University of Illinois at Urbana-Champaign and Harvard University

Abstract. Data in the form of ranking lists are frequently encountered, and combining ranking results from different sources can potentially generate a better ranking list and help understand behaviors of the rankers. Of interest here are the rank data under the following settings: (i) covariate information available for the ranked entities; (ii) rankers of varying qualities or having different opinions; and (iii) incomplete ranking lists for non-overlapping subgroups. We review some key ideas built around the Thurstone model family by researchers in the past few decades and provide a unifying approach for Bayesian Analysis of Rank data with Covariates (BARC) and its extensions in handling heterogeneous rankers. With this Bayesian framework, we can study rankers' varying quality, cluster rankers' heterogeneous opinions, and measure the corresponding uncertainties. To enable an efficient Bayesian inference, we advocate a parameter-expanded Gibbs sampler to sample from the target posterior distribution. The posterior samples also result in a Bayesian aggregated ranking list, with credible intervals quantifying its uncertainty. We investigate and compare performances of the proposed methods and other rank aggregation methods in both simulation studies and two real-data examples.

Key words and phrases: Thurstone model, rank aggregation, heterogeneous rankers, infinite mixture model, parameter-expanded data augmentation.

Rank data are rather prevailing these days, and combining ranking results from different sources is a common problem. Well-known rank aggregation problems range from the election problem back in the 18th century (Borda, 1781) to search engine results aggregation in modern days (Dwork et al., 2001;

Xinran Li, Department of Statistics, University of Illinois, Champaign, IL 61820 (e-mail: xinranli@illinois.edu). Dingdong Yi, Department of Statistics, Harvard University, Cambridge, MA 02138 (e-mail: yidingdong@gmail.com). Jun S. Liu, Department of Statistics, Harvard University, Cambridge, MA 02138 (e-mail: jliu@stat.harvard.edu).

Liu, 2009). In many cases, there are variations and complications associated with rank data. Sometimes, there are relevant covariates of the ranked entities while the ranking lists are highly incomplete. Also, the rankers are likely heterogeneous. Here, we illustrate the problem in detail using the following two examples.

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 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, most websites aggregate them using arithmetic means. Besides rankings, some summary statistics of the NFL players are also available online. For example, Table 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

Ranking lists of NFL starting quarterbacks from 13 different experts, as of week 12 in the 2014 season. The first column shows the players' names, and the remaining columns show the ranked positions of these players from the 13 experts.

Player	1	2	3	4	5	6	7	8	9	10	11	12	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: fantasy.nfl.com/research/rankings, www.fantasypros.com/nfl/rankings/qb.php.

In Example 1, according to Table 1, most experts give very similar ranking lists, with a few exceptions such as experts 1 and 5. Besides, some rankers do not place the players in Table 1 on their top 24 lists, making the ranking lists incomplete. Therefore, it is of interest to understand how the rankings may be dependent of the available summary statistics (i.e., covariates), whether the

Table 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: www.nfl.com/stats.

experts (i.e., rankers) are consistent in using these covariates when ranking the players, and whether they (the rankers) have different qualities or different opinions. Another goal is to obtain an aggregated ranking list of all players taking into account the covariate information of the players and the potential heterogeneity of the experts, which hopefully can improve the accuracy of rank aggregation compared to the simple arithmetic means.

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 the experts, cases were divided into 9 groups, each containing 12 cases. For each group of the cases, each expert evaluated the conditions of all 12 cases and provided a within-group ranking list, mostly based on their personal experiences and judgments of the cases' 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 3 shows 15 of the 69 ranking lists for two groups, and Table 4 shows the corresponding features for these two groups.

Understanding how each orthodontist used the available covariates to arrive at his/her rank list and how to form a consensus ranking are important issues

TABLE 3

Ranking lists for Groups A and H, two of the 9 groups in Example 2. The first column shows the groups and indices for the cases, and the remaining columns show the within-group ranked positions of these cases from 15 experts.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	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

in this example, 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 differs from Example 1 and prevailing rank aggregation applications in that it contains many "local" rankings among non-overlapping subgroups. Having been demonstrated to be associated with ranking outcomes by Song et al. (2015), the covariate information can not only help generating a consensus full ranking list, but also potentially reveal inhomogeneity among these experts in their ranking "qualities" as well as their way of using the covariates, (Liu et al., 2012; Song et al., 2014). As shown in Table 3 and our later analysis, there are clearly heterogeneous qualities or opinions among rankers. For example, the ranking position of case A9 from the listed 15 experts in Table 3 ranges from 2 to 12.

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. Van Erp and 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 and 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,

TABLE 4

Eleven 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.

	Urs	Uas	Uls	Lrs	Las	Lls	Rbo	Lbo	Oj	Ob	Cl
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

1957). Recently, Vitelli et al. (2017) proposed an efficient MCMC approach to conduct Bayesian inference for the Mallows model and its extension allowing mixture heterogeneous subgroups of rankers, which naturally provides uncertainty quantification for the resulting quantities of interest. Li et al. (2020) formulated a different extension of the Mallows model and provided both new theoretical results and an EM algorithm for the inference. To overcome the computational difficulty and relax model assumptions, Švendová and Schimek (2017) proposed an indirect inference approach that tries to minimize the difference between the empirical distribution functions of entities' ranks and the corresponding true ones, and used non-parametric bootstrap to quantify uncertainty.

The second type builds statistical models to characterize the data generating process of the rank data and uses the estimated models to generate the aggregated ranking list (Block and Marschak, 1960; McFadden, 1980; Diaconis, 1988; Critchlow et al., 1991; Marden, 1996; Alvo and Yu, 2014). The most popular model for rank data is the Thurstone order statistics model, which includes the Thurstone–Mosteller–Daniels (TMD) model (Thurstone, 1927; Mosteller, 1951; Daniels, 1950) and Plackett–Luce (PL) model (Bradley and Terry, 1952; Luce, 1959; Plackett, 1975) as special cases. Together with variants and extensions (Stern, 1990; Böckenholt, 1992; Walker and Ben-Akiva, 2002), the Thurston model family has been successfully applied to a wide range of problems (e.g., Gormley and 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 versions determine the rankings. In the

TMD and PL models, the noises are assumed to follow the normal and Gumbel distributions, respectively. The PL model can be equivalently viewed as a multi-stage 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. A closely related literature to the development of rank data analysis is the analysis of pairwise comparison data; see Bradley and Terry (1952) and David (1963) for earlier development, Davidson and Farquhar (1976) and Hastie and Tibshirani (1998) for applications, Luce (1959); Rao and Kupper (1967); Plackett (1975); Agresti (1990) and Huang et al. (2006) for various extensions, and Hunter (2004); Guiver and Snelson (2009); Gormley and Murphy (2009) and Caron and Doucet (2012) for efficient Bayesian computation.

Challenges arise in the analysis of ranking data when (a) rankers are of different qualities or belong to groups with different opinions; (b) covariate information are available for either the rankers or the ranked entities or both; and (c) there are incomplete ranking lists. Gormley and Murphy (2006, 2008b,a, 2010) developed the finite mixture of PL 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) introduced the finite mixture of Thurstone models to allow for heterogeneous subgroups of rankers, with limited explorations; Yu (2000) attempted to incorporate the covariate information for both ranked entities and rankers; Johnson et al. (2002) examined qualities of several known subgroups of rankers; and Lee et al. (2014) represented qualities of rankers by letting them have different noise levels. See Böckenholt (2006) for a review of developments in Thurstonian-based analysis with some further extensions. In the presence of incomplete ranking lists, Ailon (2010); Xia and Conitzer (2011); Meila and Chen (2012) and Liu et al. (2019) studied rank aggregation under various model assumptions. Recently, Deng et al. (2014) proposed a Bayesian approach that can distinguish high-quality rankers from low-quality ones, and Bhowmik and Ghosh (2017) proposed a method that utilizes covariates of ranked entities to assess qualities of all rankers. See also Badgeley et al. (2014); Li et al. (2017, 2018) for rank aggregation with application to genomic studies.

Although various Thurstonian models have been proposed in the past, their inferences are mostly based on the maximum likelihood approach and the EM algorithm (with a few exceptions). The way of quantifying uncertainties and dealing with incomplete information has been limited. In this paper, we propose a unified framework built upon the classic Thurstone model family to deal with incomplete ranking lists, to accommodate rankers with different qualities or opinions, and to incorporate covariate information of ranked entities. 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 addition to providing a full Bayesian inference procedure for parameter estimation of the proposed models, we also pay special attention to rank aggregation and the uncertainty evaluation of the resulting aggregated ranking lists.

In this paper, we mainly focus on the TMD model and its extension to accommodate various complications. The Thurstone-type model is probably one of the most natural data generating model for rank data. It contains a rich family of statistical models including both the TMD model and the PL model, and is widely used in practice. We focus on the TMD model with Gaussian errors since parametric Gaussian regression models are more intuitive and interpretable, are easier to manipulate in terms of model development, and have rich literature support. Similar extensions can be made for other Thurstone models as well. The estimation for the TMD 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 one involving mean and covariance structures with dichotomous indicators, Yao and Böckenholt (1999) proposed a Bayesian approach based on Gibbs sampler, and Johnson (2013) advocated the JAGS software to implement the Bayesian posterior sampling. Our new model is even more challenging than the classic Thurstone family of models because of its inclusion of new components for dealing with heterogeneous rankers. We design an efficient parameter-expanded Gibbs sampler algorithm (Liu and Wu, 1999), which facilitates group moves of the latent variables and greatly improves the computational efficiency.

Our extension of the TMD model is similar in spirit to Gormley and Murphy (2006, 2008b,a, 2010)'s extension of the PL model, but we allow infinite mixture components using the Dirichlet process prior. Moreover, unlike the PL model, the TMD model does not have a closed-form likelihood, and thus impose additional computational challenges. The rest of this article is organized as follows. Sections 2 and 3 elaborate on our Bayesian models for rank data with covariates. Section 4 provides details of our Markov Chain Monte Carlo (MCMC) algorithms. Section 5 introduces multiple analysis tools using MCMC samples. Section 6 displays simulation results to validate our approaches. Section 7 describes the two real-data applications using the proposed methods. Section 8 concludes with a short discussion.

Let $\mathcal{U}$ $\{1, 2, \ldots, N\}$ be the set of all entities in consideration, and 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* is a set of non-contradictory pairwise relations in $\mathcal{U}$, which gives rise to an ordered preference list for entities in $\mathcal{U}$. We call a *full ranking list* if identifies all pairwise relations in $\mathcal{U}$, otherwise a *partial ranking list*. When is a full ranking list, we can equivalently write as $i_1 \succ i_2 \succ \ldots \succ i_N$ for notational simplicity, and further define i as the *ranked position* of an entity $i \in \mathcal{U}$. Specifically, a higher ranked entity has a smaller numbered position in the list, i.e. i_1 $<$ i_2 if and only if $i_1 \succ i_2$. For example, Tables 1 and 3 show the ranked positions of the entities in each ranking list. Furthermore, for any vector $z_1, \ldots, z_N$ $^\top \in \mathbb{R}^N$, we use rank $i_1 \succ i_2 \succ \ldots \succ i_N$ 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 $_i \in \mathbb{R}^L$ be the L dimensional covariate vector of ranked entity i, and $\quad _1, \ _2, \ldots, \ _N \ ^\top \in \mathbb{R}^{N\times L}$ be the covariate matrix for all N entities. In the remaining discussion, for clarification, we use index $1 \leq i \leq N$ for ranked entities and index $1 \leq j \leq M$ for rankers, with N and M denoting the total numbers of ranked entities and rankers, respectively.

Suppose we have M full ranking lists $_1, \ _2, \ldots, \ _M$ for entities in $\mathcal{U} \quad \{1,2, \ldots, N\}$. Thurstone (1927) postulated that the ranking outcome $_j$ 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 $_j \quad Z_{1j}, \ldots, Z_{Nj} \ ^\top$ as ranker j's evaluations of all entities, and rank $_j$ as the associated full ranking list based on $_j$. Under the TMD model, $_j$ follows a multivariate Gaussian distribution with mean $\quad _1, \ldots, \ _N \ '$ representing the underlying true score of the ranked entities:

$$\text{(2.1)} \qquad \begin{aligned} Z_{ij} \quad _i \quad _{ij}, \quad _{ij} &\sim N\ 0, \ ^2 \qquad && 1 \leq i \leq N; 1 \leq j \leq M \\ _j \quad \text{rank} \quad _j\ , & && 1 \leq j \leq M \end{aligned}$$

where $_{ij}$'s are mutually independently across all i and j. Because we only observe the ranking lists $_j$, multiplying $\ ,$ by a constant or adding a constant to all the $_i$'s does not affect the likelihood function. Therefore, to ensure identifiability of the parameters, we fix $^2 \quad 1$ and impose the constraint that lies in the space $\boldsymbol{\Theta} \quad \{ \ \in \mathbb{R}^n : \mathbf{1}_N^\top \quad 0\}$, where $\mathbf{1}_N$ is an N dimensional vector with all coordinates being 1.

Model (2.1) implies that the $_j$'s are independent and identically distributed (i.i.d.) conditional on . The likelihood function is then $p \ _1, \ _2, \ldots, \ _M \mid$ $\prod_{j \ 1}^M p \ _j \mid$ with

$$\text{(2.2)} \qquad \begin{aligned} p \ _j \mid \quad & \int_{\mathbb{R}^N} p \ _j \mid \ _j, \quad p \ _j \mid \quad \mathrm{d} \ _j \\ & \int_{\mathbb{R}^N} 1_{\{\text{rank}\ \boldsymbol{Z}_j \quad _j\}} \cdot \ 2 \quad ^{-N/2} e^{-\|\boldsymbol{Z}_j - \boldsymbol{\mu}\|^2/2} \mathrm{d} \ _j . \end{aligned}$$

The goals are to estimate the parameter and then generate an aggregated ranking based on the estimated . A common approach is the maximum likelihood method, which is computationally challenging due to the integral in the likelihood. Besides, it is also nontrivial to quantify the uncertainty of the resulting rank aggregation. We focus on the Bayesian approach, which is more convenient to incorporate prior information, to quantify estimation uncertainties, and to utilize efficient MCMC algorithms including data augmentation (Tanner and Wong, 1987) and parameter expansion strategies (Liu and Wu, 1999). With a reasonable prior on the $_i$'s, we can get the corresponding posterior means of $_i$'s, based on which we can generate an aggregated ranking list.

Recalling that is restricted to the space $\boldsymbol{\Theta}$, we define $_N \quad _N - N^{-1}\mathbf{1}_N\mathbf{1}_N^\top$ as the projection matrix that maps any vector in $\mathbb{R}^N$ to $\boldsymbol{\Theta}$, where $_N$ is an $N \times N$ identity matrix and $\mathbf{1}_N$ is an N dimensional vector with all elements being 1. We choose the prior of to be $\mathcal{N}\left(\mathbf{0}, \ ^2 \ _N\right)$. The intuition for choosing this prior is that when $\sim \mathcal{N} \ \mathbf{0}, \ ^2 \ _N$, we have $_N \quad \in \boldsymbol{\Theta}$ and $_N \quad \sim$

$\mathcal{N}\ \mathbf{0},\ {}^{2}\ \ {}_{N}$. For computational convenience, it is equivalent to using the prior $\sim \mathcal{N}\ \mathbf{0},\ {}^{2}\ \ {}_{N}$ and considering the posterior mean of ${}_{N} \equiv \ \ - \bar{}\mathbf{1}_{N}$, where $\bar{}\ \ n^{-1}\sum_{i}^{n}\ {}_{1}\ \ {}_{i}$. In other words, $f\ {}_{1}\ \ \mid\ {}_{1}, \cdots,\ {}_{M}\ \ \ f\ {}_{2}\ \ - \bar{}\mathbf{1}_{N} \mid$ ${}_{1}, \cdots,\ {}_{M}$, where ${}_{1}\ \ \sim \mathcal{N}\left(\mathbf{0},\ {}^{2}\ \ {}_{N}\right)$ and ${}_{2}\ \ \sim \mathcal{N}\ \mathbf{0},\ {}^{2}\ \ {}_{N}$ denote the priors of . More generally, although we restrict to the parameter space $\mathbf{\Theta}$, we only need to specify a prior for the unconstrained and make inference based on the posterior distribution of $\ - \bar{}\mathbf{1}_{N}$. Therefore, in the following discussion, we relax the constraint that $\in \mathbf{\Theta}$, and emphasize that what matters are the relative values of the ${}_{i}$'s.

As in both examples, each ranked entity is associated with some covariates that may be relevant to how the entity is perceived and thus ranked by rankers. To incorporate the covariate information into model (2.1), we assume that the score of each entity i depends linearly on the L-dimensional covariate vector ${}_{i}$, for $i\ \ 1, \ldots, N$. To avoid being too restrictive, we allow the intercept term for each entity to be different. In particular, we have the following over-parameterized model:

$$\begin{array}{lr} {}_{i}\quad {}_{i}\quad {}_{i}^{\top}\ , & 1 \le i \le n \\ Z_{ij}\quad {}_{i}\quad {}_{ij},\quad {}_{ij} \sim \mathcal{N}\ 0, 1\ , & 1 \le i \le N; 1 \le j \le M \\ {}_{j}\quad \text{rank}\quad {}_{j}\ , & 1 \le j \le M \end{array} \tag{2.3}$$

where the ${}_{ij}$'s are jointly independent across all i and j.

Model (2.3) is over-parameterized because is invariant if we add a constant vector to and change ${}_{i}$ to ${}_{i} - \ {}_{i}^{\top}$. However, the structure between and , can help us construct some informative priors on , incorporating the covariate information. Intuitively, entities with similar ${}_{i}$'s should be close in the underlying ${}_{i}$'s. Such intuition is conformed by model (2.3) with suitable priors on , , because similar entities will have higher correlation among their ${}_{i}$'s *a priori*. Model (2.3) can be helpful when the ranking information is weak and incomplete, and the covariate information is strongly related to the ranking mechanism. More generally, some covariates of the rankers may also be available, and they can be similarly incorporated into the ${}_{i}$'s in (2.3). For example, with covariates ${}_{j}$ for each ranker j, we can model the evaluation score of ranker j for entity i as $Z_{ij}\quad {}_{i}\quad {}_{i}^{\top}\quad {}_{j}^{\top}\ {}_{i}\quad {}_{ij}$, similar to Yu (2000) and Gormley and Murphy (2010). In this paper, we focus only on the covariates of the ranked entities mainly due to our applications.

We further illustrate model (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 ${}_{i}\quad {}_{i}^{\top}$ quantifies a hypothetically universal underlying "score" of the quarterback, and each ranker evaluates it with a personal variation modeled by ${}_{ij}$. The linear term ${}_{i}^{\top}$ can explain 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 ${}_{i}$ can capture the remaining "random effect". Without ${}_{i}$, model (2.3) reduces to a rank regression model in Johnson (2013), which can be too restrictive in some applications.

We set the prior $\ ,\ \equiv\ $, where is $\mathcal{N}\ 0,\ ^{2}\ _{N}$ and p is $\mathcal{N}\ 0,\ ^{2}\ _{L}$, where $_{N}$ and $_{L}$ are $N \times N$ and $L \times L$ identity matrices, respectively. The hyper-parameter and can reflect our prior belief on the relevance of covariate information to ranking mechanism. Intuitively, the stronger the belief on the role of covariates, the smaller the ratio $^{2}/\ ^{2}$ should be chosen. The Bayesian procedure based on the generalized regression model (2.3) is henceforth referred to as Bayesian Analysis of Rank data with Covariates (BARC).

The PL model is another popular Thurstone model for rank data, and it differs from the TMD model described in Sections 2.2 and 2.3 in the distributional assumption for the noises $_{ij}$'s. In particular, the PL model assumes that all the noises $_{ij}$'s follow i.i.d. Gumbel distribution. Importantly, the probability of observing a rank list $_{j}$ under the parameter , which can be written as an integral as in (2.2), now has an equivalent closed-form expression:

$$p\ \ _{j} \mid \qquad \prod_{i\ \ 1}^{N} \frac{\exp\left(\ \ _{j\ i}\right)}{\sum_{i'\ \ i}^{N} \exp\left(\ \ _{j\ i'}\right)}. \tag{2.4}$$

Therefore, compared to the TMD model, the PL model requires less computational cost, because the likelihood for the observed rankings has a closed-form expression that does not involve any integral. Hunter (2004) proposed a minorization-maximization (MM) algorithm for finding the MLE of the parameters, Guiver and Snelson (2009) worked out a Bayesian approach based on a message-passing algorithm (Expectation-Propagation), Caron and Doucet (2012) proposed simple Gibbs samplers for Bayesian inference, and Azari Soufiani et al. (2013) proposed a class of generalized method-of-moments algorithms. The covariate information can be similarly included as in (2.3), see, e.g., Hausman and Ruud (1987); Allison and Christakis (1994).

The performance of the TMD model and the PL model depend on the nature of the data, and thus is generally case-by-case; see Azari Soufiani (2014, Chapter 2) for a simulation study. In this paper, we mainly focus on the TMD model, and the implementation can be helpful for general noise distributions that may not be easily integrated out.

In practice, the rankers in consideration may have different quality or reliability. In these cases, it is of interest to distinguish high-quality rankers from low-quality ones, and a weighted rank aggregation method is often preferred, where each ranker j has a weight w_j reflecting the quality of its ranking list. However, it is generally difficult to design a proper weighting scheme in practice, especially when little or no prior knowledge of the rankers is available. To deal with this difficulty, one can accommodate weighting through the variance parameters in the 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 (2.3) to the following weighted version:

$$
\begin{aligned}
& {}_{i} \qquad {}_{i} \qquad {}^{\top}_{i}\ , && 1 \le i \le N\\
& Z_{ij} \qquad {}_{i} \qquad {}_{ij}, \qquad {}_{ij} \sim N\ \ 0, w_j^{-1}\ \ , && 1 \le i \le N; 1 \le j \le M\\
& {}_{j} \qquad \text{rank} \qquad {}_{j}\ , && 1 \le j \le M
\end{aligned}
\tag{3.1}
$$

where $w_j > 0$ for all j and the ${}_{ij}$'s are mutually independent across all i and j.

The prior for the w_j's can be any distribution with support on positive real numbers, such as uniform and truncated chi-squared distributions. In the absence of covariates, Lee et al. (2014) considered model (3.1) and assumed that the $w_j^{1/2}$'s are i.i.d. uniform on $\ 0, 20\ $ *a priori*. Here we consider a more restrictive choice where the weights take on only a few values, which can lead to a less sticky MCMC sampler without compromising much precision in the rankers' quality evaluation and the aggregated ranking list. Specifically, we restrict w_j to three values, 2, 1 and 0.5, standing for reliable, mediocre, and low-quality rankers, respectively, with equal probabilities *a priori*, i.e.,

$$
P\ \ w_j \quad 0.5 \qquad P\ \ w_j \quad 1 \qquad P\ \ w_j \quad 2 \qquad 1/3, \qquad 1 \le j \le M
\tag{3.2}
$$

and assume the w_j's are mutually independent across all j. We call the resulting rank analysis method the Bayesian Analysis of Rank data with entities' Covariates and rankers' (unknown) Weights (BARCW, henceforth).

All previously described models assume that the underlying score is universal to all rankers, which can sometimes be too restrictive. Böckenholt (1993) and Gormley and Murphy (2006, 2008b,a) 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' reliability, this mixture model focuses on the heterogeneity in rankers' opinions while assuming that all rankers are equally reliable.

Böckenholt (1993) considered finite mixtures of TMD models. However, in finite mixture models, a common issue is how to determine the number of mixture components. Here we employ the Dirichlet process mixture model, which overcomes this issue by using mixture distributions with countably infinite number of components via a Dirichlet process prior (Antoniak, 1974; Ferguson, 1983). We first extend model (2.3) so that the underlying score of each entity is ranker-specific:

$$
\begin{aligned}
& {}^{j} \qquad {}^{j} \qquad {}^{j}\ , && 1 \le j \le M\\
& {}_{j} \qquad {}^{j} \qquad {}_{j}, \qquad {}_{j} \sim \mathcal{N}\ \ \mathbf{0},\ \ {}_{n}\ \ , && 1 \le j \le M\\
& {}_{j} \qquad \text{rank}\ (\ \ {}_{j})\ , && 1 \le j \le M
\end{aligned}
\tag{3.3}
$$

where $\ \in \mathbb{R}^{N \times L}$ is the covariate matrix for all ranked entities, ^{j} represents the underlying true score for ranker j, and ${}_{j}$'s are mutually independent. We then assume that the $\;^{j}, \;^{j}$ follow a distribution G that is drawn from a Dirichlet process, i.e.,

$$(3.4) \qquad \;^{j}, \;^{j} \mid G \overset{iid}{\sim} G, \quad G \sim DP \quad , G_0 \; ,$$

where G_0 defines a baseline distribution on $\mathbb{R}^{N \quad L}$, and is a concentration parameter. For the ease of understanding, we can equivalently view model (3.3)-(3.4) as the limit of the following finite mixture model with K components when $K \to \infty$:

$$\begin{aligned} {}^{\langle k\rangle}, \;{}^{\langle k\rangle} &\overset{i.i.d.}{\sim} G_0, && 1 \le k \le K\\ {}^{\langle k\rangle} \qquad {}^{\langle k\rangle} \qquad &{}^{\langle k\rangle}, && 1 \le k \le K\\ {}_1, \dots, \;{}_K &\sim \text{Dir} \quad /K, \dots, \quad /K \; , && \\ c_j \mid \quad &\overset{i.i.d.}{\sim} \text{Multinomial} \quad {}_1, \dots, \;{}_K \; , && 1 \le j \le M\\ {}_j \qquad {}^{\langle c_j\rangle} \qquad &{}_j, \quad {}_j \sim \mathcal{N} \;\; \mathbf{0}, \;\; {}_n \; , && 1 \le j \le M\\ {}_j \qquad &\text{rank} \left(\;\; {}_j \right), && 1 \le j \le M \end{aligned}$$

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

We choose the baseline distribution G_0 on $\mathbb{R}^{N \quad L}$ using two independent zero-mean Gaussian distributions with covariances $\;^{2} \;{}_N$ and $\;^{2} \;{}_L$, i.e., $G_0 \sim \mathcal{N} \;\; \mathbf{0}$, diag $\;^{2} \;{}_N, \;^{2} \;{}_L$. Section 4.4 provides more details on how the prior for these parameters in the model might be set. Obviously, G_0 is the same as the prior distribution of , we use in the previous models, and the conjugacy between G_0 and the distribution of $\;{}_j$'s leads to a straightforward Gibbs sampler (Neal, 1992; Liu, 1994; MacEachern, 1994). Parameter represents the degree of concentration of G around G_0 and is related to the number of distinct clusters. According to the Pólya urn scheme representation of the Dirichlet process (Blackwell and MacQueen, 1973), the expected number of clusters with in total M rankers is $\sum_{j \; 1}^{M} \quad / \; j \qquad - 1$ *a priori*. We discuss the sensitivity of these hyper-parameters in the simulation studies.

Under this Dirichlet process mixture model, we are interested in understanding the heterogeneous opinions among rankers and rank aggregation within each cluster as well as across all clusters. The aggregated ranking in each cluster is determined by those $\;^{j}$'s with identical values. The aggregated ranking list across all clusters depends on the underlying score of all rankers, i.e., $M^{-1} \sum_{j \; 1}^{M} \;\; ^{j}$. We regard this rank analysis method as BARCM, standing for Bayesian Analysis of Rank data with Covariates of entities and Mixture of rankers with different opinions. Furthermore, it is also straightforward to further incorporate varying weights for all rankers as in BARCW using model (3.1). To avoid being too lengthy, we skip the detailed description for the mixture model with varying weights, and simply denote it as BARCMW, standing for Bayesian Analysis of Rank data with Covariates of entities and Mixture of rankers with different opinions and Weights.

Models (2.1), (2.3), (3.1) and (3.3)-(3.4) can all be extended to cases where the observations are partial ranking lists. Because we define a 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 $_j$ $1 \leq j \leq M$, we also observe the $_j$'s that indicate which pairwise relationship is missing. Under the latent variable models, we denote $_j \simeq$ rank $_j$ if the partial ranking list $_j$ is consistent with the full ranking list rank $_j$. Our models, BARC, BARCW and BARCM, for the observed individual partial ranking lists are the same as in (2.1), (2.3), (3.1) and (3.3)-(3.4), except that $_j$ rank $_j$ is replaced by $_j \simeq$ rank $_j$. Let and denote the parameters for missing indicators $_j$'s and ranking lists $_j$'s, respectively. We can then write the likelihood of $_j,$ $_j$ as

$$p \quad _j, \; _j \,| \quad , \quad , \qquad \sum_{r:r\simeq \; _j} \int_{\mathbb{R}^N} p \quad _j \,|\, r, \quad _j, \quad , \quad 1_{\{r \quad \text{rank} \; \boldsymbol{z}_j \;\}} p \quad _j \,| \quad , \quad \mathrm{d} \quad _j.$$

If the pairwise relations are missing at random, in the sense that $p \quad _j \,|\, r, \quad _j, \quad ,$ $p \quad _j \,|\, \tilde{r}, \;\tilde{}\; _j, \quad ,$ for all possible $r,$ $_j, \tilde{r},$ $\tilde{}\;_j$ such that r rank $_j \simeq$ $_j$ and $\tilde{r}$ rank $\tilde{}\;_j \simeq$ $_j$, then the likelihood of $_j,$ $_j$ can be simplified as

$$p \quad _j, \; _j \,| \quad , \quad , \qquad p \quad _j \,| \quad _j, \quad , \qquad \int_{\mathbb{R}^N} 1_{\{ \; _j \simeq \text{rank} \; \boldsymbol{z}_j \;\}} p \quad _j \,| \quad , \quad \mathrm{d} \quad _j$$

If further the priors for the parameters and are mutually independent, we can ignore the $_j$'s when conducting Bayesian inference for the parameter of ranking mechanisms.

Here we give two additional remarks. First, we consider a special type of partial list, the top-K list, from which we can observed only the top K entities in a ranking list; see e.g., Schimek et al. (2015) for more detailed discussion. When K is fixed, it is not difficult to see that the corresponding pairwise comparison induced from a top-K list is missing at random. Second, we consider rank data containing ties. Generally, under the Thurstone-type model with continuous errors, the ranking list will have ties with zero probability. Practically, to mitigate this issue, we may view observed ranking lists with ties as partial ranking lists. More explicitly, we may treat ties as missing pairwise comparisons. However, such missing is not at random, implying that treating the information contained in "regarding the two entities as a tie" the same as "providing no comparisons between the two entities" may incur a little information loss, though it may not be of any practical importance.

The discussion above mainly focuses on the extension of the TMD model with normal noise. Similar extension can also be made for the PL model with Gumbel noise. Gormley and Murphy (2006) extended the PL model to allow a finite mixture of PL models. They further extended the model to mixture of experts, allowing the dependence of mixture probabilities on the ranker's covariates (Gormley and Murphy, 2008b), as well as the dependence of ranking mechanism on the ranker's covariates (Gormley and Murphy, 2010). A unique feature of the PL model is that it can be viewed as a multistage model as in (2.4). Benter (1994) extended this multistage model by allowing the probability ratios of being top among remaining entities to vary across different stages. Intuitively, under Benter's model, the choice of the top entity becomes increasingly random along the stages, and an entity with a smaller $_{i}$ in (2.4) has a greater probability to be ranked at a higher position. Gormley and Murphy (2008b,a) also studied the extension of Benter's model to mixture model and mixture of experts. For these finite mixture models, the number of mixture components are selected based on some information criteria such as BIC. Furthermore, Gormley and Murphy (2010) studied the choice of these models, including mixture model and mixture of experts, also using some model selection criteria, and illustrates how the ranker's covariate information should be incorporated into the PL model in practice. Recently, Liu et al. (2019) studied the PL mixture model for partial ranking lists, and proposed MCMC-based computation tools.

Our discussion for the TMD model involves only the covariates of the ranked entities, mostly due to our application. The covariate information of the rankers, if available, can also be incorporated similarly as in Gormley and Murphy (2010). Again, we focus mainly on the computation for the TMD model, which can be useful for general noise distributions that are difficult to integrate out.

We advocate the use of Gibbs sampler with parameter expansion (Liu and Wu, 1999) for Bayesian inference with general latent variable models, and in particular the class of TMD-based models we introduced in the previous sections. The parameter-expansion idea has been applied to ordered data and rank data analysis in previous studies (e.g., Liu and Wu, 1999; Hoff, 2009; Fong et al., 2016). Here we provide a unified parameter-expanded Gibbs sampling algorithm for the TMD model with rankers of varying qualities or heterogeneous opinions. We start with model (2.3) and then generalize this MCMC strategy to the extended models (3.1) and (3.3)-(3.4). We also provide an R package for implementing the proposed Bayesian analysis, with detailed information relegated to the Supplementary Material. For notational convenience, we define $_{1}, \ldots, _{M} \in \mathbb{R}^{N\times M}$, $\boldsymbol{\mathcal{T}}$ $\{ _{j}\}_{j}^{M}{}_{1}$, $_{N}$, $\in \mathbb{R}^{N\times N L}$, and $\boldsymbol{\Lambda}$ diag $^{2}{}_{N}, ^{2}{}_{L} \in \mathbb{R}^{N L\times N L}$.

The most computationally expensive part in our model is to sample all the Z_{ij}'s from the truncated Gaussian distributions. Moreover, because and , are intertwined together due to the posited regression model, they tend to correlate highly, similar to the difficulty in the data augmentation method for probit regression models (Albert and Chib, 1993).

To speed up the convergence of the MCMC algorithm, we follow Scheme 2 in Liu and 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 , which contains the evaluation scores of all rankers for all entities, indexed by a positive parameter , i.e., $t \equiv /$. For $1 \leq i \leq N$ and $1 \leq j \leq M$, let $_{-j}$ denote the evaluation scores of all rankers except ranker j, and $_{-i,j}$ denote ranker j's evaluation scores of all entities except entity i. The

PX-DA algorithm updates the missing data and the expanded parameters $, \ ,$ iteratively as follows:

(i) For $j \quad 1, \ldots, M$ and $i \quad 1, \ldots, N$, draw $[Z_{ij} \mid \quad _{-i,j}, \quad _{-j}, \quad , \quad]$ from truncated $\mathcal{N} \quad _i \quad _i' \ , 1$, where the truncation points are determined by $\quad _{-i,j}$ and $\quad _j$ such that rank $\quad _j \ \simeq \ _j$.

(ii) Draw $\quad \sim p \quad | \quad , \mathcal{T} \ \propto p \ t \quad |J \quad |H \ d$, and then update to be t . Here, $|J \quad | \quad ^{-NM}$ is the Jacobian of scale transformation, $H \ d \quad ^{-1}d$ is the Haar measure on a scale group up to a constant, and

$$p \ t \quad \propto \int p \ t \quad | \ , \quad p \quad p \quad \mathrm{d} \ \mathrm{d} \ \propto \exp\left\{-\frac{S}{2 \ ^2}\right\},$$

is the marginal density of latent variables evaluated at t , where

$$S \quad \sum_{j \ 1}^{M} \ _j^{\top} \ _j - \sum_{j \ 1}^{M} \sum_{j' \ 1}^{M} \ _j^{\top} \quad \boldsymbol{\Lambda}^{-1} \quad M \ ^{\top} \quad ^{-1} \ ^{\top} \quad _{j'}.$$

We can derive that $\ ^2 \sim S / \ ^2_{NM}$.

(iii) Draw $\quad , \quad \sim p \quad , \quad | \quad \sim \mathcal{N} \quad , \boldsymbol{\Sigma}$, where

$$\boldsymbol{\Lambda}^{-1} \quad M \ ^{\top} \quad ^{-1} \ ^{\top} \sum_{j \ 1}^{M} \ _j \quad \text{and} \quad \boldsymbol{\Sigma} \quad \boldsymbol{\Lambda}^{-1} \quad M \ ^{\top} \quad ^{-1}.$$

Below we give some intuition on why the PX-DA algorithm improves efficiency. Without Step (ii), the algorithm reduces to the standard Gibbs sampler, which updates the missing data and parameters iteratively. The scale group move of under the usual Gibbs sampler is slow due to both the Gibbs update for in Step (i) and the high correlation between and $, \ $. To overcome such difficulty, the PX-DA algorithm introduces a scale transformation of to facilitate its group move based on its marginal conditional distribution with $, \ $ integrated out. Thus, together with Step (iii), PX-DA effectively achieves the conditional sampling of $, \ $ and a scale group move of jointly. To ensure the validity of the MCMC algorithm, the scale transformation parameter has to be drawn from a carefully specified distribution, such that the move is invariant under the target posterior distribution, i.e., t follows the same distribution as the original under stationarity. To aid in understanding, we provide a proof in the Supplementary Material that the specified distribution of in Step (ii) satisfies this property.

Under model (3.1) for BARCW, the Gibbs step for $\quad , \ , \quad | \mathcal{T}, \quad$ is very similar to that for $\quad , \ , \quad | \mathcal{T}$ under model (2.3) for BARC, with details relegated to the Supplementary Material. The additional step is to draw w_j given all other variables. For $j \quad 1, \ldots, M$, let $\quad _{-j}$ be the weights associated with all rankers except ranker j. We draw discrete random variable w_j from the following conditional posterior probability mass function:

$$p \ w_j \mid \quad , \ _{-j}, \ , \ , \mathcal{T} \ \propto p \ w_j \ p \quad | \ , \ , \quad \propto w_j^{\frac{N}{2}} e^{-w_j \|\boldsymbol{Z}_j - \boldsymbol{\alpha} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/2}.$$

Under model (3.3)-(3.4), we first represent the parameters $\{\ _j,\ _j\}_{j\ 1}^{M}$ by a cluster allocation vector $\ c_1,\ldots,c_M\ $ and a set of cluster-wise parameters $\{\ ^{\langle k\rangle},\ ^{\langle k\rangle} : k \in \{c_1,\ldots,c_M\}\}$, and then use an MCMC algorithm to sample $\ ,\ ^{\langle k\rangle},\ ^{\langle k\rangle}$'s and $\ _1,\ldots,\ _M\ .$

We introduce $\mathcal{R}_k\ \ \{m : c_m\ \ k, 1 \leq m \leq M\}$ to denote the set of rankers that belong to cluster k given cluster allocation . Similarly, let $_{-j}$ be the subvector of excluding the jth element, and $\mathcal{R}_k\ \ _{-j}\ \ \{m : c_m\ \ k, m\ /\ j, 1 \leq m \leq M\}$ be the set of rankers except j that belong to cluster k. Due to the conjugacy between G_0 and the distribution of $\ _j$'s, we can integrate out $\ ^{\langle k\rangle},\ ^{\langle k\rangle}$'s when sampling , and the Gibbs sampling of given follows from Algorithm 3 in Neal (2000). Specifically, the Gibbs steps are as follows:

(1) For $j\ \ 1,\ldots,M$, draw c_j from

$$
\begin{aligned}
&P\left(c_j\ \ k \mid\ ,\ _{-j}, \boldsymbol{\mathcal{T}}\right)\\
&\propto P\left(c_j\ \ k \mid\ _{-j}\right) \int p\left(\ _j \mid\ ^{\langle k\rangle},\ ^{\langle k\rangle}\right) p\left(\ ^{\langle k\rangle},\ ^{\langle k\rangle} \mid\ _{-j}\right) \mathrm{d}\ ^{\langle k\rangle} \mathrm{d}\ ^{\langle k\rangle}\\
&\propto P\left(c_j\ \ k \mid\ _{-j}\right) \cdot \exp\left\{-\frac{1}{2} h\left(\{j\} \cup \mathcal{R}_k\ \ _{-j}\ \right)\ \ \frac{1}{2} h\left(\mathcal{R}_k\ \ _{-j}\ \right)\right\},
\end{aligned}
$$

where $P\left(c_j \mid\ _{-j}\right)$ has the following form:

$$
\begin{aligned}
P\left(c_j\ \ k \mid\ _{-j}\right)\ \ &\frac{|\mathcal{R}_k\ \ _{-j}\ |}{M-1}, \qquad \text{if } k \in \{c_m : m\ /\ j\}\\
P\left(c_j \notin \{c_m : m\ /\ j\} \mid\ _{-j}\right)\ \ &\frac{\ }{M-1},
\end{aligned}
$$

and $h\ \cdot\ $ is defined as

$$
\begin{aligned}
h\ \mathcal{R}\ \ \ &\sum_{m\in\mathcal{R}}\ _m^{\top}\ _m - \sum_{m\in\mathcal{R}} \sum_{m'\in\mathcal{R}}\ _m^{\top}\ \left(\boldsymbol{\Lambda}^{-1}\ \ |\mathcal{R}|\ \ ^{\top}\ \right)^{-1}\ ^{\top}\ _{m'}\\
&\log\left|\boldsymbol{\Lambda}^{-1}\ \ |\mathcal{R}|\ \ ^{\top}\ \right|,
\end{aligned}
$$

with $|\cdot|$ denoting the cardinality of a set or the determinant of a matrix.

(2) For each $k \in \{c_1,\ldots,c_M\}$, we sample $\{\ _j\}_{j\in\mathcal{R}_k\ \boldsymbol{c}}\ ,\ ^{\langle k\rangle},\ ^{\langle k\rangle} \mid \boldsymbol{\mathcal{T}},$ using Gibbs sampling steps similar to that for $\ ,\ ,\ \mid \boldsymbol{\mathcal{T}}$ under the BARC model; see the Supplementary Material for details.

Below we discuss the choice of variance parameters $\ ^2$ and $\ ^2$ for our Bayesian model BARC in (2.3) and its extensions, as well as the concentration parameter for the Dirichlet process in the mixture model. For both variance parameters, we impose priors following scaled inverse chi-squared distributions with parameters $\ ^2,\ $ and $\ ^2,\ $, i.e., $\ \ ^2\ /\ ^2$ and $\ \ ^2\ /\ ^2$ follow chi squared distributions with degrees of freedom and , respectively. For the concentration parameter, we impose a Gamma prior with shape parameter $a\ $ and rate parameter $b\ $. The Gibbs update for $\ ^2$ and $\ ^2$ given and under BARC and

BARCW, as well as that given $^{\langle k\rangle}$'s and $^{\langle k\rangle}$'s under BARCM and BARCMW, are straightforward and still involve sampling from scaled inverse chi-squared distributions. The Gibbs update for involves a two-step sampling from Beta and mixture Gamma distributions as described in Escobar and West (1995). We relegate the computation details to the Supplementary Material. The choice of hyperparameters ^{2}, , ^{2}, and a , b are discussed in the simulation studies.

Following the Bayesian computation in the previous section, we can obtain MCMC samples from the posterior distribution of , under BARC or BARCW, and from the posterior distribution of ^{j} , ^{j} 's under BARCM. Based on the posterior samples, we can then obtain the following results from Bayesian inference.

Under BARC in (2.3) or BARCW in (3.1), we use the posterior means of $_i \equiv$ $_i$ $_i^\top$'s to generate the aggregated ranking list. Under BARCM in (3.3)–(3.4) and BARCMW, we use the posterior means of $M^{-1}\sum_{j\ 1}^{M}\ _i^j$ $M^{-1}\sum_k |\mathcal{R}_k\ | \cdot$ $_i^{\langle k\rangle}$ $_i^\top$ $^{\langle k\rangle}$'s to generate the aggregated ranking list.

Most existing rank aggregation methods seek only one aggregated rank, but ignore the uncertainty of the aggregation result. When we observe $i \succ j$ in a single aggregated ranking list, 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 $\{\ ^s\}_{s\ 1}^S$ from the posterior distribution p $|\ _1, \cdots,\ _M$. For each sample ^{s} , we calculate a ranking list ^{s} rank ^{s} . We use $^s\ i$ to denote the position of entity i in ranking list ^{s} , and define the $1-$ credible interval for entity i's rank as

$$_{\mathrm{L}}\ i\ ,\ _{\mathrm{U}}\ i \qquad \left[\ _{\overline{2}}\ i\ ,\ _{1-\overline{2}}\ i\ \right],$$

where $_{\overline{2}}\ i$ and $_{1-\overline{2}}\ i$ are the $_{\overline{2}}$th and $1-\ _{\overline{2}}$ th sample quantiles of $\{\ ^s\ i\ \}_{s\ 1}^S$. The construction of credible intervals for entities' ranks under BARCM is very similar, and thus omitted here.

In BARCW and BARCM, as well as their combination BARCMW, we aim to learn the heterogeneity in rankers and subsequently improve and better understand 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 $|\ \mathcal{T}$ 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 $|\ \mathcal{T}$ estimate ranker clusters with different opinions. The number of clusters is determined by the number of distinct values in cluster allocation . The opinion of rankers in cluster k can be aggregated by the posterior means of $_i^{\langle k\rangle}$ $_i^{\top}$ $^{\langle k\rangle}$'s. We compare both methods later in simulation and application.

As discussed in Section 2.3, the interpretation of and is difficult due to over-parameterization. However, noting that the $_i$'s are modeled as i.i.d Gaussian random variables with mean zero *a priori*, the posterior distribution of still provides some meaningful information about the role of covariates in the ranking mechanism. Intuitively, for each ranked entity i, $_i^{\prime}$ can be viewed as the part of the evaluation score $_i$ linearly explained by the covariates, and $_i$ as the corresponding residual. The sign and magnitude of the coefficient $_k$ 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.

To compare the BARC-based methods with other rank aggregation methods, we adopt the normalized *Kendall tau distance* (Kendall, 1938) between ranking lists, which calculates the percentage of pairwise disagreements between two ranking lists. To measure the clustering accuracy, we adopt the adjusted Rand index (Rand, 1971), which calculates the percentage of pairwise clustering decisions that are correct after adjusting for chance.

Recall that $\mathcal{U}$ is the set $\{1, \ldots, N\}$ of entities, and entity i has a true score $_i$. We generate i.i.d. covariate vectors $_i$ $x_{i1}, \ldots, x_{ip}$ $^{\top}$'s for the N ranked entities from the multivariate Normal distribution with mean 0 and covariance Cov x_{is}, x_{it} $^{|s-t|}$ for $1 \leq s, t \leq p$, and generate M full ranking lists $\{ \ _j\}_j^M \ 1$ via the following model:

$$(6.1) \qquad _j \quad \text{rank} \quad _j , \quad _j \overset{i.i.d.}{\sim} \mathcal{N} \quad , \ ^2 \ _N , \qquad 1 \leq j \leq M \ .$$

We consider three different ways to generate the underlying true score vector , depending on the role of covariates. 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. That is,

1. $_i$ $_i^{\top}$, where $3, 2, -1, -0.5$ $^{\top}$, L 4, and 0.2.
2. $_i$ $_i^{\top}$ $\| \ _i\|^2$, where $3, 2, 1$ $^{\top}$, L 3, and 0.5.
3. $_i$ $\| \ _i\|^2$, where L 4, and 0.5.

We then compare the performance of BARC with other rank aggregation methods under varying noise levels for in (6.1). Fixing N 50 and M 10, we tried five different values of (1, 5, 10, 20, 40). For each configuration, we generated 100 simulated datasets. We applied Borda Count, Markov-Chain based methods (MC$_1$, MC$_2$, MC$_3$) (Dwork et al., 2001), Cross Entropy Monte

Carlo based method (CEMC) (Lin and Ding, 2009), PL model, and our BARC model with or without covariates. Specifically, Borda Count uses the arithmetic means of the average ranking positions over all ranking lists, the Markov-Chain based methods are based on the stationary distributions of Markov chains whose transition matrices are constructed based on the ranking lists, the CEMC approach tries to find the ranking list minimizing the average distance from all ranking lists using a stochastic search method, and the PL model is as introduced in Section 2.4. A brief review of the aforementioned methods can be found in the Supplementary Material. When employing BARC and its extensions, we input standardized covariates, fix the degrees of freedom for scaled inverse chi-squared priors on both ^{2} and ^{2} to be 3, i.e., 3, and consider three choices for the scale parameters ^{2}, ^{2} : $1^2, 1^2$, $1^2, 10^2$ and $10^2, 10^2$, denoted as BARC_1, BARC_2, BARC_3 respectively. We also consider BARC model without involving any covariates (denoted as BAR) and choose hyperparameters ^{2} 1 and 3.

Table 5 shows the (scaled) Kendall tau distances between the true rank and the aggregated ranks produced by different methods, averaged over the 100 simulated datasets for each scenario and each noise level. Specifically, the 3rd column for the Borda Count shows the Kendall tau distances between the estimated and true rank lists, averaged over the 100 replications, which serves as the baseline. The remaining columns for other methods show the ratios of their average Kendall tau distances over the corresponding values for the Borda Count. First, from Table 5, our BARC models generally perform better than other competing methods, especially in Scenarios 1 and 2 when certain linear combination of covariates is relevant for ranking. Second, comparing the BARC models with and without covariates in the last four columns of Table 5, utilizing covariates can improve the precision of the rank aggregation when covariates are useful as in Scenarios 1 and 2, and provide little harm on precision when covariates are irrelevant as in Scenario 3. Third, comparing the BARC models with different prior specification in the last three columns of Table 5, we find that the results are robust to the choice of ^{2} but sensitive to the choice of ^{2}, which is related to the non-identifiability issue for and as discussed in Section 2.3. Based on Table 5, we suggest to choose ^{2} 1 and ^{2} 10^2, which can not only exploit the use of covariates when they are indeed relevant for ranking but also provide robust rank aggregation even when these covariates are irrelevant.

We then consider the role of covariates based on our BARC models. Figure 1 shows the box plots of the posterior means of the coefficients $_k$'s over all 100 simulated data sets for each of the three scenarios at noise level 5. The results from the BARC models with different prior specifications are very similar. Note that we fix the noise level of our BARC models at 1. It is expected that the absolute scale of our estimated coefficient may be different from the truth, even under Scenario 1 where the true score is indeed linear in the covariates. However, the relative magnitudes and the signs of the estimated coefficients are still informative in telling the importance of covariates for linearly explaining the ranking mechanisms, as demonstrated in Figure 1.

TABLE 5

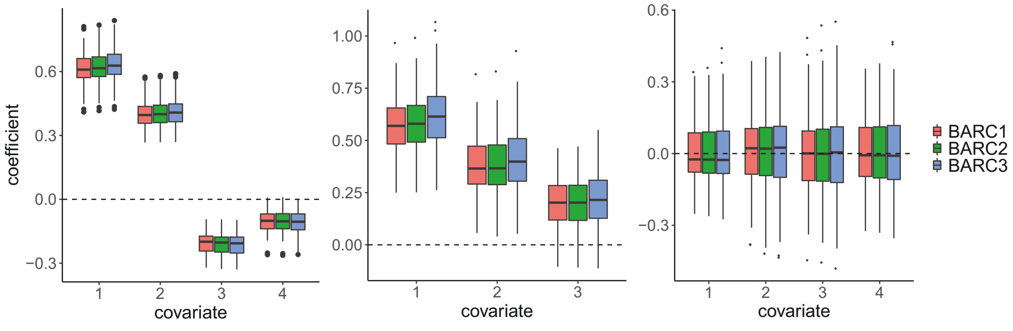
Comparison between BARC and other ranking methods. The first and second columns indicate the scenario and noise level for simulating the data. The 3rd column with parentheses shows the average Kendall tau distances between estimated ranking lists and corresponding true ones using Borda Count, and the remaining columns without parentheses show the corresponding values for different methods but standardized by the one using Borda Count. Specifically, MC_1–MC_3 denote the three forms of Markov-Chain based methods, CEMC denotes the Cross Entropy Monte Carlo based method (CEMC), PL denotes the Plackett–Luce model, BAR denotes our BARC mode without using any covariate, and $BARC_1$–$BARC_3$ denotes our BARC model using different priors.

Scenario	σ	Borda	MC_1	MC_2	MC_3	CEMC	PL	BAR	$BARC_1$	$BARC_2$	$BARC_3$
1	1	(0.026)	1.329	1.083	1.002	2.749	4.824	0.999	0.646	**0.643**	0.985
	5	(0.124)	1.285	1.018	1.009	1.145	1.098	0.985	0.653	**0.651**	0.972
	10	(0.216)	1.498	1.010	1.001	1.050	1.068	0.993	0.705	**0.704**	0.983
	20	(0.327)	1.475	1.004	0.999	1.024	1.037	0.996	0.795	**0.793**	0.990
	40	(0.408)	1.193	1.001	1.000	1.014	1.000	0.993	0.884	**0.883**	0.990
2	1	(0.028)	1.280	1.055	1.000	2.609	1.093	0.999	**0.981**	0.982	0.988
	5	(0.117)	1.313	1.016	1.006	1.120	1.089	0.989	**0.831**	0.832	0.977
	10	(0.200)	1.332	1.013	1.004	1.056	1.054	0.993	0.773	**0.772**	0.985
	20	(0.299)	1.406	1.011	1.000	1.018	1.042	0.994	0.813	**0.812**	0.988
	40	(0.388)	1.265	1.002	1.000	1.005	1.021	0.996	0.875	**0.874**	0.993
3	1	(0.046)	1.300	1.043	1.004	1.658	1.111	0.991	0.990	0.995	**0.988**
	5	(0.186)	1.376	1.011	1.001	1.054	1.071	**0.991**	1.007	1.007	0.992
	10	(0.270)	1.427	1.008	1.001	1.033	1.058	0.996	1.009	1.011	**0.996**
	20	(0.369)	1.274	1.005	1.001	1.023	1.022	0.996	1.002	1.002	**0.996**
	40	(0.421)	1.196	1.000	1.000	1.005	1.028	**0.997**	0.997	0.998	0.997

6.2 Computational gain from parameter expansion

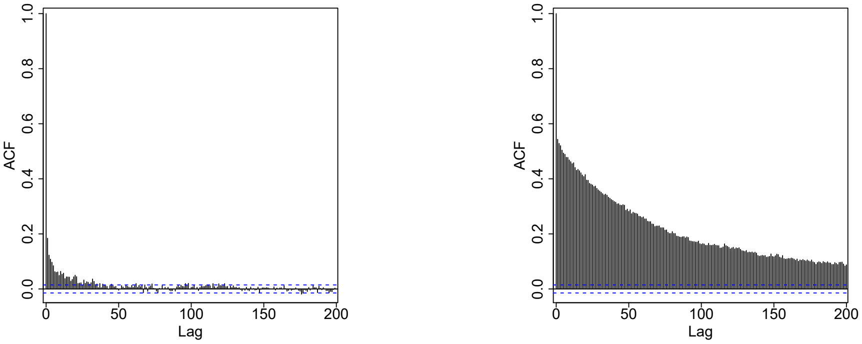
Before we move on to more complex settings, we use Scenario 2 with noise level $\sigma = 0.1$ to demonstrate the dramatic power of parameter expansion in dealing with rank data. In particular, we fit the BARC model with σ_α fixed at 1 and σ_β fixed at 10 *a priori*. From Figure 2, Gibbs sampler with parameter expansion reduces the auto-correlation in MCMC samples compared to regular Gibbs sampler. We also note that the parameter expansion step takes about the same amount of computational time as one conditional update of a simple Gibbs sampler, which is negligible.

Here we also briefly study the computational cost of BARC, and in particular



(a) Scenario 1 (b) Scenario 2 (c) Scenario 3

Fig 1: Box plots of the posterior samples of the coefficients β_k's for the standardized covariates in Scenarios 1–3 under BARC with different prior specifications.



(a) Parameter expanded

(b) Regular

Fig 2: Auto-correlation plots of MCMC samples for $_1$ in parameter expanded Gibbs sampler and regular Gibbs sampler.

Table 6

Run time (in seconds) of BARC with 100 Gibbs iterations for data from Scenario 2 under various values of N, M, L

Number of entities (N)	Numbers of rankers (M) and of covariates (L)								
		M 30			M 100			M 500	
	L 5	L 10	L 20	L 5	L 10	L 20	L 5	L 10	L 20
N 30	1.59	1.36	1.42	4.89	5.09	5.37	24.68	25.49	25.52
N 100	5.99	5.83	6.32	23.20	23.69	22.95	105.52	107.96	107.73
N 500	96.40	127.03	143.21	282.68	283.32	275.85	1012.75	1003.53	1044.60

its dependence on the number of entities N, number of rankers M and number of covariates L. Table 6 shows the run times in seconds of BARC with 100 MCMC iterations for data from Scenario 2 in Section 6.1 with various values of N, M, L, using a laptop with 2.9 GHz Intel Core i9. From Table 6, the number of covariates L does not affect the run time much, but both the number of entities N and number of rankers M affect the run time considerably. In particular, the run time increases about linearly with M, and it increases faster than linearly with N, which implies that the number of entities matters more for the computational cost. The results are intuitive by noting that each Gibbs update step of BARC mainly involves sampling NM truncated normal random variables for $_j$'s and an N L multivariate normal random vector for , . In practice, we can parallelize the sampling for $_j$'s into M machines, which can reduce the run time of BARC.

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 Example 2. We simulate data from Scenario 2 with N 80, M 10 and L 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 pairwise comparisons in a partial ranking list. Table 7 displays the Kendall tau distances between the true and the aggregated ranking lists inferred by BARC models in different

TABLE 7

BARC for partial ranking lists. The first column shows the numbers of non-overlapping subgroups of equal sizes, under which we only observe ranks within each subgroup. The second to fifth columns show the average Kendall tau distances between estimated ranking lists and corresponding true ones using different forms of our BARC model. Specifically, BAR denotes our BARC mode without using any covariate, and BARC$_1$–BARC$_3$ denotes our BARC model using different priors.

Number of subgroups	BAR	BARC$_1$	BARC$_2$	BARC$_3$
1	0.120	0.101	**0.101**	0.117
2	0.127	0.101	**0.100**	0.126
4	0.136	**0.103**	0.103	0.135
8	0.162	**0.112**	0.112	0.144
10	0.173	**0.111**	0.111	0.146
16	0.203	**0.117**	0.117	0.154

cases. Specifically, we consider four Bayesian models BAR, BARC$_1$, BARC$_2$ and BARC$_3$, whose models and priors are defined the same as in Section 6.1. Table 7 shows that BARC is quite robust with respect to partial ranking lists when the unobserved pairwise comparisons are missing completely at random and the input ranking lists have moderate dependence on the available covariates, in the sense that the precision of aggregated ranking lists is relatively stable when the partial ranking lists become more and more incomplete, as demonstrated by BARC$_1$ and BARC$_2$. In contrast, the BARC method without using covariates, denoted by BAR in Table 7, is susceptible to missing information in the partial lists. Specifically, when K 16, the average Kendall tau distance between true and aggregated ranking lists using BAR increases by about 70%, while that using BARC$_1$ and BARC$_2$ increases by about 15%.

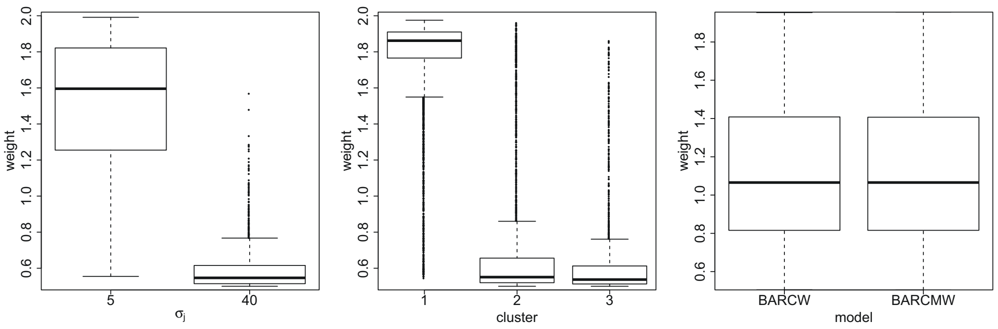
We investigate the setting where the rankers have consistent opinions but various qualities. The data are simulated from Scenarios 1–3 in Section 6.1 with N 80 and M 10, except that the noise level in (6.1) is allowed to be ranker-specific. Specifically, half of the rankers have noise level $_j$ 5 and the remaining half have noise level $_j$ 40, which represent high-quality and low-quality rankers respectively. Table 8 shows the (scaled) Kendall tau distances between the true rank and the aggregated ranks produced by different methods, averaged over the 100 simulated datasets for each scenario, where

TABLE 8

Comparison between BARC and other ranking methods when there are rankers of varying qualities. The first column shows the scenario for simulating the data. The second column with parentheses shows the average Kendall tau distances between estimated ranking lists and corresponding true ones using Borda Count, and the remaining columns without parentheses show the corresponding values for different methods but standardized by the one using Borda Count. Specifically, MC$_1$–MC$_3$ denote the three forms of Markov-Chain based methods, CEMC denotes the Cross Entropy Monte Carlo based method (CEMC), PL denotes the Plackett–Luce model, BARC denotes our BARC mode, and BARCW denotes our BARCW model.

Scenario	Borda	MC$_1$	MC$_2$	MC$_3$	CEMC	PL	BARC	BARCW
1	0.214	1.922	0.972	1.009	1.024	1.041	0.676	**0.579**
2	0.200	1.866	0.950	1.015	1.007	1.052	0.753	**0.651**
3	0.275	1.625	0.989	1.005	1.008	1.044	0.998	**0.943**

the values for all methods other than Borda Count are standardized by the corresponding values for Borda Count. Moreover, we use the same prior specification for BARC and BARCW models, i.e., $\tau_\alpha = 1$, $\tau_\beta = 10$ and $\nu_\alpha = \nu_\beta = 3$. From Table 8, BARC shows more precise rank aggregation than other methods under comparison, especially when the covariates can linearly explain the underlying true scores for ranking. By exploring the varying qualities of the rankers, BARCW further improves BARC, and the improvement increases with the role of covariates for linearly explaining the ranking mechanisms. Figure 3(a) shows the box plots of the posterior means of the weights for high-quality and low-quality rankers, separately. From Figure 3(a), the weights for rankers with different qualities are well separated, and thus BARCW is able to identify rankers of different qualities.



(a) Section 6.4 (b) Section 6.5 (c) Section 6.6

Fig 3: Box plots of the posterior means of weights for certain subgroup of rankers. Specifically, (a) shows the box plots for rankers with noise levels $\sigma_j = 5$ and $\sigma_j = 40$, respectively, over all simulations from Scenarios 1–3 in Section 6.4. (b) shows the box plots for rankers in different clusters over all simulations from Scenario I in Section 6.5. (c) shows the box plots for all rankers over all simulations in Section 6.6 using models BARCW and BARCMW, respectively.

6.5 BARCM for rankers with heterogeneous opinions

In many applications, there can be multiple groups of rankers with different opinions, despite all being reliable rankers. The Dirichlet process mixture model (3.3)-(3.4) can be used to determine the total number of clusters and to cluster the rankers. Here, we use simulations to test the sensitivity of BARCM to hyperparameter settings in the Dirichlet process prior. In particular, following the discussion in Section 4.4, we fix the degrees of freedom ν_α and ν_β at 3, and study the sensitivity of BARCM to the scale hyperparameters τ_α^2 and τ_β^2 for the variances σ_α^2 and σ_β^2 of latent variables and hyperparameters a_γ and b_γ for the concentration parameter γ of the Dirichlet process. For the variances, we consider two choices of $(\tau_\alpha, \tau_\beta)$: $(1, 10)$ and $(10, 10)$. For the concentration parameter, we consider two choices of (a_γ, b_γ) suggested by Escobar and West (1995) and Frühwirth-Schnatter and Malsiner-Walli (2019): $(2, 4)$ and $(1, 20)$, where the latter implies a smaller number of clusters *a priori*; see Frühwirth-Schnatter and Malsiner-Walli (2019) for more detailed discussion. The combination of these choices results in four prior specifications for the BARCM model, and we denote them by BARCM$_1$, BARCM$_2$, BARCM$_3$ and BARCM$_4$ respectively, as shown in Table 9.

We consider two simulation scenarios under the BARC model as in (2.3) but with three mixture components. For each scenario, we have N entities evenly divided into G non-overlapping subgroups, L covariates for each entity, and M rankers who rank only entities within the same subgroup. The categories of rankers are generated from a Multinomial distribution with probabilities $0.5, 0.3, 0.2$. The covariates $_i$'s are generated from a multivariate Normal distribution with mean zero and pairwise covariances $Cov\ x_{is}, x_{it}\qquad 0.2^{\ |s-t|}$ for $1 \leq s, t \leq L$, the coefficients are generated from $\ ^{\langle k \rangle} \overset{i.i.d.}{\sim} \mathcal{N}\ 0, 4\ _N$ and $\ ^{\langle k \rangle} \overset{i.i.d.}{\sim} \mathcal{N}\ 0,\ _L$ for $1 \leq k \leq 3$, and the noise level is fixed at 1. In Scenario I, we choose N 20, G 1, L 3 and M 100, i.e., each ranker provides a full ranking list for all the entities. In Scenario II, we choose N 108, G 9, L 11 and M 69, i.e., each ranker provides only a partial ranking list comparing units within the same subgroup of N/G 12 units. Scenario II mimics the dataset in Example 2.

Table 9 shows the accuracy (measured by the adjusted Rand index) of the *maximum a posteriori* (MAP) estimate of the clustering indicators and the posterior expected number of clusters from the BARCM model with different hyperparameters, averaged over 100 simulated datasets. From Table 9, the results are relatively robust with respect to different choices of priors, although a large value of leads to a slight overestimation of the number of clusters. Intuitively, this may be due to the fact that a larger value of implies a larger signal noise ratio, thus requesting more consistent rankings among rankers in the same cluster and encouraging more clusters of rankers.

TABLE 9
Accuracy of the clustering assignments and posterior expected number of clusters under BARCM with different prior specifications.

Scenario		BARCM$_1$	BARCM$_2$	BARCM$_3$	BARCM$_4$
	$,\quad, a\ , b$	(1, 10, 2, 4)	(10, 10, 2, 4)	(1, 10, 1, 20)	(10, 10, 1, 20)
I	Clustering accuracy	1.000	0.998	1.000	0.999
	Expected # of clusters	3.000	3.007	3.001	2.991
II	Clustering accuracy	0.998	0.973	0.998	0.972
	Expected # of clusters	3.010	3.300	3.010	3.312

Here we also explore the performance of BARCW when there are indeed mixture subgroups of rankers with different ranking opinions, i.e., using a weighting strategy to construct a "consensus". We fit the BARCW model under the prior that 1, 10 and 1. Figure 3(b) shows the box plots of the posterior means of weights for rankers in different clusters over all 100 simulated data sets from Scenario I, which demonstrates that the majority opinions are up-weighted by BARCW, while the other opinions are down-weighted. This is intuitive and expected since BARCW assumes that all rankers share the same opinion. As a result, BARCW reinforces the majority's opinion in rank aggregation. By studying rankers' heterogeneity using either BARCW or BARCM, we can better understand our ranking data even if we seek only one aggregated ranking list.

In contrast to the simulation with heterogeneous rankers, we also simulated the BARC model under the homogeneous setting to verify the robustness of BARCM and BARCW, as well as BARCMW. The simulation is the same as Scenario I in Section 6.5 except that all rankers are from one component with equal qualities. We fit the BARCW, BARCM and BARCMW models with hyperparameters 1, 10, 3, a 2 and b 4. The clustering accuracy for BARCM and BARCMW averaged over all 100 simulated datasets are, respectively, 0.99 and 1. Thus, both models classify the rankers into one cluster, i.e., the rankers have consist opinions. Figure 3(c) shows the box plots of the posterior means of weights for all rankers from BARCW and BARCMW over all 100 simulated datasets. From Figure 3(c), all rankers have similar qualities, which is consistent with the true data generation models.

Below we will analyze the two applications in Examples 1 and 2 using our Bayesian models. Based on the simulation studies in Section 6, we set the hyperparameters for the Bayesian models to be 1, 10, 3, a 2 and b 4.

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 and 2 enables us to generate aggregated lists using both rank data and the covariate information, as shown in Table 10. For quarterbacks at the top and bottom of the list, these methods mostly agree with each other. Besides only looking at the aggregated ranking lists, as suggested in Section 5.1, it is important to investigate the uncertainty in rank aggregation, which can help mitigate and explain the discrepancy across different methods. Using BARCW as an example, Figure 4(a) shows the 95% credible intervals for all quarterbacks' ranked positions under BARCW. We can see that the interval width is large for mediocre quarterbacks, which 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. In practice, this may suggest an aggregated ranking list with ties or a bucket order for several subgroups of entities, instead of a full ranking list with considerable uncertainties; see, e.g., D'Ambrosio et al. (2019), Kenkre et al. (2011) and Gionis et al. (2006) for related discussions.

All methods except BARCW, BARCM and BARCMW assume equal reliability for all rankers. After analyzing the data using BARCW, we show in Figure 5(a) the box plots as well as the means of the posterior samples of the weights for all rankers. Out of the 13 rankers, six are inferred to have significantly higher quality than the others with a majority of their posterior samples of weights being greater than or equal to 1. The second ranker seems to have medium quality with weight close to 1, while the remaining rankers all have weights close to 0.5. We further validate our weight estimation using the prediction accuracy of the experts at the end of the season. Figure 5(b) plots this prediction accuracy against the posterior mean weight of each ranker resulting from BARCW,

Table 10

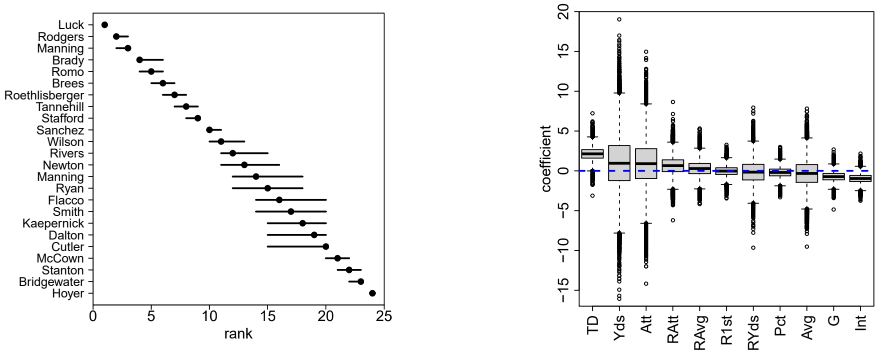
NFL rank aggregation using different methods. The first column shows the players' names, and the remaining columns show their ranked positions using different methods. BARC, BARCW, BARCM and BARCMW denote our Bayesian models for rank data, Borda denotes Borda Count, and MC_3 denotes the Markov-Chain based method.

Player	BARC	BARCW	BARCM	BARCMW	Borda	MC_3
Andrew Luck	1	1	1	1	1	1
Aaron Rodgers	2	2	2	2	2	2
Peyton Manning	3	3	3	3	3	3
Tom Brady	4	4	4	4	4	4
Tony Romo	5	5	5	5	5	5
Drew Brees	6	6	6	6	6	6
Ben Roethlisberger	7	7	7	7	7	7
Ryan Tannehill	8	8	8	8	8	8
Matthew Stafford	9	9	9	9	9	9
Mark Sanchez	10	10	10	10	10	10
Russell Wilson	11	11	11	11	11	11
Philip Rivers	12	12	12	12	12	12
Cam Newton	13	13	13	13	13	13
Eli Manning	14	14	14	14	14	14
Matt Ryan	15	15	15	15	15	15
Joe Flacco	19	16	19	16	19	19
Alex Smith	17	17	17	17	17	17
Colin Kaepernick	16	18	16	18	16	16
Andy Dalton	20	19	20	19	20	20
Jay Cutler	18	20	18	20	18	18
Josh McCown	21	21	21	21	21	21
Drew Stanton	22	22	22	22	22	22
Teddy Bridgewater	23	23	23	23	23	23
Brian Hoyer	24	24	24	24	24	24

which shows that rankers with higher weights predicts more accurately on average, and the correlation between these two measures is quite high at 0.784.

Under either BARCM or BARCMW, the 13 rankers are clustered into subgroups with different ranking opinions. To avoid the impact of multimodal posterior distributions, we run 100 MCMC chains with different random initial starts under either BARCM or BARCMW, and then choose the *maximum a posteriori* (MAP) estimate (i.e., the one with the highest joint posterior density). The resulting MAP estimates of clustering under both BARCM and BARCMW suggest that all the 13 experts belong to the same cluster, strongly suggesting that these experts share the same ranking opinion but have different qualities. Consequently, BARCW seems to be the most appropriate model for this application.

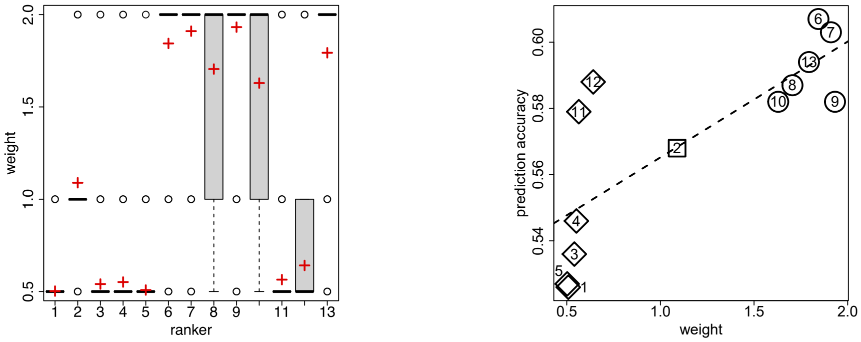
We further investigate the role of covariates in ranking these players. Figure 4(b) shows the posterior means and 95% posterior credible intervals for the coefficients of the eleven standardized covariates listed in Table 2. TD and Int, which stand for percentage of touchdowns and interceptions thrown when attempting to pass, are the most significant covariates; touchdowns have a positive effect, while interceptions have a negative effect. Based on the football common sense, touchdowns and interceptions can directly impact the result of a game.



(a) Ranked position

(b) Covariates' coefficients

Fig 4: (a) shows the credible intervals of the ranked position of the NFL quarterbacks under BARCW, and (b) shows the corresponding box plots of the posterior samples of the coefficients $_l$'s for the standardized covariates.



(a) Box plot of weights

(b) Accuracy versus weight

Fig 5: (a) shows the box plots of the posterior samples of the weights for all rankers under BARCW, and (b) plots the prediction accuracy against the posterior mean of the weight for all rankers.

As mentioned in Section 1, the orthodontics data set contains 69 partial 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 regarded as the cornerstone of systems for the evaluation of orthodontic treatment outcome (Liu et al., 2012; Song et al., 2014, 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 3 shows that the rankers do have significantly different 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 a deeper understanding of these ranking lists, we first study the heterogeneity among rankers using BARCM and BARCMW. To avoid the impact of multimodal posterior distributions, similar to Section 7.1, we run 100 MCMC chains with random initial starts under either BARCM or BARCMW, and choose the one that gives rise to the MAP estimate. Figure 6 shows the MAP estimates of the clusters, demonstrating that the 69 experts are clustered into 2 subgroups of sizes $\ 45, 24\ $, and the clustering of rankers using BARCM and BARCMW are quite consistent, with only one individual (the 25th in Figure 6) clustered differently. Specifically, this expert was estimated by BARCM and BARCMW to have 24% and 81% probabilities, respectively, to be in the bigger cluster (red dots in Figure 6), indicating a moderate amount of clustering uncertainty. See Supplementary Material for posterior clustering probabilities for all the 69 experts. From Figure 6, about $45/69\quad 65.2\%$ of the experts share consistent opinions about the ranking of the 108 patients, while the remaining experts rank the patients in a different way. This implies that most discrepancy among the experts for ranking the patients should not be explained by the quality variations of the experts, but are attributable to their different opinions.

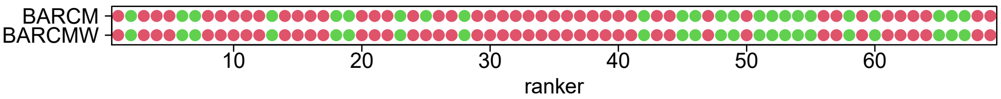
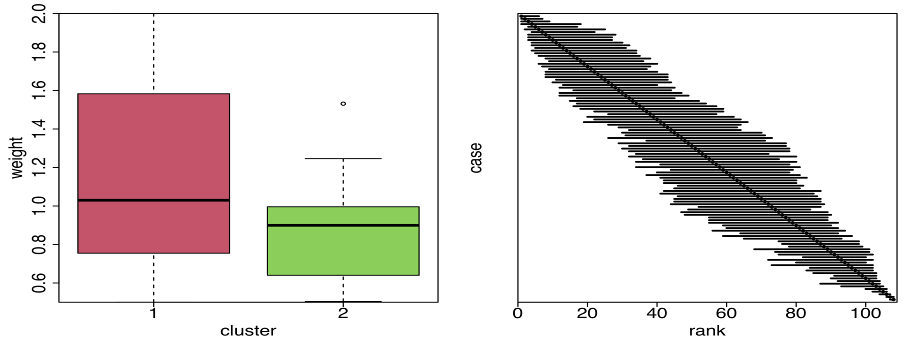


Fig 6: The MAP estimates of the clustering of the rankers under BARCM and BARCMW.

Figure 7(a) shows the box plot of rankers' weights resulting from BARCW by their estimated clusters from BARCM. A majority of rankers in the larger cluster are labeled as reliable rankers, and most rankers in the smaller cluster are labeled as mediocre or low-quality rankers. This result is similar to our simulation results in Section 6.5, i.e., in order to form a "consensus", BARCW down-weights the minority opinions when heterogeneous opinions exist.

We then study rank aggregation using our Bayesian models. The key to aggregating 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. Table 11 shows the top and bottom cases in aggregated ranking lists using different models, as well as aggregated ranking lists for each cluster under BARCM. 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 employ different assumptions. The Kendall tau distance between these two aggregated lists is 0.043. Figure 7(b) shows the 95% credible intervals for the ranked positions of the 108 cases, demonstrating that there is a substantial amount of uncertainty in the aggregated ranking list, especially around the middle of the ranking list.

Figure 8(a) shows the posterior means and 95% credible intervals of the coefficients $\ _l$'s under BARC and BARCW, as well as the average coefficients $m^{-1}\sum_{j\ 1}^{m}\ {}_l^j$'s under BARCM and BARCMW. Under these four models, the co-



(a) Ranker's weight

(b) Rank interval

Fig 7: (a) shows the box plot of the posterior means of weights for rankers in different clusters estimated by BARCM. (b) shows the 95% credible intervals of the ranked positions of the cases under BARCM.

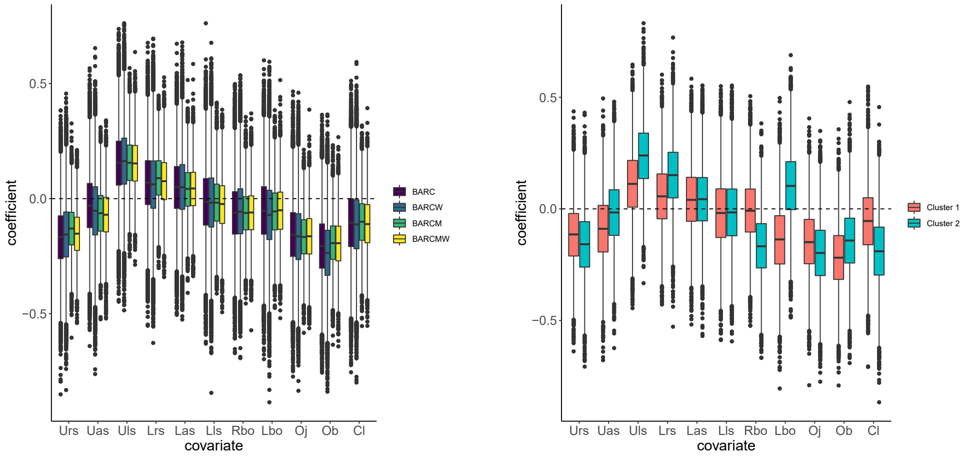
Table 11

The five cases that are considered to have the best and worst conditions based on rank aggregation. The first column denotes the ranked position. The second to fifth columns show the top and bottom five cases in the aggregated ranking lists from our Bayesian models for rank data. The last two columns show the top and bottom five cases in the aggregated ranking lists for the two clusters of rankers estimated from BARCM.

	BARC	BARCW	BARCM	BARCWM	Cluster 1	Cluster 2
1	G7	G7	G7	G7	H2	G7
2	H2	E2	E2	E2	G7	E2
3	E2	H2	H2	H2	E2	A1
4	H3	H3	F8	F8	F8	E10
5	H4	H4	H3	H3	H3	E1
104	E6	D11	D11	D11	E6	D11
105	D11	E6	E6	E6	D11	E6
106	F10	F10	F10	F10	F10	F10
107	H5	H5	H5	F4	F4	H5
108	F4	F4	F4	H5	H5	F4

variates have very similar roles in determining the rank, and are crucial for positioning patients in those non-overlapping groups. In particular, among these 11 covariates, overjet, overbite and centerline all measure certain types of overall displacement, and are thus generally considered to have stronger negative effect compared with the other local displacements in this study. This intuition is further confirmed by our analysis results. Figure 8(b) shows the posterior means and 95% credible intervals of the coefficients $_{l}^{\langle k\rangle}$'s for the two clusters under BARCM. Rankers in clusters 1 and 2 differ by putting different signs on the effect of left buccal occlusion.

Motivated by two examples we encountered in practice, we reviewed existing literature on the statistical inference of the Thurstone family models for ranking data, which include the celebrated Thurstone-Mosteller-Daniels model, the Plackett-Luce model, and their various extensions for handling more com-



(a) Coefficients under four models

(b) Coefficients for the two clusters from BARCM

Fig 8: Posterior means and 95% probability intervals of the coefficients for standardized covariates in orthodontics data under the four Bayesian models and for the two clusters under BARCM. Please refer Table 4 for the covariate information.

plex structures and situations, such as when some covariates for the ranked entities are observed and/or when the ranking lists may be composed of heterogeneous groups (or equivalently, the rankers may be clustered into different opinion subgroups).

In addition, we described three novel model-based Bayesian rank analysis methods (BARC, BARCW, BARCM), which are based on the TMD modeling framework, unified and extended existing models and methods, and proposed efficient MCMC algorithms for their needed computations. With the help of covariates, our new 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 overweights high-quality ones in rank aggregation. BARCM, on the other hand, investigates the possibility of having heterogeneous opinion groups among the rankers. All three methods evaluate the roles of covariates and generate aggregated ranking lists with uncertainty measures. Our simulation studies and real-data applications validate the importance of covariate information and the estimation of rankers' qualities as well as their heterogeneous opinions.

Our extension to the Thurstone model is similar in spirit to Vitelli et al. (2017)'s extension of the Mallows model, another popular model for rank data, but we additionally consider the incorporation of covariate information. Comparing the Thurstone and Mallows models, the former can be more general in modeling the differences among entities. For example, any miss ordering of two consecutive entities will have the same probability under the Mallows model with Kendall tau distance, but its probability depends on the underlying true score as well as the noise distribution under the Thurstone model. However, the Mallows model can be more robust since the aggregated ranking list minimizes certain average distance from all individual ranking lists, regardless of the underlying data generating process. To make the Thurstone model more robust to significant heterogeneity across individual ranking lists, it is of interest to extend the Thurstone model to accommodate more heavy-tailed error distributions, and to develop more efficient MCMC algorithms to deal with the Thurstone mixture models.

In this paper we consider only the covariate information of the ranked entities. It is of interest to further incorporate covariate information of the rankers if such data are available. Rankers' covariates can be helpful for detecting rankers' qualities and clustering rankers into subgroups with different opinions. We leave this extension of BARC and BARCM for a future study.

This research is supported in part by the NSF Grants DMS-1712714 and DMS-1903139.

Agresti, A. (1990). *Categorical Data Analysis*. New York: Wiley.

Ailon, N. (2010). Aggregation of partial rankings, p-ratings and top-m lists. *Algorithmica*, 57:284–300.

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

Allison, P. D. and Christakis, N. A. (1994). Logit models for sets of ranked items. *Sociological Methodology*, 24:199–228.

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

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

Azari Soufiani, H. (2014). *Revisiting Random Utility Models*. PhD thesis, Harvard University.

Azari Soufiani, H., Chen, W., Parkes, D. C., and Xia, L. (2013). Generalized method-of-moments for rank aggregation. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 2706–2714. Curran Associates, Inc.

Badgeley, M. A., Chikina, M. D., and Sealfon, S. C. (2014). Hybrid Bayesian-rank integration approach improves the predictive power of genomic dataset aggregation. *Bioinformatics*, 31:209–215.

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

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

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

Block, H. D. and Marschak, J. (1960). Random orderings and stochastic theories of responses. In *Contributions to Probability and Statistics*, pages 97–132.

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 Fligner, M. A. and Verducci, J. S., editors, *Probability Models and Statistical Analyses for Ranking Data*, pages 157–172. Springer New York, New York, NY.

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

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

Bradley, R. A. and Terry, M. E. (1952). Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39:324–345.

Caron, F. and Doucet, A. (2012). Efficient bayesian inference for generalized bradleyterry models. *Journal of Computational and Graphical Statistics*, 21:174–196.

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

D'Ambrosio, A., Iorio, C., Staiano, M., and Siciliano, R. (2019). Median constrained bucket order rank aggregation. *Computational Statistics*, 34:787–802.

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

David, H. A. (1963). *The method of paired comparisons*, volume 12. Oxford: Oxford University Press.

Davidson, R. R. and Farquhar, P. H. (1976). A bibliography on the method of paired comparisons. *Biometrics*, 32:241–252.

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

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

Diaconis, P. (1988). Group representations in probability and statistics. *Lecture Notes-Monograph Series*, 11:1–192.

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

Escobar, M. D. and West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, 90:577–588.

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

Fong, D. K. H., Kim, S., Chen, Z., and DeSarbo, W. S. (2016). A bayesian multinomial probit model for the analysis of panel choice data. *Psychometrika*, 81:161–183.

Frühwirth-Schnatter, S. and Malsiner-Walli, G. (2019). From here to infinity: sparse finite versus dirichlet process mixtures in model-based clustering. *Advances in Data Analysis and Classification*, 13:33–64.

Gionis, A., Mannila, H., Puolamäki, K., and Ukkonen, A. (2006). Algorithms for discovering bucket orders from data. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 561–566.

Gormley, I. C. and 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. and 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. and 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. and Murphy, T. B. (2009). A grade of membership model for rank data. *Bayesian Analysis*, 4:265–295.

Gormley, I. C. and 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*, pages 543–569.

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

Guiver, J. and Snelson, E. (2009). Bayesian inference for plackett-luce ranking models. In *proceedings of the 26th annual international conference on machine learning*, pages 377–384. ACM.

Hastie, T. and Tibshirani, R. (1998). Classification by pairwise coupling. *The Annals of Statistics*, 26:451–471.

Hausman, J. A. and Ruud, P. A. (1987). Specifying and testing econometric models for rank-ordered data. *Journal of Econometrics*, 34:83–104.

Hoff, P. D. (2009). *A first course in Bayesian statistical methods*. Springer, New York.

Huang, T.-K., Weng, R., and Lin, C.-J. (2006). Generalized bradleyterry models and multi-class probability estimates. *Journal of Machine Learning Research*, 7:85–115.

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

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., and 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.

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

Kenkre, S., Khan, A., and Pandit, V. (2011). On discovering bucket orders from preference data. In *Proceedings of the 2011 SIAM International Conference on Data Mining*, pages 872–883. SIAM.

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

Li, H., Xu, M., Liu, J. S., and Fan, X. (2020). An extended mallows model for ranked data aggregation. *Journal of the American Statistical Association*, 115(530):730–746.

Li, X., Choudhary, P. K., Biswas, S., and Wang, X. (2018). A bayesian latent variable approach to aggregation of partial and top-ranked lists in genomic studies. *Statistics in Medicine*, 37:4266–4278.

Li, X., Wang, X., and Xiao, G. (2017). A comparative study of rank aggregation methods for partial and top ranked lists in genomic applications. *Briefings in Bioinformatics*, 20:178–189.

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

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

Liu, A., Zhao, Z., Liao, C., Lu, P., and Xia, L. (2019). Learning plackett-luce mixtures from partial preferences. In *Proceedings of the Thirty-third AAAI Conference on Artificial Intelligence (AAAI-19)*.

Liu, J. S. (1994). The collapsed gibbs sampler in bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, 89(427):958–966.

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

Liu, S.-Q., Shen, G., Bai, D., Zhou, H., Li, S., Chen, W.-J., Wang, D.-W., Li, W.-R., Geng, Z., and Xu, T.-M. (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.

Liu, T.-Y. (2009). Learning to rank for information retrieval. *Foundations and Trends in Information Retrieval*, 3:225–331.

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

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*, pages 114–130.

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

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

McFadden, D. (1980). Econometric models for probabilistic choice among products. *The Journal of Business*, 53:S13–S29.

Meila, M. and Chen, H. (2012). Dirichlet process mixtures of generalized mallows models. *arXiv preprint arXiv:1203.3496*.

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.

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

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

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

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

Rao, P. V. and Kupper, L. L. (1967). Ties in paired-comparison experiments: A generalization of the bradley-terry model. *Journal of the American Statistical Association*, 62:194–204.

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

Rubinstein, R. Y. and Kroese, D. P. (2004). *The cross-entropy method: a unified approach to combi-*

natorial optimization, Monte-Carlo simulation and machine learning. Springer Science & Business Media.

Schimek, M. G., Budinská, E., Kugler, K. G., Švendová, V., Ding, J., and Lin, S. (2015). Topklists: a comprehensive r package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists. *Stat Appl Genet Mol Biol*, pages 311–6.

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., and 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., and Xu, T.-M. (2014). Reliability assessment and correlation analysis of evaluating orthodontic treatment outcome in chinese patients. *International journal of oral science*, 6(1):50–55.

Stern, H. (1990). Models for distributions on permutations. *Journal of the American Statistical Association*, 85:558–564.

Švendová, V. and Schimek, M. G. (2017). A novel method for estimating the common signals for consensus across multiple ranked lists. *Computational Statistics & Data Analysis*, 115:122–135.

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

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

Van Erp, M. and 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*.

Vitelli, V., Sørensen, Ø., Crispino, M., Frigessi, A., and Arjas, E. (2017). Probabilistic preference learning with the mallows rank model. *The Journal of Machine Learning Research*, 18:5796–5844.

Walker, J. and Ben-Akiva, M. (2002). Generalized random utility model. *Mathematical Social Sciences*, 43:303–343.

Xia, L. and Conitzer, V. (2011). A maximum likelihood approach towards aggregating partial orders. In *Twenty-Second International Joint Conference on Artificial Intelligence*.

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

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

Supplementary Material for "Bayesian Analysis of Rank Data with Covariates and Heterogeneous Rankers"

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 to denote the marginal posterior distribution of given all the observed ranking lists $\mathcal{T}$, i.e.,

$$p \qquad | \mathcal{T} \;\; \propto p \qquad p \; \mathcal{T} \; | \qquad\quad p \qquad 1\{\text{rank} \qquad\quad \mathcal{T}\}.$$

In order to show the validity of parameter expansion, it suffices to prove that for any following the marginal posterior distribution , its transformation t also follows the same distribution , as long as is draw from the distribution with density proportional to $t \qquad |J \qquad | \;{}^{-1}$. The proof is as follows.

By construction, the joint density of , is

$$p \quad , \qquad p \qquad p \quad | \qquad\qquad \cdot \frac{t \qquad\quad {}^{-nm-1}}{\int_{\mathbb{R}} \; t \qquad\quad {}^{-nm-1}\mathrm{d}},$$

which immediately implies the joint density of , $\equiv t$, :

$$p \quad , \qquad p \quad , \; |J \qquad |^{-1} \qquad\qquad \cdot \frac{t \qquad {}^{-1}}{\int_{\mathbb{R}} \; t \qquad\quad {}^{-nm-1}\mathrm{d}}$$

$$\text{(A1)} \qquad\qquad t^{-1} \qquad\qquad \cdot \frac{{}^{-1}}{\int_{\mathbb{R}} \; t \;\; t^{-1} \qquad\qquad {}^{-nm-1}\mathrm{d}}.$$

Note that $t \quad t^{-1}$ $/$ t^{-1} , where $/$. We can then simplify the denominator in (A1) as

$$\int_{\mathbb{R}} \; t \;\; t^{-1} \qquad\qquad {}^{-nm-1}\mathrm{d} \qquad \int_{\mathbb{R}} \; t^{-1} \qquad\quad / \quad {}^{-nm-1}\mathrm{d} \quad /$$

$$\qquad\qquad {}^{-nm} \int_{\mathbb{R}} \; t^{-1} \qquad\quad \cdot \;\; {}^{nm-1}\mathrm{d} \;\; ,$$

and thus further simplify p , as

$$p \quad , \qquad\qquad \cdot \frac{t^{-1} \qquad\quad {}^{-1}}{{}^{-nm} \int_{\mathbb{R}} \; t^{-1} \qquad \cdot \;\; {}^{nm-1}\mathrm{d}} \qquad\qquad \cdot \frac{t^{-1} \qquad\quad {}^{nm-1}}{\int_{\mathbb{R}} \; t^{-1} \qquad \cdot \;\; {}^{nm-1}\mathrm{d}}.$$

Therefore, the marginal density of is

$$p \qquad\qquad\qquad \cdot \frac{\int_{\mathbb{R}} \; t^{-1} \qquad\quad {}^{nm-1}\mathrm{d}}{\int_{\mathbb{R}} \; t^{-1} \qquad \cdot \;\; {}^{nm-1}\mathrm{d}} \qquad\qquad ,$$

i.e., $\equiv t$ follows the distribution with density .

The Gibbs sampler with parameter expansion for BARC model is accomplished by iterating the following steps.

(1) For j $1,\ldots,M$ and i $1,\ldots,N$, draw $[Z_{ij} \mid {}_{-i,j}, {}_{-j}, ,]$ from truncated $\mathcal{N}\ {}_i \quad {}_i^{\top}\ , 1$, where the truncation points are determined by ${}_{-i,j}$ and ${}_j$ such that rank ${}_j \simeq {}_j$.

(2) Draw $\sim S/\ {}_{NM}^{2}\ {}^{1/2}$ and then update to be $/$, where

$$S \quad \sum_{j\ 1}^{M} {}_j^{\top}\ {}_j - \sum_{j\ 1}^{M}\sum_{j'\ 1}^{M} {}_j^{\top} \quad \mathbf{\Lambda}^{-1} \quad M \ {}^{\top} \quad {}^{-1} \ {}^{\top} \quad {}_{j'}.$$

(3) Draw $,\ \sim \mathcal{N}\ , \mathbf{\Sigma}$, where

$$\mathbf{\Sigma} \ {}^{\top}\sum_{j\ 1}^{M} {}_j \quad \text{and} \quad \mathbf{\Sigma} \quad \mathbf{\Lambda}^{-1} \quad M \ {}^{\top} \quad {}^{-1}.$$

(4) Draw ${}^2 \sim \quad {}^2 \ \sum_{i\ 1}^{N} {}_i^2 \ / \ {}_N^2$ and ${}^2 \sim \quad {}^2 \ \sum_{l\ 1}^{L} {}_l^2 \ / \ {}_L^2$.

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

(1) For i $1,\ldots,N$ and j $1,\ldots,M$, draw $[Z_{ij} \mid {}_{-i,j}, {}_{-j}, ,]$ from truncated $\mathcal{N}\ {}_i \quad {}_i^{\top}\ , w_j^{-1}$ where the truncation points are determined by ${}_{-i,j}$ and ${}_j$ such that rank ${}_j \simeq {}_j$.

(2) Draw $\sim S^{1/2}/\ {}_{NM}$ and then update to be $/$, where

$$S \quad \sum_{j\ 1}^{M} w_j \ {}_j^{\top}\ {}_j - \sum_{j\ 1}^{M}\sum_{j'\ 1}^{M} w_j w_{j'} \ {}_j^{\top} \left(\mathbf{\Lambda}^{-1} \quad \sum_{m\ 1}^{M} w_m \quad {}^{\top} \right)^{-1} \ {}^{\top} \quad {}_{j'}.$$

(3) Draw $,\ \sim \mathcal{N}\ , \mathbf{\Sigma}$, where

$$\mathbf{\Sigma} \ {}^{\top}\sum_{j\ 1}^{m} w_j \ {}_j, \quad \mathbf{\Sigma} \quad \left(\mathbf{\Lambda}^{-1} \quad \sum_{j\ 1}^{M} w_j \quad {}^{\top} \right)^{-1}.$$

(4) For j $1,\ldots,M$, draw w_j from a probability mass function proportional to $w_j^{\frac{N}{2}} e^{-w_j \|\boldsymbol{Z}_j - \boldsymbol{\alpha} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2/2}$.

(5) Draw ${}^2 \sim \quad {}^2 \ \sum_{i\ 1}^{N} {}_i^2 \ / \ {}_N^2$ and ${}^2 \sim \quad {}^2 \ \sum_{l\ 1}^{L} {}_l^2 \ / \ {}_L^2$.

The Gibbs sampler with parameter expansion for BARCM model is accomplished by iterating the following steps.

(1) For each $k \in \{c_1,\ldots,c_M\}$, draw ${}_k \sim S_k^{1/2}/\ {}_{N\cdot|\mathcal{R}_k\ \boldsymbol{c}\ |}$ and then update ${}_j$ to be ${}_j/\ {}_k$ for j with c_j k, where

$$S_k \quad \sum_{j\in\mathcal{R}_k\ \boldsymbol{c}} {}_j^{\top}\ {}_j - \sum_{j\in\mathcal{R}_k\ \boldsymbol{c}}\ \sum_{j'\in\mathcal{R}_k\ \boldsymbol{c}} {}_j^{\top} \left(\mathbf{\Lambda}^{-1} \quad |\mathcal{R}_k \quad | \ {}^{\top} \right)^{-1} \ {}^{\top} \quad {}_{j'}.$$

(2) For each $k \in \{c_1, \ldots, c_M\}$, draw $\quad ^{\langle k\rangle}, \quad ^{\langle k\rangle} \sim \mathcal{N} \quad _k, \mathbf{\Sigma}_k$, where

$$_k \quad \mathbf{\Sigma}_k \quad ^\top \sum_{j\in\mathcal{R}_k \ \boldsymbol{c}} \quad _j, \quad \text{and} \quad \mathbf{\Sigma}_k \quad \left(\mathbf{\Lambda}^{-1} \quad |\mathcal{R}_k \quad | \quad ^\top \right)^{-1}.$$

(3) For $i \quad 1, \ldots, N$ and $j \quad 1, \ldots, M$, draw $[Z_{ij} \mid \quad _{-i,j}, \quad _{-j}, \quad ^{\langle c_j\rangle}, \quad ^{\langle c_j\rangle}]$ from truncated $\mathcal{N} \quad ^{\langle c_j\rangle}_i \quad ^\top_i \ ^{\langle c_j\rangle}, \ 1$, where the truncation points are determined by $\quad _{-i,j}$ and $\quad _j$ such that rank $\quad _j \simeq \quad _j$.

(4) Let $\mathcal{K} \quad \{c_1, c_2, \ldots, c_M\}$ be the set of cluster labels of all units, where $\mathcal{K}$ does not contain replicable elements, and $K \quad |\mathcal{K}|$ be the cardinality of the set $\mathcal{K}$. Draw $\quad ^2 \sim \quad ^2 \ \sum_{k\in\mathcal{K}} \| \quad ^{\langle k\rangle}\|_2^2 \ / \ ^2_{KN}$, and $\quad ^2 \sim \quad ^2 \ \sum_{k\in\mathcal{K}} \| \quad ^{\langle k\rangle}\|_2^2 \ / \ ^2_{KL}$. Then draw $\quad \sim$ Beta $\quad 1, n$, and $\quad$ from a mixture Gamma distribution

$$\text{Gamma} \ a \quad K, b \ -\log \qquad 1- \quad \text{Gamma} \ a \quad K-1, b \ -\log \quad ,$$

where the weight is defined as $\quad / \ 1 - \qquad a \quad K-1 \ / \{N \ b \ - \log \quad \}$.

(5) For $j \quad 1, \ldots, M$, draw c_j from

$$\begin{aligned} & P\left(c_j \quad k \mid \quad , \quad _{-j}, \boldsymbol{\mathcal{T}}\right) \\ & \propto P\left(c_j \quad k \mid \quad _{-j}\right) \int p\left(\quad _j \mid \quad ^{\langle k\rangle}, \quad ^{\langle k\rangle}\right) p\left(\quad ^{\langle k\rangle}, \quad ^{\langle k\rangle} \mid \quad _{-j}\right) \mathrm{d} \quad ^{\langle k\rangle} \mathrm{d} \quad ^{\langle k\rangle} \\ & \propto P\left(c_j \quad k \mid \quad _{-j}\right) \cdot \exp\left\{ -\frac{1}{2} h(\{j\} \cup \mathcal{R}_k \quad _{-j} \ ) \quad \frac{1}{2} h(\mathcal{R}_k \quad _{-j} \ ) \right\}, \end{aligned}$$

where $P\left(c_j \mid \quad _{-j}\right)$ has the following form:

$$\begin{aligned} P\left(c_j \quad k \mid \quad _{-j}\right) & \quad \frac{|\mathcal{R}_k \quad _{-j} \ |}{M-1}, \qquad \text{if } k \in \{c_m : m \ / \ j\} \\ P\left(c_j \notin \{c_m : m \ / \ j\} \mid \quad _{-j}\right) & \quad \frac{\qquad}{M-1}, \end{aligned}$$

and $h \ \cdot$ is defined as

$$\begin{aligned} h \ \mathcal{R} \quad & \sum_{m\in\mathcal{R}} \quad ^\top_m \quad _m - \sum_{m\in\mathcal{R}} \sum_{m'\in\mathcal{R}} \quad ^\top_m \quad \left(\mathbf{\Lambda}^{-1} \quad |\mathcal{R}| \quad ^\top \right)^{-1} \quad ^\top \quad _{m'} \\ & \log \left| \mathbf{\Lambda}^{-1} \quad |\mathcal{R}| \quad ^\top \quad \right|, \end{aligned}$$

with $|\cdot|$ denoting the cardinality of a set or the determinant of a matrix.

The Gibbs sampler with parameter expansion for BARCMW model is accomplished by iterating the following steps.

(1) For each $k \in \{c_1, \ldots, c_M\}$, draw $\quad _k \sim S_k^{1/2} / \quad _{N\cdot|\mathcal{R}_k \ \boldsymbol{c} \ |}$ and then update $\quad _j$ to be $\quad _j / \quad _k$ for j with $c_j \quad k$, where

$$S_k \quad \sum_{j\in\mathcal{R}_k \ \boldsymbol{c}} w_j \quad ^\top_j \quad _j - \sum_{j,j'\in\mathcal{R}_k \ \boldsymbol{c}} w_j w_{j'} \quad ^\top_j \quad \left(\mathbf{\Lambda}^{-1} \quad \sum_{m\in\mathcal{R}_k \ \boldsymbol{c}} w_m \quad ^\top \right)^{-1} \quad ^\top \quad _{j'}.$$

(2) For each $k \in \{c_1, \dots, c_M\}$, draw $\quad ^{\langle k\rangle}, \quad ^{\langle k\rangle} \sim \mathcal{N} \quad _k, \boldsymbol{\Sigma}_k$, where

$$_k \quad \boldsymbol{\Sigma}_k \quad ^\top \sum_{j\in\mathcal{R}_k \ \boldsymbol{c}} w_j \quad _j, \quad \text{and} \quad \boldsymbol{\Sigma}_k \quad \left(\boldsymbol{\Lambda}^{-1} \quad \sum_{j\in\mathcal{R}_k \ \boldsymbol{c}} w_j \quad ^\top \quad \right)^{-1}.$$

(3) For $i \quad 1, \dots, N$ and $j \quad 1, \dots, M$, draw $[Z_{ij} \mid \quad _{-i,j}, \quad _{-j}, \quad ^{\langle c_j\rangle}, \quad ^{\langle c_j\rangle}]$ from truncated $\mathcal{N} \quad ^{\langle c_j\rangle}_i \quad ^\top_i \ ^{\langle c_j\rangle}, w_j^{-1}$, where the truncation points are determined by $\quad _{-i,j}$ and $\quad _j$ such that rank $\quad _j \simeq \quad _j$.

(4) For $j \quad 1, \dots, M$, draw w_j from a probability mass function proportional to $w_j^{\frac{N}{2}} e^{-w_j\left\|\boldsymbol{Z}_j-\boldsymbol{\alpha}^{\langle c_j\rangle}-\boldsymbol{X}\boldsymbol{\beta}^{\langle c_j\rangle}\right\|_2^2/2}$.

(5) Let $\mathcal{K} \quad \{c_1, c_2, \dots, c_M\}$ be the set of cluster labels of all units, where $\mathcal{K}$ does not contain replicable elements, and $K \quad |\mathcal{K}|$ be the cardinality of the set $\mathcal{K}$. Draw $\quad ^2 \sim \quad ^2 \quad \sum_{k\in\mathcal{K}} \| \quad ^{\langle k\rangle}\|_2^2 \ / \quad ^2_{KN}$, and $\quad ^2 \sim \quad ^2 \quad \sum_{k\in\mathcal{K}} \| \quad ^{\langle k\rangle}\|_2^2 \ / \quad ^2_{KL}$. Then draw $\quad \sim$ Beta $\quad 1, n$, and $\quad$ from a mixture Gamma distribution

$$\text{Gamma} \ a \quad K, b \ -\log \qquad 1 - \quad \text{Gamma} \ a \quad K-1, b \ -\log \quad ,$$

where the weight is defined as $\quad / \ 1 - \qquad a \quad K-1 \ / \{N \ b \ - \log \quad \}$.

(6) For $j \quad 1, \dots, M$, draw c_j from

$$\begin{aligned} & P\left(c_j \quad k \mid \quad , \quad _{-j}, \boldsymbol{\mathcal{T}}\right) \\ \propto \ & P\left(c_j \quad k \mid \quad _{-j}\right) \int p\left(\quad _j \mid \quad ^{\langle k\rangle}, \quad ^{\langle k\rangle}\right) p\left(\quad ^{\langle k\rangle}, \quad ^{\langle k\rangle} \mid \quad _{-j}\right) \mathrm{d} \quad ^{\langle k\rangle} \mathrm{d} \quad ^{\langle k\rangle} \\ \propto \ & P\left(c_j \quad k \mid \quad _{-j}\right) \cdot \exp\left\{-\frac{1}{2}h(\{j\} \cup \mathcal{R}_k \quad _{-j} \ ) \quad \frac{1}{2}h(\mathcal{R}_k \quad _{-j} \ )\right\}, \end{aligned}$$

where $P\left(c_j \mid \quad _{-j}\right)$ has the following form:

$$P\left(c_j \quad k \mid \quad _{-j}\right) \quad \frac{|\mathcal{R}_k \quad _{-j} \ |}{M-1}, \qquad \text{if } k \in \{c_m : m \ / \ j\}$$

$$P\left(c_j \notin \{c_m : m \ / \ j\} \mid \quad _{-j}\right) \quad \frac{\qquad}{M-1},$$

and $h \quad \cdot$ is defined as

$$\begin{aligned} h \ \mathcal{R} \quad & \sum_{m\in\mathcal{R}} w_m \quad ^\top_m \quad _m - \sum_{m\in\mathcal{R}} \sum_{m'\in\mathcal{R}} w_m w_{m'} \quad ^\top_m \left(\boldsymbol{\Lambda}^{-1} \quad \sum_{j'\in\mathcal{R}} w_{j'} \quad ^\top \quad \right)^{-1} \quad ^\top \quad _{m'} \\ & \log\left|\boldsymbol{\Lambda}^{-1} \quad \sum_{m\in\mathcal{R}} w_m \quad ^\top \quad \right|, \end{aligned}$$

with $|\cdot|$ denoting the cardinality of a set or the determinant of a matrix.

We provide an R package BayesRankAnalysis for implementing the proposed Bayesian models for rank data. A detailed description for installation and usage of the package can be found on the website https://github.com/li-xinran/BayesRankAnalysis.

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 $\{ \ _j \ i \ \}_{1\leq j\leq m}$ be the ranking positions of entity i received from all m rankers. The Borda Count method aggregates ranks based on their arithmetic mean, $\sum_{j}^{m} \ 1 \ \ _j \ i \ /m$.

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 $P \quad \{p_{i_1 i_2}\}_{i_1,i_2\in U}$, 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 $\{ \ _1,\ldots, \ _m\}$. 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 uniformly from all the partial lists $\{ \ _1,\ldots, \ _m\}$ containing entity i_1 then picking an entity i_2 uniformly from the set $\{i_2 \mid \quad i_2 \ \leq \quad i_1 \ \}$.

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

$$\text{sort } i \in U \text{ by } \quad _i \downarrow ,$$

where $\quad\quad _1,\ldots, \ _{|U|}$ satisfies $\ P \quad\quad$, and the symbol "↓" means that the entities are sorted in descending order.

PL model assumes that a ranking list $\quad\quad i_1 \succ i_2 \succ \ldots \succ i_n$ is observed with probability

$$P \quad \mid \quad\quad\quad \frac{\ _{i_1}}{\sum_{l}^{n} \ 1 \ \ _{i_l}} \times \frac{\ _{i_2}}{\sum_{l}^{n} \ 2 \ \ _{i_l}} \times \cdots \times \frac{\ _{i_1}}{\ _{i_{n-1}} \quad\ _{i_n}},$$

where $\ _i \in \ \ 0,1 \ $ and $\sum_{i}^{n} \ 1 \ \ _i \quad\ 1$. Each ranking list from $\{ \ _1,\ldots, \ _m\}$ follows the above distribution independently. We apply the classical Minorize-Maximization (MM) algorithm for PL model estimation (Hunter, 2004).

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

$$\text{(A1)} \qquad \arg \min_{\in \mathcal{S} \ U} d \quad , \ _1,\ldots, \ _m$$

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 and Ding (2009) used a stochastic search method to optimize (A1) by adopting the cross entropy Monte Carlo (CEMC) approach (Rubinstein and Kroese, 2004). In the paper we use the CEMC approach based on the Kendall tau distance.

Figure A1 shows the posterior probability of being in the bigger cluster (red dots in Figure 6) for each expert under BARCM and BARCMW. From Figure A1, the results from BARCM and BARCMW are mostly consistent, resulting in similar MAP estimates for clustering. The only exception is the 25th ranker, whose probabilities under the two models lie in different sides of 0.5, leading to different MAP estimates for clustering.

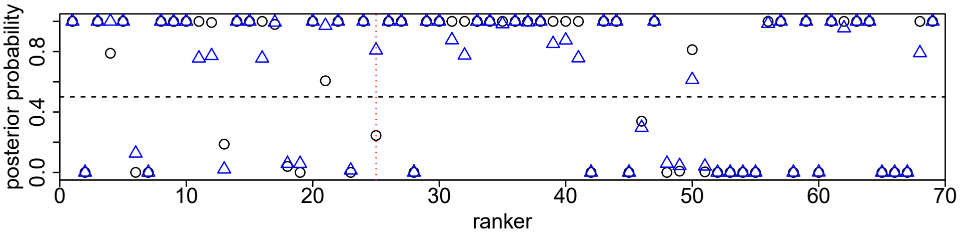


Fig A1: The posterior probability of being in the bigger cluster (red dots in Figure 6) for each ranker under BARCM (black circle) and BARCMW (blue triangle).