(R)^c} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{v}\|^2} \mathrm{d}\boldsymbol{Z} \\
& + \left| \int_{\mathcal{A}\cap\mathcal{B}(R)} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{\mu}\|^2} \mathrm{d}\boldsymbol{Z} - \int_{\mathcal{A}\cap\mathcal{B}(R)} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{v}\|^2} \mathrm{d}\boldsymbol{Z} \right| < \epsilon.
\end{aligned}$$

We conclude that $\int_{\tau(\boldsymbol{Z})=\tau'} e^{-\frac{1}{2}\|\boldsymbol{Z}-\boldsymbol{\mu}\|^2} \mathrm{d}\boldsymbol{Z}$ is continuous with respect to $\boldsymbol{\mu}$.

Second, for every sufficiently small $V \subset \boldsymbol{\Theta}$ the function $\tau \mapsto \sup_{\boldsymbol{\mu} \in U} m_{\boldsymbol{\mu}}(\tau)$ is measurable and satisfies

$$E_{\boldsymbol{\mu}_0} \sup_{\boldsymbol{\mu} \in V} m_{\boldsymbol{\mu}}(\tau) < \infty.$$

The domain of τ a finite set containing all possible ranking lists of n entities. Thus, for any $t \in R$, the preimage of (t, ∞) under $\tau \mapsto \sup_{\boldsymbol{\mu} \in U} m_{\boldsymbol{\mu}}(\tau)$ is a finite set. Thus, $\tau \mapsto \sup_{\boldsymbol{\mu} \in U} m_{\boldsymbol{\mu}}(\tau)$ is a measureable function.

Third, because $P_{\boldsymbol{\mu}}(\tau) \leq 1$, we have $\sup_{\boldsymbol{\mu} \in V} m_{\boldsymbol{\mu}}(\tau) \leq 0$. Since the domain of τ is a finite set, there exists a lower bound c such that $\sup_{\boldsymbol{\mu} \in V} m_{\boldsymbol{\mu}}(\tau) \geq c$ for every τ. Thus, $E_{\boldsymbol{\mu}_0} \sup_{\boldsymbol{\mu} \in V} m_{\boldsymbol{\mu}}(\tau)$ exists and $E_{\boldsymbol{\mu}_0} \sup_{\boldsymbol{\mu} \in V} m_{\boldsymbol{\mu}}(\tau) < \infty$.

Due to the identifiability of $P_{\boldsymbol{\mu}}$ in Lemma 2, $E_{\boldsymbol{\mu}_0} m_{\boldsymbol{\mu}}(\tau)$ attains its maximum uniquely at $\boldsymbol{\mu}_0$. Then according to Wald's consistency proof, for every $\epsilon > 0$ and every compact set $K \subset \boldsymbol{\Theta}$,

$$P\left(\{\|\hat{\boldsymbol{\mu}}_m - \boldsymbol{\mu}_0\| \geq \epsilon\} \cap \{\hat{\boldsymbol{\mu}}_m \in K\}\right) \to 0, \text{ as } n \text{ is fixed}, m \to \infty.$$

A.2 Validity of the parameter-expanded Gibbs sampler

Below we show the validity of parameter expanded Gibbs sampler under BARC, and the validity under BARCW and BARCM follows by the same logic. We use π to denote the marginal posterior distribution of $\boldsymbol{Z}$ given all the observed ranking lists $\mathcal{T}$, i.e.,

$$\pi(\boldsymbol{Z}) = p(\boldsymbol{Z} \mid \mathcal{T}) \propto p(\boldsymbol{Z}) p(\mathcal{T} \mid \boldsymbol{Z}) = p(\boldsymbol{Z}) 1\{\tau(\boldsymbol{Z}) = \mathcal{T}\}.$$

In order to show the validity of parameter expansion, it suffices to prove that for any $\boldsymbol{Z}$ following the marginal posterior distribution $\pi(\boldsymbol{Z})$, its transformation $t_\theta(\boldsymbol{Z})$ also

follows the same distribution π, as long as θ is draw from the distribution with density proportional to $\pi(t_\theta(\boldsymbol{Z}))|J_\theta(\boldsymbol{Z})|\theta^{-1}$. The proof is as follows.

By construction, the joint density of $(\boldsymbol{Z}, \theta)$ is

$$p(\boldsymbol{Z}, \theta) = p(\boldsymbol{Z})p(\theta \mid \boldsymbol{Z}) = \pi(\boldsymbol{Z}) \cdot \frac{\pi(t_\theta(\boldsymbol{Z}))\theta^{-nm-1}}{\int_{\mathbb{R}} \pi(t_\gamma(\boldsymbol{Z}))\gamma^{-nm-1}\mathrm{d}\gamma},$$

which immediately implies the joint density of $(\boldsymbol{Y}, \theta) \equiv (t_\theta(\boldsymbol{Z}), \theta)$:

$$\begin{aligned} p(\boldsymbol{Y}, \theta) = p(\boldsymbol{Z}, \theta)|J_\theta(\boldsymbol{Z})|^{-1} &= \pi(\boldsymbol{Z}) \cdot \frac{\pi(t_\theta(\boldsymbol{Z}))\theta^{-1}}{\int_{\mathbb{R}} \pi(t_\gamma(\boldsymbol{Z}))\gamma^{-nm-1}\mathrm{d}\gamma} \\ &= \pi(t_\theta^{-1}(\boldsymbol{Y})) \cdot \frac{\pi(\boldsymbol{Y})\theta^{-1}}{\int_{\mathbb{R}} \pi(t_\gamma(t_\theta^{-1}(\boldsymbol{Y})))\gamma^{-nm-1}\mathrm{d}\gamma}. \end{aligned} \tag{A.2.1}$$

Note that $t_\gamma(t_\theta^{-1}(\boldsymbol{Y})) = \theta\boldsymbol{Y}/\gamma = t_\kappa^{-1}(\boldsymbol{Y})$, where $\kappa = \theta/\gamma$. We can then simplify the denominator in (A.2.1) as

$$\begin{aligned} \int_{\mathbb{R}} \pi(t_\gamma(t_\theta^{-1}(\boldsymbol{Y})))\gamma^{-nm-1}\mathrm{d}\gamma &= \int_{\mathbb{R}} \pi(t_\kappa^{-1}(\boldsymbol{Y}))(\theta/\kappa)^{-nm-1}\mathrm{d}(\theta/\kappa) \\ &= \theta^{-nm} \int_{\mathbb{R}} \pi(t_\kappa^{-1}(\boldsymbol{Y})) \cdot \kappa^{nm-1}\mathrm{d}\kappa, \end{aligned}$$

and thus further simplify $p(\boldsymbol{Y}, \theta)$ as

$$p(\boldsymbol{Y}, \theta) = \pi(\boldsymbol{Y}) \cdot \frac{\pi(t_\theta^{-1}(\boldsymbol{Y}))\theta^{-1}}{\theta^{-nm}\int_{\mathbb{R}} \pi(t_\kappa^{-1}(\boldsymbol{Y})) \cdot \kappa^{nm-1}\mathrm{d}\kappa} = \pi(\boldsymbol{Y}) \cdot \frac{\pi(t_\theta^{-1}(\boldsymbol{Y}))\theta^{nm-1}}{\int_{\mathbb{R}} \pi(t_\kappa^{-1}(\boldsymbol{Y})) \cdot \kappa^{nm-1}\mathrm{d}\kappa}.$$

Therefore, the marginal density of $\boldsymbol{Y}$ is

$$p(\boldsymbol{Y}) = \pi(\boldsymbol{Y}) \cdot \frac{\int_{\mathbb{R}} \pi(t_\theta^{-1}(\boldsymbol{Y}))\theta^{nm-1}\mathrm{d}\theta}{\int_{\mathbb{R}} \pi(t_\kappa^{-1}(\boldsymbol{Y})) \cdot \kappa^{nm-1}\mathrm{d}\kappa} = \pi(\boldsymbol{Y}),$$

i.e., $\boldsymbol{Y} \equiv t_\theta(\boldsymbol{Z})$ follows the distribution with density π.

A.3 GIBBS SAMPLER FOR BARCW

The Gibbs sampling with parameter expansion for BARCW model is accomplished by iterating the following steps.

1. For $i = 1, \ldots, n$, $j = 1, \ldots, m$: Draw $[Z_{ij} \mid Z_{[-i],j}, \boldsymbol{Z}_{[-j]}, \boldsymbol{\alpha}, \boldsymbol{\beta}]$ from $\mathcal{N}(\alpha_i + \boldsymbol{x}_i'\boldsymbol{\beta}, 1)$ with truncation points determined by $Z_{[-i],j}$, such that Z_{ij} falls in the correct position according to τ_j^P.

2. Draw $\theta \sim S^{1/2}/\chi_{nm}$ where

$$S = \sum_{j=1}^{m} w_j \boldsymbol{Z}_j'\boldsymbol{Z}_j - \sum_{j,k}^{m} w_j w_k \boldsymbol{Z}_j'\boldsymbol{V}\left(\boldsymbol{\Lambda}^{-1} + \sum_{j=1}^{m} w_j \boldsymbol{V}'\boldsymbol{V}\right)^{-1} \boldsymbol{V}'\boldsymbol{Z}_k.$$

3. Draw $(\boldsymbol{\alpha}', \boldsymbol{\beta}')' \sim p(\boldsymbol{\alpha}, \boldsymbol{\beta} \mid t_\theta(\boldsymbol{Z})) \equiv \mathcal{N}\left(\hat{\boldsymbol{\eta}}/\theta, \hat{\boldsymbol{\Sigma}}\right)$, where

$$\hat{\boldsymbol{\eta}} = \left(\boldsymbol{\Lambda}^{-1} + \sum_{j=1}^{m} w_j \boldsymbol{V}'\boldsymbol{V}\right)^{-1} \boldsymbol{V}' \sum_{j=1}^{m} w_j \boldsymbol{Z}_j \text{ and } \hat{\boldsymbol{\Sigma}} = \left(\boldsymbol{\Lambda}^{-1} + \sum_{j=1}^{m} w_j \boldsymbol{V}'\boldsymbol{V}\right)^{-1}$$

.

4. For $j = 1, \ldots, m$: Draw w_j from

$$p(w_j \mid \boldsymbol{Z}, \boldsymbol{w}_{[-j]}, \boldsymbol{\alpha}, \boldsymbol{\beta}, \mathcal{T}) \propto w_j^{\frac{n}{2}} \exp\left(-\frac{w_j}{2} \sum_{i=1}^{n} \left(Z_{ij} - \boldsymbol{x}_i'\boldsymbol{\beta} - \alpha_i\right)^2\right).$$

A.4 DETAILED STEP 2 IN GIBBS SAMPLING OF BARCM

For each $k \in \{q_1, \ldots, q_m\}$, we sample $\boldsymbol{Z}_{\mathcal{A}_k(\boldsymbol{q})}, \boldsymbol{\alpha}^{\langle k \rangle}, \boldsymbol{\beta}^{\langle k \rangle} \mid \mathcal{T}, \boldsymbol{q}$ as following:

1. For $i = 1, \ldots, n$, $j = 1, \ldots, m$: Draw $[Z_{ij} \mid Z_{[-i],j}, \boldsymbol{Z}_{[-j]}, \boldsymbol{\alpha}^{\langle q_j \rangle}, \boldsymbol{\beta}^{\langle q_j \rangle}]$ from $\mathcal{N}(\alpha_i^{\langle q_j \rangle} +$

$\boldsymbol{x}_i'\boldsymbol{\beta}^{\langle q_j \rangle}, 1)$ with truncation points determined by $Z_{[-i],j}$, such that Z_{ij} falls in the correct position according to τ_j^P.

2. Draw $\theta \sim S^{1/2}/\chi_{nm}$ where

$$S = \sum_k \left[\sum_{j \in \mathcal{A}_k(\boldsymbol{q})} \boldsymbol{Z}_j'\boldsymbol{Z}_j - \sum_{i,j \in \mathcal{A}_k(\boldsymbol{q})} \boldsymbol{Z}_j'\boldsymbol{V} \left(\boldsymbol{\Lambda}^{-1} + |\mathcal{A}_k(\boldsymbol{q})| \boldsymbol{V}'\boldsymbol{V} \right)^{-1} \boldsymbol{V}'\boldsymbol{Z}_i \right].$$

3. For each $k \in \{q_1, \ldots, q_m\}$: Draw $\left(\boldsymbol{\alpha}^{\langle k \rangle'}, \boldsymbol{\beta}^{\langle k \rangle'} \right)' \sim \mathcal{N}\left(\hat{\boldsymbol{\eta}}_k/\theta, \hat{\boldsymbol{\Sigma}}_k \right)$, where

$$\hat{\boldsymbol{\eta}}_k = \left(\boldsymbol{\Lambda}^{-1} + |\mathcal{A}_k(\boldsymbol{q})| \boldsymbol{V}'\boldsymbol{V} \right)^{-1} \boldsymbol{V}' \sum_{j \in \mathcal{A}_k(\boldsymbol{q})} \boldsymbol{Z}_j \text{ and } \hat{\boldsymbol{\Sigma}}_k = \left(\boldsymbol{\Lambda}^{-1} + |\mathcal{A}_k(\boldsymbol{q})| \boldsymbol{V}'\boldsymbol{V} \right)^{-1}.$$

A.5 Rank aggregation methods in comparison

A.5.1 Methods based on summary statistics

Rank aggregation methods based on summary statistics (e.g. average ranking position) are easily understood and widely used. Suppose we have m full ranking lists. Let $\{\tau_j(i)\}_{1 \le j \le m}$ be the ranking positions of entity i received from all m rankers. The Borda Count method aggregates ranks based on their arithmetic mean, $\sum_{j=1}^m \tau_j(i)/m$.

A.5.2 Markov Chain Based Methods

Dwork et al. (2001) proposed three Markov Chain based methods (MC_1, MC_2, MC_3) to solve the rank aggregation problem. The basic idea behind these methods is to construct a Markov chain with transition matrix $P = \{p_{i_1 i_2}\}_{i_1, i_2 \in U}$, where $p_{i_1 i_2}$ is the transition probability from entity i_1 to entity i_2, based on the pairwise comparison information from $\{\tau_1, \ldots, \tau_m\}$. For example, the transition rule of MC_2 is:

If the current state is i_1 then the next state is chosen by first picking a list τ uniformly from all the partial lists $\{\tau_1, \ldots, \tau_m\}$ containing entity i_1 then picking an entity i_2 uniformly from the set $\{i_2 \mid \tau(i_2) \leq \tau(i_1)\}$.

Then, the authors use the stationary distribution of this Markov chain to generate the aggregated ranking list ρ. Explicitly,

$$\rho = \text{sort}(i \in U \text{ by } \pi_i \downarrow),$$

where $\pi = (\pi_1, \ldots, \pi_{|U|})$ satisfies $\pi P = \pi$, and the symbol "$\downarrow$" means that the entities are sorted in descending order.

A.5.3 Plackett-Luce based method

PL model assumes that a ranking list $\tau = [i_1 \succ i_2 \succ \ldots \succ i_n]$ is observed with probability

$$P(\tau \mid \boldsymbol{\gamma}) = \frac{\gamma_{i_1}}{\sum_{l=1}^{n} \gamma_{i_l}} \times \frac{\gamma_{i_2}}{\sum_{l=2}^{n} \gamma_{i_l}} \times \cdots \times \frac{\gamma_{i_1}}{\gamma_{i_{n-1}} + \gamma_{i_n}},$$

where $\gamma_i \in (0, 1)$ and $\sum_{i=1}^{n} \gamma_i = 1$. Each ranking list from $\{\tau_1, \ldots, \tau_m\}$ follows the above distribution independently. We apply the classical Minorize-Maximization (MM) algorithm for PL model estimation (Hunter, 2004).

A.5.4 Stochastic optimization-based rank aggregation

Optimization-based rank aggregation methods are proposed to minimize the average distance between a candidate list and each of the input lists, i.e.,

$$\rho = \arg \min_{\sigma \in \mathcal{S}(U)} d(\sigma, \tau_1, \ldots, \tau_m) \tag{A.5.1}$$

where $\mathcal{S}(U)$ represents all allowable rankings, and $d(\cdot)$ is either the average Kendall tau distance or the average Spearman's footrule distance.

Lin & Ding (2009) used a stochastic search method to optimize (A.5.1) by adopting the cross entropy Monte Carlo (CEMC) approach (Rubinstein & Kroese, 2004). The corresponding optimization methods based on these two distance measures are denoted as CEMC_K and CEMC_F.

A.6 MCMC Diagnostic

In Figure A.1, we present convergence diagnostics for MCMC samples of $\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}$, which BARCW uses to generate aggregated ranking list. In Figure A.2, we present convergence diagnostics for MCMC samples of $\boldsymbol{\mu}^{(j)} - \bar{\boldsymbol{\mu}}^{(j)}$ in BARCM method when applied to orthodontics example.

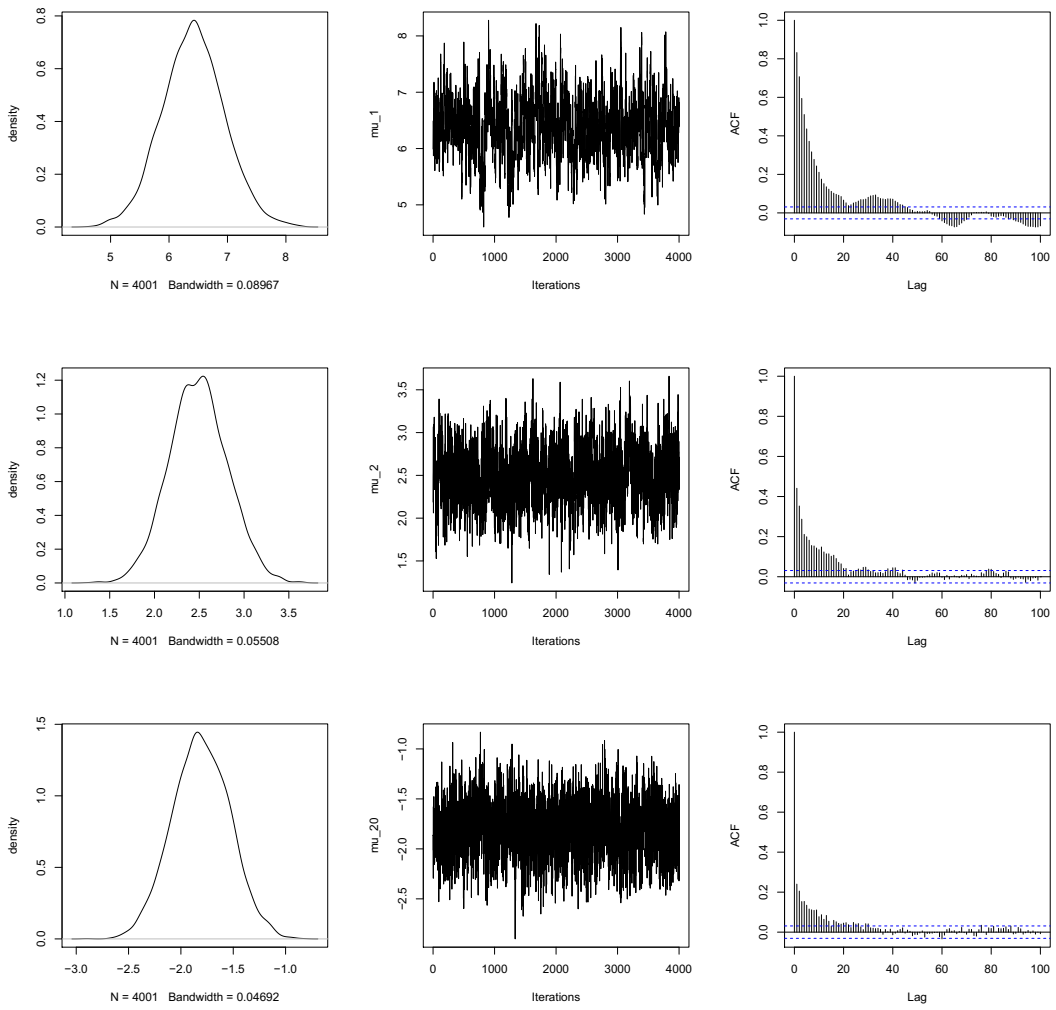


Figure A.1: Convergence diagnostics from the fit of quarterback ranking data by BARCW. We explores the convergence of MCMC samples for three typical dimensions of $\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}$. The left panel shows density plots; The middle panel shows trace-plots; The right panel shows autocorrelation plots. The effective sample size is above 300 (per 1000 saved samples) for any dimension of $\boldsymbol{\mu} - \bar{\boldsymbol{\mu}}$.

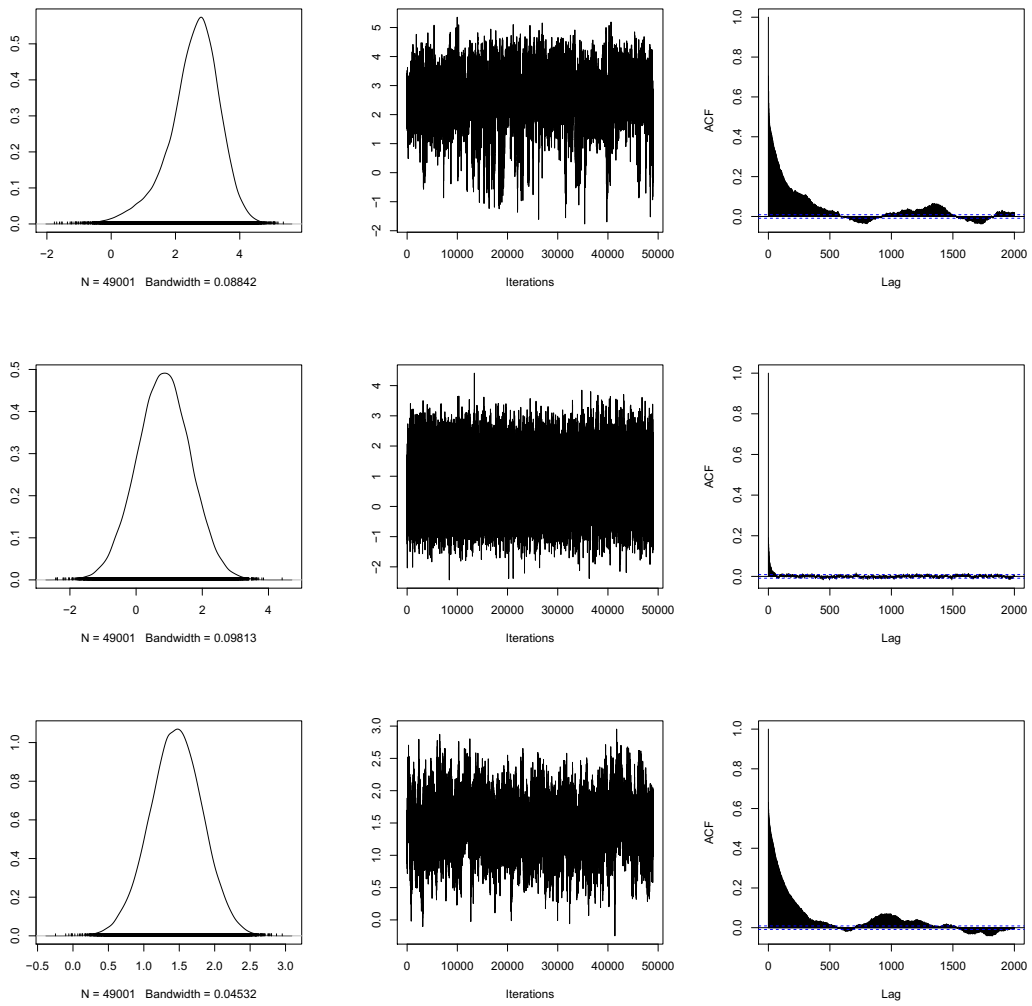


Figure A.2: Convergence diagnostics from the fit of orthodontics data by BARCM. We explores the convergence of MCMC samples for the underlying evaluation of the same entity A_1 by three different rankers via $\boldsymbol{\mu}^{(j)} - \bar{\boldsymbol{\mu}}^{(j)}$ $(j = 1,2,3)$. The left panel shows density plots; The middle panel shows trace-plots; The right panel shows autocorrelation plots.

B

Appendix to Chapter 2

B.1 Proof of propositions

Proof of Proposition 1: Let $\boldsymbol{r}_t = (z_t, \gamma_t)'$. Since $y_t = z_t + \gamma_t$ for all t, so $y_t \mid y_{1:(t-1)}, \gamma_{1:(t-1)}$ is equivalent to $z_t + \gamma_t \mid z_{1:(t-1)}, \gamma_{1:(t-1)}$. By treating $\boldsymbol{r}_t$ as 2-dimensional observed data and $\boldsymbol{\alpha}_t = (\boldsymbol{h}_t', \boldsymbol{s}_t')'$ as state vector in a state space model, we can rewrite equation

(2.2.1b)-(2.2.1e) as

$$\begin{aligned} \boldsymbol{r}_t &= \boldsymbol{\Phi}'\boldsymbol{\alpha}_t + \boldsymbol{\xi}_t \\ \boldsymbol{\alpha}_t &= \boldsymbol{\Lambda}\boldsymbol{\alpha}_{t-1} + \boldsymbol{\tau}_t \end{aligned} \tag{B.1.1}$$

where $\boldsymbol{\Phi} = \begin{bmatrix} \boldsymbol{w} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{v} \end{bmatrix}$, $\boldsymbol{\Lambda} = \begin{bmatrix} \boldsymbol{F} & \boldsymbol{O} \\ \boldsymbol{O} & \boldsymbol{P} \end{bmatrix}$ and $(\boldsymbol{\xi}_t', \boldsymbol{\tau}_t')' \overset{iid}{\sim} \mathcal{N}(\boldsymbol{0}, \boldsymbol{H})$.

We let $\boldsymbol{r}_{1:t} = (\boldsymbol{r}_1', \ldots, \boldsymbol{r}_t')'$ and $\boldsymbol{\alpha}_{1:t} = (\boldsymbol{\alpha}_1', \ldots, \boldsymbol{\alpha}_t')'$. According to the property of Gaussian linear state space model, $\boldsymbol{r}_{1:t}$ and $\boldsymbol{\alpha}_{1:t}$ jointly follows multivariate normal distribution. Therefore, the sub-vector $\boldsymbol{r}_{1:t}$ also follows multivariate normal distribution, and $\boldsymbol{r}_t \mid \boldsymbol{r}_{1:(t-1)}$ follows bivariate normal distribution with mean linear in $\boldsymbol{r}_{1:(t-1)}$, i.e.

$$\boldsymbol{r}_t \mid \boldsymbol{r}_{1:(t-1)} \sim \mathcal{N}(\boldsymbol{\Gamma}_t \boldsymbol{r}_{1:(t-1)}, \boldsymbol{\Sigma}_t), \tag{B.1.2}$$

where $\boldsymbol{\Gamma}_t$ and $\boldsymbol{\Sigma}_t$ are determined by $\boldsymbol{\Phi}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{H}$. For any given parameters, this above distribution can be numerically evaluated through Kalman filter. Here, we focus only on the general analytical formulation. Following (B.1.2) and $y_t = \boldsymbol{1}'\boldsymbol{r}_t$, we have

$$y_t \mid \boldsymbol{r}_{1:(t-1)} \sim \mathcal{N}(\boldsymbol{1}'\boldsymbol{\Gamma}_t \boldsymbol{r}_{1:(t-1)}, \boldsymbol{1}'\boldsymbol{\Sigma}_t \boldsymbol{1}).$$

Thus, given $\boldsymbol{r}_{1:(t-1)}$, or equivalently $z_{1:(t-1)}$ and $\gamma_{1:(t-1)}$, y_t follows univariate normal distribution with mean linear in $z_{1:(t-1)}$ and $\gamma_{1:(t-1)}$.

When taking exogenous variable $\boldsymbol{x}_t$ into account, we have $p(y_t \mid \boldsymbol{r}_{1:(t-1)}, \boldsymbol{x}_{1:t}) \propto p(\boldsymbol{x}_t \mid y_t) p(y_t \mid \boldsymbol{r}_{1:(t-1)})$. Under model (2.3.1),

$$p(\boldsymbol{x}_t \mid y_t) \propto \exp\left(-\frac{1}{2} (\boldsymbol{x}_t - \boldsymbol{\mu}_x - y_t \boldsymbol{\beta})' \boldsymbol{Q}^{-1} (\boldsymbol{x}_t - \boldsymbol{\mu}_x - y_t \boldsymbol{\beta}) \right).$$

Hence,

$$p(y_t \mid \boldsymbol{r}_{1:(t-1)}, \boldsymbol{x}_{1:t}) \propto \exp\left(-\frac{1}{2}(\boldsymbol{x}_t - \boldsymbol{\mu}_x - y_t\boldsymbol{\beta})'\boldsymbol{Q}^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_x - y_t\boldsymbol{\beta}) - \frac{1}{2}(\mathbf{1}'\boldsymbol{\Sigma}_t\mathbf{1})^{-1}\left(y_t - \mathbf{1}'\boldsymbol{\Gamma}_t\boldsymbol{r}_{1:(t-1)}\right)^2\right).$$

$y_t \mid \boldsymbol{r}_{1:(t-1)}, \boldsymbol{x}_{1:t}$ follows normal distribution, since the above equation is an exponential function of a quadratic form of y_t. By reorganizing the terms in above equation, we have

$$E\left(y_t \mid \boldsymbol{r}_{1:(t-1)}, \boldsymbol{x}_{1:t}\right) = \left(\left(\mathbf{1}'\boldsymbol{\Sigma}_t\mathbf{1}\right)^{-1} + \boldsymbol{\beta}'\boldsymbol{Q}^{-1}\boldsymbol{\beta}\right)^{-1}\left(\left(\mathbf{1}'\boldsymbol{\Sigma}_t\mathbf{1}\right)^{-1}\mathbf{1}'\boldsymbol{\Gamma}_t\boldsymbol{r}_{1:(t-1)} + \boldsymbol{\beta}'\boldsymbol{Q}^{-1}(\boldsymbol{x}_t - \boldsymbol{\mu}_x)\right)$$

and

$$Var\left(y_t \mid \boldsymbol{r}_{1:(t-1)}, \boldsymbol{x}_{1:t}\right) = \left(\left(\mathbf{1}'\boldsymbol{\Sigma}_t\mathbf{1}\right)^{-1} + \boldsymbol{\beta}'\boldsymbol{Q}^{-1}\boldsymbol{\beta}\right)^{-1}.$$

Therefore, $y_t \mid z_{1:(t-1)}, \gamma_{1:(t-1)}, \boldsymbol{x}_{1:t}$ also follows normal distribution with mean linear in $z_{1:(t-1)}$, $\gamma_{1:(t-1)}$ and $\boldsymbol{x}_t$.

Proof of Proposition 2: Under the same notation as our previous proof, we now consider the following predictive distribution

$$\begin{aligned} p\left(y_{t+l} \mid \boldsymbol{x}_{1:t}, \boldsymbol{r}_{1:(t-1)}\right) &\propto \int p\left(y_{t+l}, \boldsymbol{r}_t \mid \boldsymbol{x}_{1:t}, \boldsymbol{r}_{1:(t-1)}\right) \mathrm{d}\boldsymbol{r}_t \\ &\propto \int p\left(y_{t+l} \mid \boldsymbol{x}_{1:t}, \boldsymbol{r}_{1:t}\right) p\left(\boldsymbol{r}_t \mid \boldsymbol{x}_{1:t}, \boldsymbol{r}_{1:(t-1)}\right) \mathrm{d}\boldsymbol{r}_t. \end{aligned}$$

Under model (2.3.1), $\boldsymbol{x}_{1:t}$ is independent of y_{t+l} conditional on $y_{1:t}$. Hence, $p(y_{t+l} \mid \boldsymbol{x}_{1:t}, \boldsymbol{r}_{1:t}) = p(y_{t+l} \mid \boldsymbol{r}_{1:t})$. Similarly, $\boldsymbol{x}_{1:(t-1)}$ is independent of $\boldsymbol{r}_t$ conditional on $\boldsymbol{r}_{1:(t-1)}$, implying $p(\boldsymbol{r}_t \mid \boldsymbol{x}_{1:t}, \boldsymbol{r}_{1:(t-1)}) = p(\boldsymbol{r}_t \mid \boldsymbol{x}_t, \boldsymbol{r}_{1:(t-1)})$. Note that $p\left(\boldsymbol{r}_t \mid \boldsymbol{x}_t, \boldsymbol{r}_{1:(t-1)}\right) \propto$

$p(\boldsymbol{x}_t \mid \boldsymbol{r}_t)p(\boldsymbol{r}_t \mid \boldsymbol{r}_{1:(t-1)})$. Thus, we have

$$\begin{aligned} p\left(y_{t+l}, \boldsymbol{r}_t \mid \boldsymbol{x}_{1:t}, \boldsymbol{r}_{1:(t-1)}\right) &\propto p\left(y_{t+l} \mid \boldsymbol{r}_{1:t}\right) p\left(\boldsymbol{r}_t \mid \boldsymbol{x}_t, \boldsymbol{r}_{1:(t-1)}\right) \\ &\propto p\left(y_{t+l} \mid \boldsymbol{r}_{1:t}\right) p(\boldsymbol{x}_t \mid \boldsymbol{r}_t) p(\boldsymbol{r}_t \mid \boldsymbol{r}_{1:(t-1)}). \end{aligned}$$

In the previous proof, we learned that $\boldsymbol{r}_{1:(t+l)}$ also follows multivariate normal distribution. Similar to equation (B.1.2), we can write $\boldsymbol{r}_{t+l} \mid \boldsymbol{x}_{1:t}$ under the following representation,

$$\boldsymbol{r}_{t+l} \mid \boldsymbol{r}_{1:t} \sim \mathcal{N}(\boldsymbol{\Gamma}_{t,l}\boldsymbol{r}_{1:t}, \boldsymbol{\Sigma}_{t,l}), \tag{B.1.3}$$

where $\boldsymbol{\Gamma}_{t,l}$ and $\boldsymbol{\Sigma}_{t,l}$ are determined by $\boldsymbol{\Phi}$, $\boldsymbol{\Lambda}$ and $\boldsymbol{H}$. Hence, $y_{t+l} \mid \boldsymbol{r}_{1:t} \sim \mathcal{N}(\mathbf{1}'\boldsymbol{\Gamma}_{t,l}\boldsymbol{r}_{1:t}, \mathbf{1}'\boldsymbol{\Sigma}_{t,l}\mathbf{1})$. Combining the above results with model (2.3.1), we have

$$\begin{aligned} p\left(y_{t+l}, \boldsymbol{r}_t \mid \boldsymbol{x}_{1:t}, \boldsymbol{r}_{1:(t-1)}\right) \propto \exp\Bigg(&-\frac{1}{2}(\mathbf{1}'\boldsymbol{\Sigma}_{t,l}\mathbf{1})^{-1}\left(y_{t+l} - \mathbf{1}'\boldsymbol{\Gamma}_{t,l}\boldsymbol{r}_{1:t}\right)^2 - \frac{1}{2}(\boldsymbol{x}_t - \boldsymbol{\mu}_x - \boldsymbol{\beta}\mathbf{1}'\boldsymbol{r}_t)'\boldsymbol{Q}^{-1} \\ &(\boldsymbol{x}_t - \boldsymbol{\mu}_x - \boldsymbol{\beta}\mathbf{1}'\boldsymbol{r}_t) - \frac{1}{2}\left(\boldsymbol{r}_t - \boldsymbol{\Gamma}_{t-1,1}\boldsymbol{r}_{1:(t-1)}\right)' \boldsymbol{\Sigma}_{t-1,1}^{-1}\left(\boldsymbol{r}_t - \boldsymbol{\Gamma}_{t-1,1}\boldsymbol{r}_{1:(t-1)}\right)\Bigg), \end{aligned} \tag{B.1.4}$$

which is a exponential function of quadratic form of y_t and $\boldsymbol{r}_t$. Therefore, $y_{t+l}, \boldsymbol{r}_t \mid \boldsymbol{x}_{1:t}, \boldsymbol{r}_{1:(t-1)}$ follows multivariate normal distribution, whereas $y_{t+l} \mid \boldsymbol{x}_{1:t}, \boldsymbol{r}_{1:(t-1)}$ follows univariate normal distribution. Moreover, the conditional expectation is

$$\begin{aligned} & E\left(y_{t+l} \mid \boldsymbol{x}_{1:t}, \boldsymbol{r}_{1:(t-1)}\right) \\ = \; & E\left(E\left(y_{t+l} \mid \boldsymbol{r}_{1:t}\right) \mid \boldsymbol{x}_{1:t}, \boldsymbol{r}_{1:(t-1)}\right) \\ = \; & E\left(\boldsymbol{\Gamma}_{t,l}(\boldsymbol{r}'_{1:(t-1)}, \boldsymbol{r}'_t)' \mid \boldsymbol{x}_{1:t}, \boldsymbol{r}_{1:(t-1)}\right) \\ = \; & \boldsymbol{\Gamma}_{t,l}\left(\boldsymbol{r}'_{1:(t-1)}, E(\boldsymbol{r}_t \mid \boldsymbol{x}_t, \boldsymbol{r}_{1:(t-1)})'\right)', \end{aligned}$$

where $E(\boldsymbol{r}_t \mid \boldsymbol{x}_t, \boldsymbol{r}_{1:(t-1)})$ is linear in $\boldsymbol{x}_t$ and $\boldsymbol{r}_{1:(t-1)}$. Therefore, $y_{t+l} \mid \boldsymbol{x}_{1:t}, \boldsymbol{r}_{1:(t-1)}$ follows univariate normal distribution with mean linear in $\boldsymbol{x}_t$ and $\boldsymbol{r}_{1:(t-1)}$.

B.2 Robustness to seasonal decomposition method choice

We compare the performance of PRISM with two different seasonal decomposition methods: STL and the classic additive decomposition. Both methods decompose target time series y_t into the trend component T_t, the seasonal component S_t and the irregular component R_t:

$$y_t = T_t + S_t + R_t.$$

In the classic additive decomposition, the trend component T_t is calculated from moving average of $\{y_t\}$ and the seasonal component S_t is simply assumed the same each period. In contrast, STL relies on a sequence of applications of loess smoother to generate the seasonal and trend components. Both STL and classic additive seasonal decomposition are options in the R package of PRISM method with STL being the default option.

Table B.1: The performance of PRISM with two different seasonal decomposition methods: STL and additive decomposition.

	real-time	forecast 1 wk	forecast 2 wk	forecast 3 wk
RMSE				
additive decompostion	**0.496**	0.497	0.462	**0.460**
STL decompostion	0.498	**0.492**	**0.453**	0.467
MAE				
additive decompostion	0.543	0.538	0.482	**0.458**
STL decompostion	**0.542**	**0.534**	**0.479**	0.465

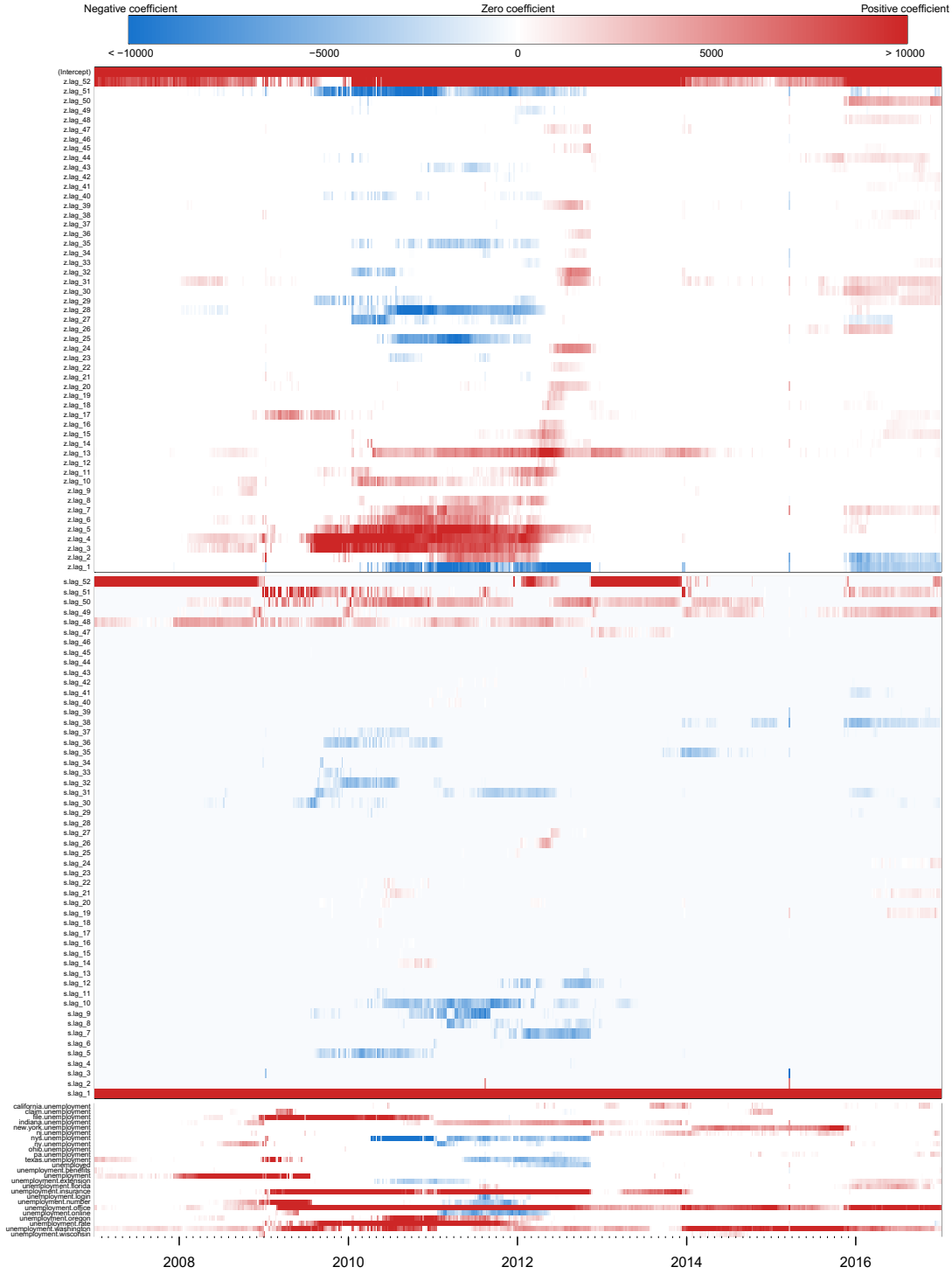
B.3 Effect of the discount factor

We test the effect of discount factor w between 0.95 and 0.995, as suggested in Lindoff (1997). Table B.2 shows the performance of PRISM with different w. The discount factor w is an option in the R package that implements PRISM method, and the default value is set to be 0.99.

Table B.2: The performance of PRISM for $w \in [0.95, 0.995]$.

	real-time	forecast 1 wk	forecast 2 wk	forecast 3 wk
RMSE				
$w = 0.995$	0.512	0.501	0.463	0.462
$w = 0.99$	0.495	0.491	**0.457**	**0.461**
$w = 0.985$	**0.494**	0.486	0.462	0.466
$w = 0.98$	0.506	0.489	0.471	0.472
$w = 0.975$	0.500	0.485	0.471	0.468
$w = 0.97$	0.513	0.488	0.475	0.495
$w = 0.965$	0.516	0.488	0.485	0.479
$w = 0.96$	0.519	0.492	0.486	0.483
$w = 0.955$	0.525	0.502	0.502	0.497
$w = 0.95$	0.540	**0.484**	0.482	0.491
MAE				
$w = 0.995$	0.556	0.546	0.487	0.467
$w = 0.99$	0.545	0.532	**0.478**	0.461
$w = 0.985$	**0.538**	0.522	0.480	**0.460**
$w = 0.98$	0.543	0.517	0.484	**0.460**
$w = 0.975$	0.540	**0.516**	0.481	0.461
$w = 0.97$	0.549	0.519	0.488	0.482
$w = 0.965$	0.555	0.523	0.495	0.465
$w = 0.96$	0.552	0.524	0.496	0.468
$w = 0.955$	0.563	0.533	0.511	0.477
$w = 0.95$	0.572	0.521	0.496	0.474

B.4 Coefficient heatmap


Negative coefficient
Zero coefficient
Positive coefficient
< −10000
−5000
0
5000
> 10000
(Intercept)
z.lag_52
z.lag_51
z.lag_50
z.lag_49
z.lag_48
z.lag_47
z.lag_46
z.lag_45
z.lag_44
z.lag_43
z.lag_42
z.lag_41
z.lag_40
z.lag_39
z.lag_38
z.lag_37
z.lag_36
z.lag_35
z.lag_34
z.lag_33
z.lag_32
z.lag_31
z.lag_30
z.lag_29
z.lag_28
z.lag_27
z.lag_26
z.lag_25
z.lag_24
z.lag_23
z.lag_22
z.lag_21
z.lag_20
z.lag_19
z.lag_18
z.lag_17
z.lag_16
z.lag_15
z.lag_14
z.lag_13
z.lag_12
z.lag_11
z.lag_10
z.lag_9
z.lag_8
z.lag_7
z.lag_6
z.lag_5
z.lag_4
z.lag_3
z.lag_2
z.lag_1
s.lag_52
s.lag_51
s.lag_50
s.lag_49
s.lag_48
s.lag_47
s.lag_46
s.lag_45
s.lag_44
s.lag_43
s.lag_42
s.lag_41
s.lag_40
s.lag_39
s.lag_38
s.lag_37
s.lag_36
s.lag_35
s.lag_34
s.lag_33
s.lag_32
s.lag_31
s.lag_30
s.lag_29
s.lag_28
s.lag_27
s.lag_26
s.lag_25
s.lag_24
s.lag_23
s.lag_22
s.lag_21
s.lag_20
s.lag_19
s.lag_18
s.lag_17
s.lag_16
s.lag_15
s.lag_14
s.lag_13
s.lag_12
s.lag_11
s.lag_10
s.lag_9
s.lag_8
s.lag_7
s.lag_6
s.lag_5
s.lag_4
s.lag_3
s.lag_2
s.lag_1
california.unemployment
claim.unemployment
file.unemployment
indiana.unemployment
new.york.unemployment
rate.unemployment
nj.unemployment
ny.unemployment
ohio.unemployment
pa.unemployment
texas.unemployment
unemployed
unemployment.benefits
unemployment
unemployment.file
unemployment.florida
unemployment.insurance
unemployment.login
unemployment.number
unemployment.office
unemployment.online
unemployment.rate
unemployment.site
unemployment.washington
unemployment.wisconsin
2008
2010
2012
2014
2016

C

Appendix to Chapter 3

C.1 THE DETAILED E-STEP

We show the details about Q functions in the E-step.

$$
\begin{aligned}
& Q_1^{(m)}\left(\tau_{1:N}, \boldsymbol{\Lambda}_{1:(N+1)}, K\right) \\
= \; & \mathrm{E}_{\boldsymbol{\Gamma}, \boldsymbol{F} \mid \boldsymbol{Y}, \boldsymbol{\Omega}^{(m)}}\left[\log p\left(\tau_{1:N}, \boldsymbol{\Lambda}_{1:(N+1)}, K, \boldsymbol{\Gamma}, \boldsymbol{F}\right)\right] \\
= \; & \left\langle \log p\left(\tau_{1:N}, K\right)\right\rangle + \left\langle \log p\left(\boldsymbol{\Lambda}_{1:(N+1)} \mid \tau_{1:N}, K\right)\right\rangle + \left\langle \log p\left(\boldsymbol{F} \mid \boldsymbol{\Lambda}_{1:(N+1)}\right)\right\rangle + \left\langle \log p\left(\boldsymbol{\Gamma} \mid K\right)\right\rangle .
\end{aligned}
$$

In the above equation, the first term $\left\langle \log p\left(\tau_{1:N}, K\right)\right\rangle$ is given by (3.3.10) as $-\frac{1}{2} NK \log T$.

In the second term, based on (3.3.5),

$$
p\left(\boldsymbol{\Lambda}_{1:(N+1)} \mid \tau_{1:N}, K\right) \propto \prod_{j=1}^{N+1} p\left(\boldsymbol{\Lambda}_j \mid K\right) \propto \prod_{j=1}^{N+1} \prod_{k=1}^{K} p\left(\lambda_{jk}^2\right),
$$

where $p\left(\lambda_{jk}^2\right) \propto \lambda_{jk}^{-2(1+\eta/2)} \exp\left(-\frac{\eta s_\lambda^2}{2\lambda_{jk}^2}\right)$. Based on (3.3.9),

$$
p\left(\boldsymbol{F} \mid \boldsymbol{\Lambda}_{1:(N+1)}\right) \propto \prod_{j=1}^{N+1} \prod_{t=\tau_{j-1}+1}^{\tau_j} p\left(\boldsymbol{f}_t \mid \boldsymbol{\Lambda}_j\right),
$$

where

$$
p\left(\boldsymbol{f}_t \mid \boldsymbol{\Lambda}_j\right) \propto \prod_{k=1}^{K} \lambda_{jk}^{-2} \exp\left(-\frac{f_{tk}^2}{2\lambda_{jk}^2}\right) \prod_{k=K+1}^{K^*} \lambda_0^{-2} \exp\left(-\frac{f_{tk}^2}{2\lambda_0^2}\right).
$$

Finally, based on (3.3.8),

$$p\left(\mathbf{\Gamma} \mid K\right) \propto \prod_{k=1}^{K}\prod_{i=1}^{d}\theta_1^{\gamma_{ik}}(1-\theta_1)^{1-\gamma_{ik}} \prod_{k=K+1}^{K^*}\prod_{i=1}^{d}\theta_0^{\gamma_{ik}}(1-\theta_0)^{1-\gamma_{ik}}.$$

Therefore, we have (3.4.4) from by combining the pieces above. The derivation of (3.4.5) is very similar to the E-step in Ročková & George (2016a), and thus omitted here.

C.2 The detailed M-step

C.2.1 The PELT method implementation

We first explicity define the cost function

$$\mathcal{C}\left(\tau_{j-1}+1, \tau_j\right) = -\max_{\mathbf{\Lambda}_j}\left[\log p\left(\mathbf{\Lambda}_j \mid K^{(m)}\right) + \sum_{t=\tau_{j-1}+1}^{\tau_j}\left\langle \log p\left(\boldsymbol{f}_t \mid \mathbf{\Lambda}_j, K^{(m)}\right)\right\rangle\right]$$

where

$$\left\langle \log p\left(\boldsymbol{f}_t \mid \mathbf{\Lambda}_j, K\right)\right\rangle = C + \frac{1}{2}\sum_{k=1}^{K}\left[\frac{\left\langle f_{tk}^2\right\rangle}{\lambda_{jk}^2} + \log\lambda_{jk}^2\right] + \frac{1}{2}\sum_{k=K+1}^{K^*}\left[\frac{\left\langle f_{tk}^2\right\rangle}{\lambda_0^2} + \log\lambda_0^2\right]$$

and

$$\log p\left(\mathbf{\Lambda}_j \mid K\right) = C' + \frac{1}{2}\sum_{k=1}^{K}\left[\frac{\eta s_\lambda^2}{\lambda_{jk}^2} + \log\lambda_{jk}^2\right]$$

with constant C and C' not changing with Λ_j. Let penalty constant $c = \frac{1}{2}K\log T$, then we initialize the process with $F(0) = -c$, $cp(0) =$ NULL and $R_1 = \{0\}$. Based on Algorithm 2 in Killick et al. (2012), we iterate following steps for $\tau^* = 1, \ldots, T$

1. Calculate $F(\tau^*) = \min_{\tau\in R_{\tau^*}}\left[F(\tau) + \mathcal{C}(\tau+1, \tau^*) + c\right]$.

2. Let $\tau^1 = \arg\min_{\tau \in R_{\tau^*}} [F(\tau) + \mathcal{C}(\tau + 1, \tau^*) + c]$.

3. Set $cp(\tau^*) = [cp(\tau^1), \tau^1]$.

4. Set $R_{\tau^*+1} = \{\tau^*\} \cup \{\tau \in R_{\tau^*} : F(\tau) + \mathcal{C}(\tau + 1, \tau^*) < F(\tau^*)\}$

Then, we update $\tau_{1:N}^{(m+1)}$ to the change points recorded in $cp(T)$, and update

$$\boldsymbol{\Lambda}_j^{(m+1/2)} = \arg\max_{\boldsymbol{\Lambda}_j} \left[\log p\left(\boldsymbol{\Lambda}_j \mid K^{(m)}\right) + \sum_{t=\tau_{j-1}^{(m+1)}+1}^{\tau_j^{(m+1)}} \left\langle \log p\left(\boldsymbol{f}_t \mid \boldsymbol{\Lambda}_j, K^{(m)}\right) \right\rangle \right]$$

for $j = 1, \ldots, N^{(m+1)} + 1$ where $N^{(m+1)}$ is determined by the length of $\tau_{1:N}^{(m+1)}$.

C.2.2 The detailed M-step for Q_2

Following Theorem 3.1 in Ročková & George (2016a), the maximization of Q_2 can be interpreted as a log-posterior arising from a series of independent penalized regressions. Note that $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T)'$, $\langle \boldsymbol{F} \rangle = (\langle \boldsymbol{f}_1 \rangle, \ldots, \langle \boldsymbol{f}_T \rangle)'$ and $\boldsymbol{B}_{p \times K} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_p)'$. We denote $\widetilde{\boldsymbol{Y}} = \begin{pmatrix} \boldsymbol{Y} \\ \boldsymbol{0} \end{pmatrix} \in \mathbb{R}^{(T+K) \times p}$ and its columns be $\widetilde{\boldsymbol{y}}^1, \ldots, \widetilde{\boldsymbol{y}}^p$. Then, let $\widetilde{\boldsymbol{F}} = \begin{pmatrix} \langle \boldsymbol{F} \rangle \\ \boldsymbol{M}^{1/2} \end{pmatrix} \in \mathbb{R}^{(T+K) \times p}$ where $\boldsymbol{M}^{1/2}$ is the square root of $\sum_{t=1}^{T} \left(\langle \boldsymbol{f}_t \boldsymbol{f}_t' \rangle - \langle \boldsymbol{f}_t \rangle \langle \boldsymbol{f}_t \rangle' \right)$. Based on this penalized regression formulation, $\boldsymbol{\beta}_i^{(m+1)}$ can be obtained conditionally on $\boldsymbol{\Sigma}^{(m)}$ by an adaptive LASSO (Zou, 2006), in which

$$\boldsymbol{\beta}_i^{(m+1)} = \arg\min_{\boldsymbol{\beta}_i} \left\{ \|\widetilde{\boldsymbol{y}}^i - \widetilde{\boldsymbol{F}}\boldsymbol{\beta}_i\|^2 + 2\sigma_i^{(m)^2} \sum_{k=1}^{K} |\beta_{ik}| \delta_{ik} \right\}.$$

Then, conditional on $\boldsymbol{B}^{(m+1)}$, we apply a closed form update $\boldsymbol{\Sigma}^{(m+1)}$ with

$$\sigma_i^{(m+1)^2} = \frac{1}{T+\xi+2}\left(\|\widetilde{\boldsymbol{y}}^i - \widetilde{\boldsymbol{F}}\boldsymbol{\beta}_i^{(m+1)}\|^2 + \xi s_\sigma^2\right).$$

References

Aguilar, O. & West, M. (2000). Bayesian dynamic factor models and portfolio allocation. *Journal of Business & Economic Statistics*, 18(3), 338–357.

Albert, J. H. & Chib, S. (1993). Bayesian analysis of binary and polychotomous response data. *Journal of the American statistical Association*, 88, 669–679.

Alvo, M. & Yu, P. L. (2014). *Statistical Methods for Ranking Data*. Springer-Verlag New York.

Ameen, J. & Harrison, P. (1984). Discount weighted estimation. *Journal of Forecasting*, 3(3), 285–296.

Antoniak, C. E. (1974). Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, (pp. 1152–1174).

Aoki, M. (1987). *State space modeling of time series*. Springer.

Asai, M., McAleer, M., & Yu, J. (2006). Multivariate stochastic volatility: a review. *Econometric Reviews*, 25(2-3), 145–175.

Bai, J. & Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica*, 70(1), 191–221.

Banbura, M., Giannone, D., Modugno, M., & Reichlin, L. (2013). Now-casting and the real-time data flow.

Barigozzi, M., Cho, H., & Fryzlewicz, P. (2016). Simultaneous multiple change-point and factor analysis for high-dimensional time series. *arXiv preprint arXiv:1612.06928*.

Barnett, I. & Onnela, J.-P. (2016). Change point detection in correlation networks. *Scientific reports*, 6.

Barry, D. & Hartigan, J. A. (1993). A bayesian analysis for change point problems. *Journal of the American Statistical Association*, 88(421), 309–319.

Bauwens, L., Laurent, S., & Rombouts, J. V. (2006). Multivariate garch models: a survey. *Journal of applied econometrics*, 21(1), 79–109.

Benter, W. (1994). Computer-Based Horse Race Handicapping and Wagering Systems: A Report. In W. T. Ziemba, V. S. Lo, & D. B. Haush (Eds.), *Efficiency Of Racetrack Betting Markets* (pp. 183–198).

Bhowmik, A. & Ghosh, J. (2017). LETOR Methods for Unsupervised Rank Aggregation. In *Proceedings of the 26th International Conference on World Wide Web* (pp. 1331–1340).: International World Wide Web Conferences Steering Committee.

Blackwell, D. & MacQueen, J. B. (1973). Ferguson distributions via pólya urn schemes. *The annals of statistics*, (pp. 353–355).

Böckenholt, U. (1992). Thurstonian representation for partial ranking data. *British Journal of Mathematical and Statistical Psychology*, 45(1), 31–49.

Böckenholt, U. (1993). Applications of Thurstonian Models to Ranking Data. In M. A. Fligner & J. S. Verducci (Eds.), *Probability Models and Statistical Analyses for Ranking Data* (pp. 157–172). New York, NY: Springer New York.

Böckenholt, U. (2006). Thurstonian-Based Analyses: Past, Present, and Future Utilities. *Psychometrika*, 71, 615–629.

Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: a multivariate generalized arch model. *The review of economics and statistics*, (pp. 498–505).

Borda, J. C. (1781). Mémoire sur les élections au scrutin.

Carvalho, C. M., Chang, J., Lucas, J. E., Nevins, J. R., Wang, Q., & West, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, 103(484), 1438–1456.

Chamberlain, G. & Rothschild, M. (1983). Arbitrage, factor structure, and mean-variance analysis on large asset markets. *Econometrica*, 51(5), 1281–1304.

Chen, H., Chiang, R. H., & Storey, V. C. (2012). Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 36(4).

Chib, S., Nardari, F., & Shephard, N. (2006). Analysis of high dimensional multivariate stochastic volatility models. *Journal of Econometrics*, 134(2), 341–371.

Choi, H. & Varian, H. (2012). Predicting the present with google trends. *Economic Record*, 88(s1), 2–9.

Cleveland, R. B., Cleveland, W. S., McRae, J. E., & Terpenning, I. (1990). Stl: A seasonal-trend decomposition procedure based on loess. *Journal of Official Statistics*, 6(1), 3–73.

Critchlow, D. E., Fligner, M. A., & Verducci, J. S. (1991). Probability models on rankings. *Journal of mathematical psychology*, 35(3), 294–318.

Daniels, H. E. (1950). Rank correlation and population models. *Journal of the Royal Statistical Society. Series B (Methodological)*, 12, 171–191.

De Livera, A. M., Hyndman, R. J., & Snyder, R. D. (2011). Forecasting time series with complex seasonal patterns using exponential smoothing. *Journal of the American Statistical Association*, 106(496), 1513–1527.

DeConde, R. P., Hawley, S., Falcon, S., Clegg, N., Knudsen, B., & Etzioni, R. (2006). Combining results of microarray experiments: a rank aggregation approach. *Statistical Applications in Genetics and Molecular Biology*, 5.

Dempster, A. P., Laird, N. M., & Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the royal statistical society. Series B (methodological)*, (pp. 1–38).

Deng, K., Han, S., Li, K. J., & Liu, J. S. (2014). Bayesian aggregation of order-based rank data. *Journal of the American Statistical Association*, 109, 1023–1039.

Diaconis, P. & Graham, R. L. (1977). Spearman's footrule as a measure of disarray. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 262–268).

Diebold, F. X. & Nerlove, M. (1989). The dynamics of exchange rate volatility: a multivariate latent factor arch model. *Journal of Applied econometrics*, 4(1), 1–21.

Durbin, J. & Koopman, S. J. (2012). *Time series analysis by state space methods*, volume 38. OUP Oxford.

Dwork, C., Kumar, R., Naor, M., & Sivakumar, D. (2001). Rank aggregation methods for the web. In *Proceedings of the 10th international conference on World Wide Web* (pp. 613–622).: ACM.

Einav, L. & Levin, J. (2014). The data revolution and economic analysis. *Innovation Policy and the Economy*, 14(1), 1–24.

Engle, R. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business & Economic Statistics*, 20(3), 339–350.

Engle, R. F. (2007). High dimension dynamic correlations.

Engle, R. F., Shephard, N., & Sheppard, K. (2008). Fitting vast dimensional time-varying covariance models.

Ettredge, M., Gerdes, J., & Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the ACM*, 48(11), 87–92.

Fan, J., Liao, Y., & Mincheva, M. (2013). Large covariance estimation by thresholding principal orthogonal complements. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 75(4), 603–680.

Fan, J., Zhang, J., & Yu, K. (2012). Vast portfolio selection with gross-exposure constraints. *Journal of the American Statistical Association*, 107(498), 592–606.

Fearnhead, P. (2006). Exact and efficient bayesian inference for multiple changepoint problems. *Statistics and computing*, 16(2), 203–213.

Ferguson, T. S. et al. (1983). Bayesian density estimation by mixtures of normal distributions. *Recent advances in statistics*, 24(1983), 287–302.

Forni, M., Hallin, M., Lippi, M., & Reichlin, L. (2000). The generalized dynamic-factor model: Identification and estimation. *The review of Economics and Statistics*, 82(4), 540–554.

Gardner Jr, E. S. & McKenzie, E. (1985). Forecasting trends in time series. *Management Science*, 31(10), 1237–1246.

George, E. I. & McCulloch, R. E. (1993). Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423), 881–889.

Giannone, D., Reichlin, L., & Small, D. (2008). Nowcasting: The real-time informational content of macroeconomic data. *Journal of Monetary Economics*, 55(4), 665–676.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232), 1012–1014.

Goel, S., Hofman, J. M., Lahaie, S., Pennock, D. M., & Watts, D. J. (2010). Predicting consumer behavior with web search. *Proceedings of the National academy of sciences*, 107(41), 17486–17490.

Gormley, I. C. & Murphy, T. B. (2006). Analysis of Irish third-level college applications data. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 169, 361–379.

Gormley, I. C. & Murphy, T. B. (2008a). Exploring Voting Blocs within the Irish Electorate: A Mixture Modeling Approach. *Journal of the American Statistical Association*, 103, 1014–1027.

Gormley, I. C. & Murphy, T. B. (2008b). A mixture of experts model for rank data with applications in election studies. *The Annals of Applied Statistics*, 2, 1452–1477.

Gormley, I. C. & Murphy, T. B. (2010). Clustering ranked preference data using sociodemographic covariates. In *Choice Modelling: The State-of-the-art and The State-of-practice: Proceedings from the Inaugural International Choice Modelling Conference* (pp. 543–569).

Gould, P. G., Koehler, A. B., Ord, J. K., Snyder, R. D., Hyndman, R. J., & Vahid-Araghi, F. (2008). Forecasting time series with multiple seasonal patterns. *European Journal of Operational Research*, 191(1), 207–222.

Gray-Davies, T., Holmes, C. C., & Caron, F. (2016). Scalable Bayesian nonparametric regression via a Plackett-Luce model for conditional ranks. *Electronic Journal of Statistics*, 10, 1807–1828.

Harvey, A. & Koopman, S. J. (1993). Forecasting hourly electricity demand using time-varying splines. *Journal of the American Statistical Association*, 88(424), 1228–1236.

Harvey, A., Ruiz, E., & Shephard, N. (1994). Multivariate stochastic variance models. *The Review of Economic Studies*, 61(2), 247–264.

Harvey, A. C. (1989). *Forecasting, structural time series models and the Kalman filter*. Cambridge University Press.

Harvey, A. C. & Koopman, S. J. (1997). Multivariate structural time series models. In C. Heij, H. Schumacher, B. Hanzon, & C. Praagman (Eds.), *Systematic Dynamics in Economic and Financial Models* (pp. 269–98). Chichester: John Wiley and Sons.

Harvey, A. C. & Shephard, N. (1993). Structural time series models. In R. C. R. Maddala, G S & H. D. Vinod (Eds.), *Handbook of Statistics. Vol. 11 : Econometrics* (pp. 261–302). Amsterdam: North–Holland.

Holt, C. (1957). *Forecasting trends and seasonals by exponentially weighted averages.* Technical report, Carnegie Institute of Technology.

Hunter, D. R. (2004). Mm algorithms for generalized bradley-terry models. *Annals of Statistics*, (pp. 384–406).

Hyndman, R., Koehler, A. B., Ord, J. K., & Snyder, R. D. (2008). *Forecasting with exponential smoothing: the state space approach*. Springer Science & Business Media.

Jacquier, É., Polson, N. G., & Rossi, P. E. (1995). Models and priors for multivariate stochastic volatility.

Johnson, T. R. ad Kuhn, K. M. (2013). Bayesian thurstonian models for ranking data using jags. *Behavior research methods*, 45(3), 857–872.

Johnson, V. E., Deaner, R. O., & Van Schaik, C. P. (2002). Bayesian analysis of rank data with application to primate intelligence experiments. *Journal of the American Statistical Association*, 97(457), 8–17.

Keller, T. A., Just, M. A., & Stenger, V. A. (2001). Reading span and the time-course of cortical activation in sentence-picture verification. In *Annual Convention of the Psychonomic Society*, volume 686: Orlando, FL.

Kendall, M. & Stuart, A. (1983). *The advanced theory of statistics. Vol. 3*. Griffin.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika*, (pp. 81–93).

Khoury, M. J. & Ioannidis, J. P. (2014). Big data meets public health. *Science*, 346(6213), 1054–1055.

Killick, R., Fearnhead, P., & Eckley, I. A. (2012). Optimal detection of changepoints with a linear computational cost. *Journal of the American Statistical Association*, 107(500), 1590–1598.

Kim, G.-H., Trimi, S., & Chung, J.-H. (2014). Big-data applications in the government sector. *Communications of the ACM*, 57(3), 78–85.

Knowles, D. & Ghahramani, Z. (2011). Nonparametric bayesian sparse factor models with application to gene expression modeling. *The Annals of Applied Statistics*, (pp. 1534–1552).

Lee, M. D., Steyvers, M., & Miller, B. (2014). A Cognitive Model for Aggregating People's Rankings. *PLOS ONE*, 9, 1–9.

Lin, S. (2010). Rank aggregation methods. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2, 555–570.

Lin, S. & Ding, J. (2009). Integration of ranked lists via cross entropy monte carlo with applications to mrna and microrna studies. *Biometrics*, 65, 9–18.

Lindoff, B. (1997). On the optimal choice of the forgetting factor in the recursive least squares estimator.

Liu, C., Rubin, D. B., & Wu, Y. N. (1998). Parameter expansion to accelerate em: The px-em algorithm. *Biometrika*, (pp. 755–770).

Liu, J. S. & Wu, Y. N. (1999). Parameter expansion for data augmentation. *Journal of the American Statistical Association*, 94, 1264–1274.

Liu, S., Shen, G., Bai, D., Zhou, H., Li, S., Chen, W., Wang, D., Li, W., Geng, Z., & Xu, T. (2012). Consistency of the subjective evaluation of malocclusion severity by the chinese orthodontic experts. *Beijing da xue xue bao. Yi xue ban= Journal of Peking University. Health sciences*, 44(1), 98–102.

Longerstaey, J. & Spencer, M. (1996). Riskmetrics technical document. *Morgan Guaranty Trust Company of New York: New York*.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Ma, S. & Su, L. (2016). Estimation of large dimensional factor models with an unknown number of breaks.

MacEachern, S. N. (1994). Estimating normal means with a conjugate style dirichlet process prior. *Communications in Statistics-Simulation and Computation*, 23(3), 727–741.

Mallows, C. L. (1957). Non-null ranking models. i. *Biometrika*, (pp. 114–130).

Marden, J. I. (1996). *Analyzing and modeling rank data*. CRC Press.

Markowitz, H. (1952). Portfolio selection. *The journal of finance*, 7(1), 77–91.

Maydeu-Olivares, A. (1999). Thurstonian modeling of ranking data via mean and covariance structure analysis. *Psychometrika*, 64, 325–340.

McAfee, A. & Brynjolfsson, E. (2012). Big data: the management revolution. *Harvard business review*, 90(10), 60–68.

McKinsey Global Institute (2011). Big data: The next frontier for innovation, competition, and productivity.

Meng, X.-L. & Rubin, D. B. (1993). Maximum likelihood estimation via the ecm algorithm: A general framework. *Biometrika*, 80(2), 267–278.

Mosteller, F. (1951). Remarks on the method of paired comparisons: I. the least squares solution assuming equal standard deviations and equal correlations. *Psychometrika*, 16(1), 3–9.

Murdoch, T. B. & Detsky, A. S. (2013). The inevitable application of big data to health care. *Jama*, 309(13), 1351–1352.

Neal, R. M. (1992). Bayesian mixture modeling. In *Maximum Entropy and Bayesian Methods* (pp. 197–211). Springer.

Neal, R. M. (2000). Markov chain sampling methods for dirichlet process mixture models. *Journal of computational and graphical statistics*, 9(2), 249–265.

Ord, J. K., Koehler, A. B., & Snyder, R. D. (1997). Estimation and prediction for a class of dynamic nonlinear statistical models. *Journal of the American Statistical Association*, 92(440), 1621–1629.

Plackett, R. L. (1975). The Analysis of Permutations. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 24, 193–202.

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using google trends. *Scientific reports*, 3, srep01684.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336), 846–850.

Richmond, S., Shaw, W. C., O'brien, K. D., Buchanan, I. B., Jones, R., Stephens, C. D., Roberts, C. T., & Andrews, M. (1992). The development of the par index (peer assessment rating): reliability and validity. *The European Journal of Orthodontics*, 14, 125–139.

Ročková, V. & George, E. I. (2016a). Fast bayesian factor analysis via automatic rotations to sparsity. *Journal of the American Statistical Association*, 111(516), 1608–1622.

Ročková, V. & George, E. I. (2016b). The spike-and-slab lasso. *Journal of the American Statistical Association*, (just-accepted).

Rubinstein, R. Y. & Kroese, D. P. (2004). *The cross-entropy method: a unified approach to combinatorial optimization, Monte-Carlo simulation and machine learning*. Springer Science & Business Media.

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2), 461–464.

Scott, S. L. & Varian, H. R. (2013). *Bayesian variable selection for nowcasting economic time series*. Technical report, National Bureau of Economic Research.

Scott, S. L. & Varian, H. R. (2014). Predicting the present with bayesian structural time series. *International Journal of Mathematical Modelling and Numerical Optimisation*, 5(1-2), 4–23.

Siegel, E. (2016). *Predictive analytics: The power to predict who will click, buy, lie, or die*. Wiley Hoboken (NJ).

Song, G.-Y., Jiang, R.-P., Zhang, X.-Y., Liu, S.-Q., Yu, X.-N., Chen, Q., Weng, X.-R., Wu, W.-Z., Su, H., Ren, C., Shan, R.-K., Geng, Z., Xu, T.-M., & Research Group of Establishing Chinese Evaluation Standard of Orthodontic Treatment Outcome (2015). Validation of subjective and objective evaluation methods for orthodontic treatment outcome. *Journal of Peking University. Health sciences*, 47, 90–97.

Song, G.-Y., Zhao, Z.-H., Ding, Y., Bai, Y.-X., Wang, L., He, H., Shen, G., Li, W.-R., Baumrind, S., Geng, Z., et al. (2014). Reliability assessment and correlation analysis of evaluating orthodontic treatment outcome in chinese patients. *International journal of oral science*, 6(1), 50–55.

Stock, J. H. & Watson, M. W. (2002). Forecasting using principal components from a large number of predictors. *Journal of the American statistical association*, 97(460), 1167–1179.

Sun, Z., Liu, X., & Wang, L. (2017). A hybrid segmentation method for multivariate time series based on the dynamic factor model. *Stochastic Environmental Research and Risk Assessment*, 31(6), 1291–1304.

Tanner, M. A. & Wong, W. H. (1987). The calculation of posterior distributions by data augmentation. *Journal of the American statistical Association*, 82, 528–540.

Taylor, J. W. (2010). Exponentially weighted methods for forecasting intraday time series with multiple seasonal cycles. *International Journal of Forecasting*, 26(4), 627–646.

Thurstone, L. L. (1927). A law of comparative judgment. *Psychological review*, 34, 273.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)*, (pp. 267–288).

Tse, Y. K. & Tsui, A. K. C. (2002). A multivariate generalized autoregressive conditional heteroscedasticity model with time-varying correlations. *Journal of Business & Economic Statistics*, 20(3), 351–362.

Van der Vaart, A. W. (2000). *Asymptotic statistics*, volume 3. Cambridge university press.

Van Erp, M. & Schomaker, L. (2000). Variants of the borda count method for combining ranked classifier hypotheses. In *Proceedings of the 7th International Workshop on Frontiers in Handwriting Recognition*.

Winters, P. R. (1960). Forecasting sales by exponentially weighted moving averages. *Management science*, 6(3), 324–342.

Wu, L. & Brynjolfsson, E. (2015). The future of prediction: How google searches foreshadow housing prices and sales. In *Economic analysis of the digital economy* (pp. 89–118). University of Chicago Press.

Yang, S., Kou, S. C., Lu, F., Brownstein, J. S., Brooke, N., & Santillana, M. (2017). Advances in using internet searches to track dengue. *PLoS computational biology*, 13(7), e1005607.

Yang, S., Santillana, M., & Kou, S. (2015). Accurate estimation of influenza epidemics using google search data via argo. *Proceedings of the National Academy of Sciences*, 112(47), 14473–14478.

Yao, G. & Böckenholt, U. (1999). Bayesian estimation of thurstonian ranking models based on the gibbs sampler. *British Journal of Mathematical and Statistical Psychology*, 52, 79–92.

Yao, Y.-C. (1984). Estimation of a noisy discrete-time step function: Bayes and empirical bayes approaches. *The Annals of Statistics*, (pp. 1434–1447).

Yu, P. L. H. (2000). Bayesian analysis of order-statistics models for ranking data. *Psychometrika*, 65(3), 281–299.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476), 1418–1429.

Zou, H., Hastie, T., & Tibshirani, R. (2006). Sparse principal component analysis. *Journal of computational and graphical statistics*, 15(2), 265–286.