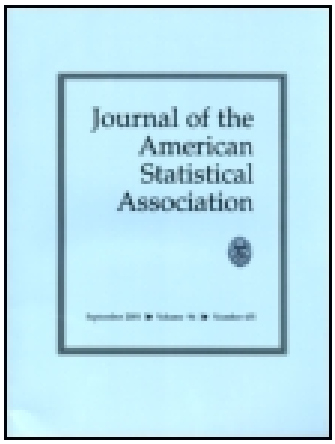


Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

Bayesian Aggregation of Order-Based Rank Data

Ke Deng, Simeng Han, Kate J. Li & Jun S. Liu

Accepted author version posted online: 14 Jan 2014.Published online: 02 Oct 2014.

To cite this article: Ke Deng, Simeng Han, Kate J. Li & Jun S. Liu (2014) Bayesian Aggregation of Order-Based Rank Data, Journal of the American Statistical Association, 109:507, 1023-1039, DOI: 10.1080/01621459.2013.878660

To link to this article: http://dx.doi.org/10.1080/01621459.2013.878660

Bayesian Aggregation of Order-Based Rank Data

Ke DENG, Simeng HAN, Kate J. LI, and Jun S. LIU

Rank aggregation, that is, combining several ranking functions (called base rankers) to get aggregated, usually stronger rankings of a given set of items, is encountered in many disciplines. Most methods in the literature assume that base rankers of interest are equally reliable. It is very common in practice, however, that some rankers are more informative and reliable than others. It is desirable to distinguish high quality base rankers from low quality ones and treat them differently. Some methods achieve this by assigning prespecified weights to base rankers. But there are no systematic and principled strategies for designing a proper weighting scheme for a practical problem. In this article, we propose a Bayesian approach, called Bayesian aggregation of rank data (BARD), to overcome this limitation. By attaching a quality parameter to each base ranker and estimating these parameters along with the aggregation process, BARD measures reliabilities of base rankers in a quantitative way and makes use of this information to improve the aggregated ranking. In addition, we design a method to detect highly correlated rankers and to account for their information redundancy appropriately. Both simulation studies and real data applications show that BARD significantly outperforms existing methods when equality of base rankers varies greatly.

KEY WORDS: Meta-analysis; Power law distribution; Rank aggregation; Spam detection.

1. INTRODUCTION

Rank aggregation aims to generate a "better" aggregated rank list (referred to as *aggregated ranker*) for a set of entities from multiple individual ranking functions (referred to as *base rankers*). Early efforts on rank aggregation can be traced back to studies on social choice theory and political elections in the eighteenth century (Borda 1781). Since mid-1990s, rank aggregation has drawn much attention with the rise of internet and web search engines. Score-based rank aggregation methods for meta-search (Shaw and Fox 1994; Manmatha, Rath, and Feng 2001; Montague and Aslam 2001; Manmatha and Sever 2002), document analysis (Hull, Pedersen, and Schütze 1996; Vogt and Cottrel 1999), and similarity search in database (Fagin, Lotem, and Naor 2001), which take score information from individual base rankers as input to generate an aggregated ranker, form the first wave of modern rank aggregation studies. However, considering that usually only order information is available in meta-search, methods that rely only on the order information from base rankers became popular quickly. The first generation of order-based methods construct the aggregated ranking function based on simple statistics of ranked lists from base rankers. For example, Van Erp and Schomaker (2000) and Aslam and Montague (2001) proposed to use a democratic voting procedure called *Borda count* (i.e., the average rank across all base rankers) to generate the aggregated rank; while Fagin, Kumar, and Sivakumar (2003b) suggested the use of median rank. To strive for better performance, more complicated methods were proposed, including Markov-chain-based methods (Dwork et al. 2001), fuzzy-logic-based method (Ahmad and Beg 2002), genetic algorithm (Beg 2004), and graph-based method (Lam and Leung 2004). As an important special case, the problem of combining the top-d lists has been given extra attention in Dwork et al. (2001) and Fagin, Kumar, and Sivakumar (2003). Freund et al. (2003) proposed a boosting method for rank aggregation with the guidance of "feedbacks" that provide information regarding relative preferences of selected pairs of entities.

Randa and Straccia (2003) compared the performance of score-based methods and rank-based methods in the context of meta-search, and found that Markov-chain-based methods performed comparably to score-based methods, but significantly outperformed methods based on Borda count. The success of Markov-chain-based methods quickly made Dwork et al. (2001) a classic. These methods were later applied to bioinformatics problems, and a number of their variations and extensions were proposed to fit more complicated situations (Sese and Morishita 2001; DeConde et al. 2006; Lin and Ding 2009).

In practice, the problem of rank aggregation can become even more challenging because of the diverse quality of base rankers. For example, in a meta-search study, some search engines are more powerful than others; in a meta-analytic bioinformatic study, some labs collect and/or analyze data more efficiently than other labs; and in a competition, some judges are more experienced and objective than others. In some extreme cases, some base rankers may be noninformative or even misleading. For example, "paid placement" and "paid inclusion" are very popular among search engines. These low quality base rankers, referred to as *spam rankers*, may disturb the rank aggregation procedure significantly if they are not treated properly. Giving different base rankers different weights, as mentioned in Aslam and Montague (2001) and Lin and Ding (2009), seems to be the only method available that takes the diverse quality of base rankers into consideration. A clear limitation of this approach, however, is that there are no systematic and principled strategies for designing a proper weighting scheme for a practical problem. Supervised rank aggregation may help learn a good weighting scheme, as proposed in Liu et al. (2007), but it needs a good set of training data, which may often be unavailable. To illustrate the main problems and ideas, consider the following simple example:

Ke Deng, Mathematical Sciences Center, Tsinghua University, Beijing 100084, China (E-mail: *kdeng@math.tsinghua.edu.cn*). Simeng Han (E-mail: *han@fas.harvard.edu*) and Jun S. Liu (E-mail: *jliu@stat.harvard.edu*), Department of Statistics, Harvard University, Cambridge, MA 02138, USA. Kate J. Li, Sawyer Business School, Suffolk University, Boston, MA 02108, USA (E-mail: *kjli@suffolk.edu*). This research was supported in part by the NSF grants DMS-0706989, DMS-1007762 and DMS-1208771, and by Shenzhen Special Fund for Strategic Emerging Industry grant (No.ZD201111080127A).

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jasa*.

Journal of the American Statistical Association
September 2014, Vol. 109, No. 507, Theory and Methods
DOI: 10.1080/01621459.2013.878660

NBA Team Ranking. In July 2012, after the 2011–2012 NBA basketball season, we sent out a questionnaire to all graduate students of the Harvard Statistics Department and a small group of summer school students who were taking the summer course STAT 100 at Harvard, asking them to select the best 8 NBA teams of the 2011–2012 season and rank them top-down based on his/her own knowledge without checking online information or consulting others. We also asked each student to classify himself/herself into one of the following four groups in the survey: (1) "Avid fans," who never missed NBA games; (2) "Fans," who watched NBA games frequently; (3) "Infrequent watchers," who watched NBA games occasionally; and (4) "Not-interested," who never watched NBA games in the past season. This extra piece of information, which is usually unavailable in most real problems, will not be used for rank aggregation, but will be used to validate the quality measure we infer from the ranking data. We received 28 responses, amounting to a 47% response rate. The data are displayed in Table 1. We also list in Table 1 six ranking results generated in December 2011 after the preseason games by professional news agencies such as NBA.com, ESPN.com, etc. Assuming that we do not know the quality of each ranker, we wish to combine these rankings into a better-quality aggregated ranking function, and to also judge each ranker's "quality" based only on these ranking results.

We propose here a Bayesian method to tackle the order-based rank aggregation problem. By reformulating the original rank aggregation problem into a Bayesian model selection problem and attaching a quality parameter to each base ranker, we can estimate the quality of base rankers jointly with rank aggregation. Compared to existing methods, our method is distinct in that it uses an explicit probabilistic model, is adaptive to the heterogeneity of base rankers, and can handle complex situations (such as with correlated rankers of different qualities) more efficiently. The remainder of the article is organized as follows. Section 2 formally defines the rank aggregation problem, and briefly reviews the existing methods. In Section 3 describes our Bayesian model for rank aggregation, explains intuitions behind the model, provides details of the corresponding Markov chain Monte Carlo algorithm, and extends the method to handle partial rankings and the issue of supervision. Section 4 proposes a tool for detecting overly correlated ranker groups and describe a hierarchical model to account for information redundancy due to ranker correlations. Section 5 provides simulation evaluations of the new Bayesian method, and Section 6 explores a few real-data applications. Section 7 concludes the article with a short discussion.

2. AN OVERVIEW OF EXISTING METHODS

Let $U = \{1, 2, \ldots, n\}$ be the "universe" (set) of n entities of interest. An ordered list (or simply, a list) τ with respect to U is a ranking of entities in a subset $S \subseteq U$, that is, $\tau = [x_1 \preceq x_2 \preceq \cdots \preceq x_d]$, where $S = \{x_1, \ldots, x_d\}$ and "$i \preceq j$" means that i is ranked better than j. Let $\tau(i)$ be the position or rank of entity $i \in \tau$ (a highly ranked element has a low-numbered position in the list). We call τ a *full list* if $S = U$; and a *partial list* otherwise. An important special case of partial lists is the *top*-d *lists*. For a list τ and a subset T of U, the projection of τ with respect to T (denoted as $\tau_{|T}$) is a new list that contains only entities from T. Note that if τ happens to contain all elements in T, then $\tau_{|T}$ is a full list with respect to T. In the NBA Team Ranking example, U is the set of all 30 NBA teams, and we observed six full lists $P_1, \ldots, P_6$ from six professional news agencies and 28 partial lists $S_1, \ldots, S_{28}$ from 28 students.

2.1 Methods Based on Summary Statistics

Many rank aggregation methods are based on simple summary statistics of the m given base rankers. Let $\{\tau_k(i)\}_{1\le k\le m}$ be the ranks that entity i receives from the m base rankers. To determine the rank of entity i in the aggregated list, the arithmetic mean, geometric mean, or median of $\{\tau_k(i)\}_{1\le k\le m}$ have all been proposed. We refer to these three methods as AriM, GeoM, and MedR, respectively. These naive methods are straightforward and perform reasonably well when rankers in consideration are full and of similar qualities. But they are easily disturbed by spam rankers and also have difficulties in dealing with partial lists. In the NBA ranking example, we may treat those unranked teams in a partial list as ranked "# 19.5," which is at the middle between 9 and 30, or an arbitrary number between 9 and 30. If the average approach AriM was used, one may end up ranking Lakers and Bulls much higher than they should be.

2.2 Optimization-Based Methods and Markov-chain-Based Methods

Dwork et al. (2001) proposed to report the list that minimizes an objective function as the aggregated rank list, that is, let

$$\alpha = \arg\min_{\sigma\in\mathscr{A}_U} d(\sigma; \tau_1, \ldots, \tau_m),$$

where $\mathscr{A}_U$ is the space of all allowable rankings of entities in U, and the objective function d can be either the *Spearman's footrule distance* (Diaconis and Graham 1977)

$$d_F(\sigma; \tau_1, \ldots, \tau_m) \triangleq \frac{1}{m}\sum_{k=1}^{m} F(\sigma_{|\tau_k}, \tau_k),$$

where $F(\sigma_{|\tau_k}, \tau_k) = \sum_{i\in\tau_k} |\sigma_{|\tau_k}(i) - \tau_k(i)|$, or the *Kendall tau distance* (Diaconis 1988)

$$d_K(\sigma; \tau_1, \ldots, \tau_m) \triangleq \frac{1}{m}\sum_{k=1}^{m} K(\sigma_{|\tau_k}, \tau_k),$$

where $K(\sigma_{|\tau_k}, \tau_k)$ is the bubble sort distance between $\sigma_{|\tau_k}$ and τ_k (i.e., the number of swaps that the *bubble sort* algorithm would make to place $\sigma_{|\tau_k}$ in the same order as τ_k). The aggregation obtained by optimizing the Kendall distance is called *Kemeny optimal aggregation*, and the one obtained by optimizing the Spearman's footrule distance is called *footrule optimal aggregation*. In fact, the idea of generating the aggregated ranking by optimizing the Kendall distance can be traced back to the Mallows model in 1950s (Mallows 1957), which is generalized by Fligner and Verducci (1986), and later by Meila et al. (2007).

Considering that it is computationally expensive to solve the above optimization problems (the Kemeny optimal aggregation is NP-Hard, and the footrule optimal aggregation needs an expensive polynomial algorithm), Dwork et al. (2001) also proposed a few Markov-chain-based methods as fast alternatives to provide suboptimal solutions. The basic idea behind these

Table 1. Power rankings of NBA teams for the 2011–2012 season collected from six professional sport-ranking web sites and a survey of 28 Harvard students

		Professional						Avid fans						Fans						Infrequent watchers							Notinterested individuals								
N.o.	Team	P_1	P_2	P_3	P_4	P_5	P_6	S_1	S_2	S_3	S_4	S_5	S_6	S_7	S_8	S_9	S_{10}	S_{11}	S_{12}	S_{13}	S_{14}	S_{15}	S_{16}	S_{17}	S_{18}	S_{19}	S_{20}	S_{21}	S_{22}	S_{23}	S_{24}	S_{25}	S_{26}	S_{27}	S_{28}
1	Heat	1	2	1	1	1	1	1	1	2	3	1	3	–	1	2	1	3	1	3	1	4	1	4	2	1	2	–	4	–	–	1	1	1	2
2	Thunder	3	3	2	3	2	3	2	2	3	2	–	2	–	–	4	2	7	4	–	–	–	2	–	–	2	–	–	–	–	2	–	–	–	–
3	Spurs	7	10	11	5	8	7	6	5	5	5	–	6	–	6	5	4	–	–	5	–	8	6	3	6	–	–	–	–	–	–	–	–	–	–
4	Celtics	5	11	10	9	9	5	4	8	1	4	2	5	2	3	1	3	4	3	–	4	2	3	2	–	4	4	–	2	–	–	–	2	–	–
5	Clippers	8	5	6	10	5	6	–	6	–	8	–	7	–	–	–	–	–	–	–	–	–	–	–	–	–	–	2	–	–	–	–	–	4	–
6	Lakers	6	7	7	6	6	8	3	7	6	1	3	1	1	2	7	7	1	2	1	2	1	4	5	1	3	1	5	1	–	–	8	4	2	–
7	Pacers	14	13	14	14	13	12	8	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	6	–	–	–	5	8	–	–
8	76ers	15	16	13	15	15	15	–	–	–	–	–	–	–	4	–	–	8	6	–	–	–	–	–	5	6	–	7	6	6	4	3	3	–	–
9	Mavericks	2	1	3	2	4	4	–	–	7	7	–	8	–	–	8	6	6	5	–	8	3	–	1	–	–	–	3	–	–	–	4	–	5	–
10	Bulls	4	4	4	4	3	2	5	4	8	6	4	4	3	–	3	–	–	8	4	–	–	–	–	–	–	3	–	3	3	5	–	5	–	1
11	Knicks	9	6	9	8	7	13	–	3	4	–	5	–	–	–	–	–	2	–	–	–	–	–	–	4	7	8	–	–	2	8	–	7	7	6
12	Grizzlies	10	8	8	7	11	10	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	8	–	–	–	–	–
13	Nuggets	19	9	5	13	10	9	–	–	–	–	8	–	–	–	–	–	–	–	–	–	–	–	7	–	–	–	–	–	–	–	–	–	8	–
14	Magic	11	12	17	11	14	11	–	–	–	–	–	–	6	–	–	–	–	–	–	5	–	–	–	–	5	–	4	–	–	–	6	6	–	–
15	Hawks	12	18	12	18	12	18	7	–	–	–	–	–	–	–	–	5	–	–	–	–	–	–	–	–	8	7	–	–	5	–	–	–	–	8
16	Jazz	18	23	26	27	28	19	–	–	–	–	–	–	8	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	3
17	TrailBlazers	13	14	15	12	16	14	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
18	Rockets	21	15	16	16	17	17	–	–	–	–	–	–	5	7	6	–	–	–	2	6	–	–	–	3	–	5	–	5	–	3	–	–	–	7
19	Bucks	16	17	20	17	20	16	–	–	–	–	–	–	–	–	–	–	–	–	–	3	–	–	–	–	–	–	–	–	–	6	–	–	–	–
20	Suns	20	22	19	21	19	21	–	–	–	–	–	–	–	8	–	–	–	–	7	7	6	8	8	–	–	–	–	–	–	–	7	–	–	–
21	Nets	17	19	24	20	24	23	–	–	–	–	–	–	–	–	–	–	5	–	–	–	–	–	–	–	–	–	–	–	7	–	–	–	3	5
22	Warriors	22	21	23	19	22	20	–	–	–	–	6	–	–	–	–	–	–	–	6	–	–	–	–	–	–	–	1	–	–	–	–	–	–	–
23	Timberwolves	23	20	22	22	23	24	–	–	–	–	–	–	–	–	–	–	–	–	–	–	7	7	–	–	–	–	–	–	4	7	–	–	–	4
24	Hornets	27	28	18	23	18	25	–	–	–	–	–	–	7	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
25	Pistons	25	25	25	24	25	22	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	8	–	–	–	8	–	–	2	–	–	–
26	Kings	29	24	21	26	21	26	–	–	–	–	–	–	4	5	–	–	–	–	8	–	5	5	–	–	–	–	–	–	1	–	–	–	–	–
27	Wizards	28	27	28	25	27	27	–	–	–	–	7	–	–	–	–	–	–	–	–	–	–	–	–	7	–	–	–	7	–	–	–	–	6	–
28	Raptors	24	26	29	28	30	28	–	–	–	–	–	–	–	–	–	8	–	–	–	–	–	–	–	–	–	–	8	–	–	1	–	–	–	–
29	Cavaliers	26	29	27	29	26	29	–	–	–	–	–	–	–	–	–	–	–	7	–	–	–	–	–	–	–	6	–	–	–	–	–	–	–	–
30	Bobcats	30	30	30	30	29	30	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	6	–	–	–	–	–	–	–	–	–	–	–

Remark: The 30 NBA teams are arranged in the table based on their performances in playoffs of the season, that is, the top 16 teams reached the playoffs, the top eight teams survived the first round of the playoffs, and so on. The six professional power rankings ($P_1, \ldots, P_6$) are downloaded from FOXSports.com, ESPN.com, SI.com, NBA.com, midwestsportsfans.com, and jsonline.com, respectively. More details about these professional power rankings are given in the Appendix. The 28 rankings by Harvard students ($S_1, \ldots, S_{28}$) are collected by a survey after the 2011–2012 season was finished, in which each student was asked to select the best eight NBA teams in the 2011–2012 season and rank them top-down based on his/her own knowledge without checking online information or consulting others. To collect information about how much the students followed NBA games in the 2011–2012 season, we also asked every student to classify himself/herself into one of the following four groups in the survey: (1) "Avid fans" who never missed NBA games, (2) "Fans" who watched NBA games frequently, (3) "Infrequent watchers" who watched NBA games occasionally, and (4) the "Not-interested" who never watched NBA games in the past season. In addition, we encouraged the students to do guess randomly if they really have no ideas about these teams.

methods is to construct a transition probability matrix $P = \{p_{ij}\}_{i,j\in U}$ based on $\{\tau_1, \ldots, \tau_m\}$, where p_{ij} is the transition probability from entity i to entity j, and use the stationary distribution of P to generate the aggregated ranked list. More precisely, we let

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

where $\pi = (\pi_1, \ldots, \pi_n)$ satisfies $\pi P = \pi$, and symbol "$\downarrow$" means that the entities are sorted in descending order. Suppose that the current state of the Markov chain is entity i, a few different transition rules were suggested by Dwork et al. (2001) and Deconde et al. (2006):

- MC_1: The next state is generated uniformly from the set of all entities that are ranked higher than (or equal to) i by some base rankers.
- MC_2: The next state is generated by first picking a base ranker τ at random from all base rankers containing entity i, and then picking a entity j at random from the set $\{j \in \tau : \tau(j) \le \tau(i)\}$.
- MC_3: A base ranker τ is chosen at random from all base rankers containing entity i and an entity j is chosen at random from all entities ranked by τ, and the next state is set at j if $\tau(j) \le \tau(i)$ and stay in i otherwise.
- MC_4: An entity j is generated uniformly from the union of all entities ranked by all base rankers. If $\tau(j) \le \tau(i)$ for a majority of base rankers that rank both i and j, then go to j; otherwise, stay in i.
- MC_T: It is almost identical to MC_4, except that the move from i to j at the last step is not a deterministic procedure based on the majority vote, but a stochastic procedure in which the probability for accepting j is proportional to the percentage of base rankers that rank j higher than i among all base rankers that rank both i and j.

2.3 Rank Aggregation of Weighted Lists

Considering that base rankers of interest may not be equally informative or reliable in practice, methods based on weighted lists are also proposed. In these methods, each base ranker τ_k is assigned a weight w_k ($0 \le w_k \le 1$ and $\sum_{1\le k\le m} w_k = 1$), and the base rankers with larger weights play more important roles in generating the aggregated list. Aslam and Montague (2001) proposed to generate the aggregated list based on the weighted average of the m lists (known as Borda Fuse), that is, let $\alpha = \text{sort}(i \in U \text{ by } \sum_{k=1}^m w_k \tau_k(i) \downarrow)$. Lin and Ding (2009) extended the objective function of Dework et al. (2001) to a weighted fashion, and generated the aggregated list as follows:

$$\alpha = \arg\min_{\sigma\in\mathcal{A}_U} d(\sigma; \tau_1, \ldots, \tau_m; w) = \arg\min_{\alpha\in\mathcal{A}_U} \sum_{k=1}^m w_k d(\alpha_{|\tau_k}, \tau_k),$$

where $d(\alpha_{|\tau_k}, \tau_k) = F(\alpha_{|\tau_k}, \tau_k)$ or $K(\alpha_{|\tau_k}, \tau_k)$. The authors also proposed using cross entropy Monte Carlo (CEMC, see Rubinstein and Kroese 2004) to solve the above optimization problem. The optimization method based on Spearman's footrule distance is denoted as CEMC_F, and that based on Kendall distance is denoted as CEMC_K.

Although assigning weights to base rankers is a sensible way of handling the quality difference among them, it can be quite difficult to design a proper weight specification scheme in practice, especially when little or no prior knowledge on base rankers is available. The *supervised rank aggregation* (SRA) of Liu et al. (2007) solves this problem at the price of extra training data. In SRA, the true relative ranks of some entities are provided as training data, and the weights $\{w_k\}_{1\le k\le m}$, which are treated as parameters instead of prespecified constants in these models, are optimized with the help of the training data as well as the aggregated list σ. A problem of SRA is that no training data are available in many applications.

2.4 Rank Aggregation Via Boosting

Another line of using training data to achieve rank aggregation in the literature is the *RankBoost* method of Freund et al. (2003). Similar to SRA, RankBoost assumes that, besides the rank lists $\{\tau_1, \ldots, \tau_m\}$, we also have a *feedback function* of the form $\Phi : U \times U \to \mathbf{R}$, where $\Phi(i, j) > 0$ means that entity i should be ranked above entity j, $\Phi(i, j) < 0$ otherwise, and $\Phi(i, j) = 0$ indicates no preference between i and j. Different from SRA, RankBoost does not assign weights to different rankings themselves. Instead, RankBoost follows the boosting idea to generate a series of *weak rankers* from $\{\tau_1, \ldots, \tau_m\}$, and construct the final ranking by a weighted average of these weak rankers.

As an illustration, we applied AriM, the four MC-based methods (MC_1, …, MC_4), and the two CEMC-based methods (CEMC_F and CEMC_K) to combine the 34 full/partial rankings listed in Table 1 without informing any method about qualities of the rankers. The new method BARD as described in the next section was also applied to the same dataset for a comparison purpose. The RankBoost method is not included in the comparison because feedbacks required by it are not available in this study. The results are displayed in Table 2 and Figure 8. Figure 8 shows that BARD properly discovered the quality difference among the 34 base rankers. This advantage in turn leads to a better performance of BARD over other tested methods, which treated all the rankers equally (Table 2). Section 6.2 gives more details on how BARD was applied to this problem.

3. A BAYESIAN MODEL FOR RANK AGGREGATION

3.1 Assumptions and the Model

Here, we propose a Bayesian approach, called *Bayesian Aggregation of Rank Data* (BARD), to tackle the problem of rank aggregation with rankers of different quality levels. This subsection focuses on the case with full lists; extensions to cases where partial lists and training data are involved will be discussed in Section 3.4. The BARD method reformulates the ranking problem as follows. We assume that the set U is composed of two nonoverlapping subsets: set U_R representing relevant entities (with true signals) and set U_B representing noisy background entities. The common goal of each base ranker is to distinguish the relevant entities from the background ones. By integrating rankings from all base rankers, we attempt to best identify the set of relevant entities.

Table 2. Performance of different methods on survey data of NBA teams

		AriM		MC1		MC2		MC3		MC4		CEMC		BARD
N.o.	Team	Mean	Rank	π_i	Rank	π_i	Rank	π_i	Rank	π_i	Rank	$d = F$	$d = K$	ρ_i
1	Heat	3.85	1	0.060	6	0.419	1	0.433	1	1.000	1	1	1	1.00
2	Thunder	10.65	5	0.060	7	0.089	3	0.125	3	0.000	15	6	6	1.00
3	Spurs	12.07	7	0.053	8	0.022	8	0.022	7	0.000	3	4	4	1.00
4	Celtics	7.88	3	0.060	3	0.089	4	0.102	4	0.000	**18**	10	10	1.00
5	Clippers	15.16	11	0.042	9	0.015	9	0.015	9	0.000	6	9	2	1.00
6	Lakers	5.22	2	0.060	4	0.160	2	0.142	2	0.000	14	2	9	1.00
7	Pacers	16.91	14	0.025	**19**	0.003	**21**	0.003	**17**	0.000	5	3	3	1.00
8	76ers	14.07	9	0.038	12	0.008	11	0.008	11	0.000	**24**	11	11	1.00
9	Mavericks	11.16	6	0.060	5	0.063	5	0.035	6	0.000	**26**	5	5	1.00
10	Bulls	9.53	4	0.060	1	0.053	6	0.046	5	0.000	8	**18**	8	1.00
11	Knicks	12.56	8	0.060	2	0.022	7	0.021	8	0.000	**27**	8	**18**	1.00
12	Grizzlies	17.31	**17**	0.035	13	0.004	16	0.004	13	0.000	**25**	15	**26**	1.00
13	Nuggets	16.93	15	0.039	10	0.004	14	0.003	16	0.000	2	**14**	**20**	1.00
14	Magic	15.79	12	0.028	16	0.005	12	0.006	12	0.000	**30**	**26**	15	1.00
15	Hawks	16.44	13	0.030	15	0.003	**19**	0.003	15	0.000	**22**	**20**	14	1.00
16	Jazz	19.38	**26**	0.015	**26**	0.001	**26**	0.001	**26**	0.000	16	12	**27**	**0.00**
17	TrailBlazers	18.53	23	0.019	24	0.001	25	0.002	23	0.000	19	23	23	0.00
18	Rockets	14.76	10	0.038	11	0.012	10	0.010	10	0.000	29	7	13	0.53
19	Bucks	18.29	21	0.026	18	0.003	20	0.003	18	0.000	28	13	21	0.00
20	Suns	17.13	16	0.020	22	0.002	23	0.002	22	0.000	11	27	12	0.00
21	Nets	18.09	20	0.032	14	0.004	17	0.004	14	0.000	23	21	22	0.00
22	Warriors	18.46	22	0.024	20	0.004	13	0.003	19	0.000	13	22	7	0.00
23	Timberwolves	17.99	19	0.020	23	0.002	24	0.002	24	0.000	21	25	25	0.00
24	Hornets	19.78	28	0.009	28	0.000	29	0.000	29	0.000	7	28	28	0.00
25	Pistons	19.16	24	0.018	25	0.002	22	0.001	25	0.000	4	29	16	0.00
26	Kings	17.76	18	0.026	17	0.004	18	0.002	20	0.000	9	16	29	0.00
27	Wizards	19.32	25	0.008	29	0.001	28	0.000	28	0.000	10	19	19	0.00
28	Raptors	19.69	27	0.022	21	0.004	15	0.002	21	0.000	12	30	30	0.00
29	Cavaliers	20.18	29	0.012	27	0.001	27	0.001	27	0.000	17	24	24	0.00
30	Bobcats	20.93	30	0.003	30	0.000	30	0.000	30	0.000	20	30	30	0.00

Remark: (1) for AriM and the MC-based methods, the "Rank" was generated based on "Mean" or "π_i" with the Tie Method = random; (2) in CEMC, $d = F$ stands for CEMC$_F$, and $d = K$ stands for CEMC$_K$, both methods were applied under the default setting where base rankers are equally weighted; (3) BARD was applied under the default setting with hyperparameter $p = \frac{16}{30}$; (4) the errors in the aggregated rankings are highlighted in **bold**, and BARD made fewest errors among all tested methods.

Let I_i be the group indicator of entity $i \in U$, where $I_i = 1$ if $i \in U_R$, and $I_i = 0$ if $i \in U_B$. We make the following assumptions for base rankers $\tau_1, \ldots, \tau_m$:

- We have independent rankers. That is, given the indicator vector $I = \{I_i\}_{i \in U}$, the rankings $\tau_1, \ldots, \tau_m$ are mutually independent. A way to detect the violation of this assumption and a remedy of the method when the violation occurs will be discussed in Section 4.
- For each base ranker τ_k, the relative ranks of all background entities $\tau_k^0 \triangleq \tau_{k|U_B}$ are purely random (i.e., uniformly distributed);
- The relative rank of a relevant entity $i \in U_R$ among the background entities $\tau_k^{1|0}(i) \triangleq \tau_{k|\{i\}\cup U_B}(i)$ follows a power law distribution, i.e. $P(\tau_k^{1|0}(i) = t) \propto t^{-r_k}$, where a larger γ_k ($\gamma_k > 0$) means that ranker τ_k can better distinguish relevant entities from the background ones. Note that by requiring that $\gamma_k > 0$, we assume that each base ranker τ_k is making a good-faith effort for the common goal.
- Given $\tau_k^{1|0} \triangleq \{\tau_k^{1|0}(i)\}_{i \in U_R}$, the relative ranks of all relevant entities $\tau_k^1 \triangleq \tau_{k|U_R}$ is purely random (i.e., uniform).

Because the triplet $(\tau_k^0, \tau_k^{1|0}, \tau_k^1)$ gives an equivalent representation of the information in a full list τ_k when I is given (the equivalency is illustrated in Figure 1 with a toy example), the above assumptions lead to the following likelihood

U	I	τ_k		τ_k^0	$\tau_k^{1\|0}$	τ_k^1
E_1	1	2		-	2	1
E_2	1	3		-	2	2
E_3	1	5		-	3	3
E_4	0	1		1	-	-
E_5	0	4	$\Longleftrightarrow$	2	-	-
E_6	0	6		3	-	-
E_7	0	7		4	-	-
E_8	0	8		5	-	-
E_9	0	9		6	-	-
E_{10}	0	10		7	-	-

Figure 1. An equivalent representation of a full rank list τ_k via the triplet $(\tau_k^0, \tau_k^{1|0}, \tau_k^1)$, where τ_k^0 and τ_k^1 give the internal rankings of the background entities and relevant entities respectively, $\tau_k^{1|0}$ gives the relative rank of each relevant entity among background entities.

function:

$$P(\tau_1, \ldots, \tau_m \mid I, \gamma) = \prod_{k=1}^{m} P(\tau_k \mid I, \gamma_k) = \prod_{k=1}^{m} P\big(\tau_k^0, \tau_k^{1|0}, \tau_k^1 \mid I, \gamma_k\big) = \prod_{k=1}^{m} P\big(\tau_k^0 \mid I\big) \times P\big(\tau_k^{1|0} \mid I, \gamma_k\big) \times P\big(\tau_k^1 \mid \tau_k^{1|0}; I\big), \quad (1)$$

where $P(\tau_k^0 \mid I)$ and $P(\tau_k^1 \mid \tau_k^{1|0}; I)$ are uniform distributions on the corresponding spaces of allowable configurations, and

$$P\big(\tau_k^{1|0} \mid I, \gamma_k\big) = \prod_{i \in U_R} P\big(\tau_k^{1|0}(i) \mid I, \gamma_k\big) \quad \text{where} \quad P\big(\tau_k^{1|0}(i) = t \mid I, \gamma_k\big) \propto t^{-\gamma_k}. \quad (2)$$

In practice, however, both I and γ are unknown, and it is our main goal to estimate them from observations $\{\tau_1, \ldots, \tau_m\}$. Letting $\pi(I, \gamma)$ denote the prior distribution, we have the following posterior distribution of (I, γ):

$$P(I, \gamma \mid \tau_1, \ldots, \tau_m) \propto P(\tau_1, \ldots, \tau_m \mid I, \gamma)\pi(I, \gamma).$$

Since the marginal probability

$$\rho_i \triangleq P(I_i = 1 \mid \tau_1, \ldots, \tau_m) \quad (3)$$

is a good measurement of the importance of entity i, we generate the aggregated list as

$$\alpha = \text{sort}(i \in U \text{ by } \rho_i \downarrow). \quad (4)$$

On the other hand, the posterior mean

$$\overline{\gamma}_k \triangleq \int \gamma_k P(\gamma_k \mid \tau_1, \ldots, \tau_m) d\gamma_k \quad (5)$$

gives the estimation of the quality of base ranker τ_k.

The identifiability of the BARD model can be argued from the following intuition. If we have a large number m of independent rankers who generate rankings from the posited model, we will be able to observe a clear gap between the average rank of a relevant entity (across all rankers) and the average rank of a background entity, which provides us strong evidence to separate relevant entities from background ones. More precisely, using a method of moment we can consistently identify relevant entities as m goes to infinity. On the other hand, knowing the set of relevant entities in turn enables consistent estimations of the quality parameters γ as both numbers of relevant and background entities go to infinity. In a practical problem where m is of moderate size, having the quality information of different rankers becomes more useful for efficiently inferring relevant entities. Thus, inferring the entity indicators vector I and estimating the quality parameters γ can help each other.

3.2 Motivations and Intuitions Behind the Model

Compared with existing methods, BARD is unique in following aspects: (1) it partitions all the entities under ranking consideration into two groups: the relevant one U_R and the background one U_B, and ignores detailed rankings within each group (i.e., using the uniform distribution for τ_k^0 and τ_k^1); (2) it uses a power-law distribution to model the relative rank of a relevant entity among all background entities; and (3) it uses an explicit parameter associated with the power-law distribution to reflect the quality of a ranker. In this subsection, we explain how these features might help us better resolve the rank aggregation problem.

The partition of U into U_R and U_B is motivated by the observation that behind a ranking problem there is often a partitioning problem. For example, in the page-ranking problem, conceptually there is a binary answer for each web page whether or not it is truly relevant to a given search task (e.g., a group of key words). In grant review processes, there is always a binary decision for each proposal whether or not it should be funded. On the other hand, however, although ranking is often an intermediate step for decision processes that aim to partition entities into "selected" or "unselected" groups, the detailed ranking of each entity is still important to have in many problems (e.g., proposal rankings). In these cases our partition model is not faithful and somewhat limited. Fortunately, as shown by our simulations and real-data applications, under our partition model the inferred posterior probability for each entity to be in the "relevant group" serves as a good measure to rank the entity.

The power-law model $\tau_k^{1|0}(i)$ is a convenient approximation reasonably reflective of reality. In a real problem, the distribution of $\tau_k^{1|0}(i)$ depends on many factors and can take different forms in different problems. But we need to find a computationally feasible model for $\tau_k^{1|0}(i)$. We reason that it needs to satisfy the following simple requirement: it should give higher probability to a better rank for a relevant item and be no worse than assigning it a random (uniform) rank. That is, the probability function should be a monotone decreasing function. Two obvious choices are exponential or polynomial, of which polynomial is more robust and therefore chosen here. A large range of numerical investigations also support the adoption of the power law distribution. For example, we generated each ranker τ_k as the order of $\{X_{k,1}, \ldots, X_{k,n}\}$, that is,

$$\tau_k = \text{sort}(i \in U \text{ by } X_{k,i} \downarrow),$$

where $X_{k,i}$ is generated from two different distributions $F_{k,0}$ and $F_{k,1}$ via the following mechanism:

$$X_{k,i} \sim F_{k,0} \cdot I(i \in U_B) + F_{k,1} \cdot I(i \in U_R), \ \forall\, i \in U.$$

Figure 2 shows that the linear trend in the log-log plot of t versus $h(t) = P(\tau_k^{1|0}(i) = t \mid \tau_k^0; I, \gamma_k)$ is quite stable across different specifications of $F_{k,0}$ and $F_{k,1}$.

Third, by modeling τ_k^1 and τ_k^0 with the uniform distribution, BARD ignores the detailed information on the internal rankings within subset U_R and U_B, and only takes relative rankings between the two subsets into consideration. In other words, we choose to ignore all information in the data that is ancillary to the task of separating relevant entities from the background ones. This strategy greatly reduces the model complexity and computation burden while losses only marginal information in the data. In some scenarios, we can even argue that internal rankings within the background group are just noise, and thus, should be ignored to stabilize the analysis.

3.3 Details of the Bayesian Computation

Let $n_I = \sum_{i=1}^{n} I_i$ be the number of relevant entities defined by I. Recall that, for entity i with $I_i = 1$ (i.e., relevant entity), $\tau_k^{1|0}(i)$ denotes the relative rank of entity i among all the background

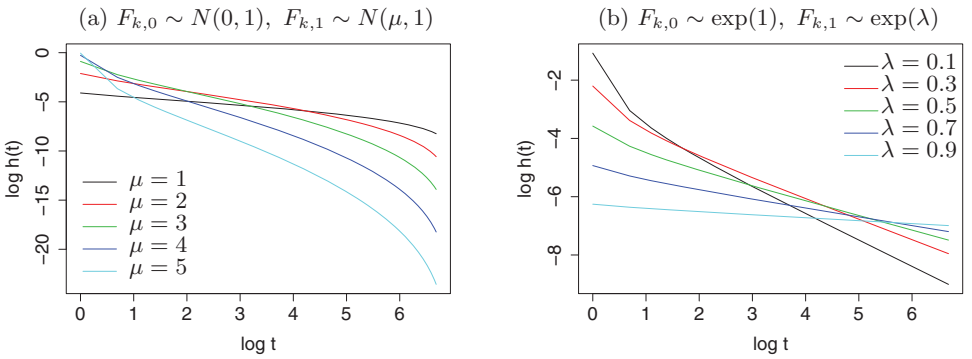


Figure 2. Log-log plots of relative rank $\tau_k^{1|0}(i) = t$ versus the corresponding probability $h(t) = P(\tau_k^{1|0}(i) = t \mid I, \gamma_k)$ under different scenarios. In each plot, we set $|U_B| = 1000$, thus the range of t is $\{1, \ldots, 1001\}$. The values of $h(t)$ are calculated via numerical integration.

entities and takes value in $\in \{1, 2, \ldots, n - n_I + 1\}$. According to model (2), we have

$$P\big(\tau_k^{1|0}(i) = t_i \mid I, \gamma_k\big) = \frac{t_i^{-\gamma_k}}{C(\gamma_k, n_I)},$$

and

$$P\big(\tau_k^{1|0} \mid I, \gamma_k\big) = \prod_{i \in U_R} P\big(\tau_k^{1|0}(i) \mid I, \gamma_k\big), \tag{6}$$

where the normalizing constant $C(\gamma_k, n_I) = \sum_{t=1}^{n-n_I+1} t^{-\gamma_k}$.

Let $\mathscr{A}_{U_R}$ be the space of all allowable rankings of entities in U_R. Let $\mathscr{A}_{U_R}(\tau_k^{1|0})$ be the configurations of τ_k^1 that are compatible with a given $\tau_k^{1|0}$. $\mathscr{A}_{U_R}(\tau_k^{1|0})$ is a subset of $\mathscr{A}_{U_R}$ due to constraints introduced by $\tau_k^{1|0}$. For example, given $\tau_k^{1|0} = (2, 2, 3)$ as shown in Figure 1, τ_k^1 has only two possible configurations: (1, 2, 3) or (2, 1, 3), since only the relative position of the first two entities E_1 and E_2 is not fixed given $\tau_k^{1|0}$. In general, we have the following assignment based on the "purely random assumption":

$$P\big(\tau_k^1 = \tau \mid \tau_k^{1|0}; I\big) = \frac{1}{\prod_{t=1}^{n-n_I+1} n_{\tau_k,t}^{1|0}!} \cdot I\big(\tau \in \mathscr{A}_{U_R}(\tau_k^{1|0})\big), \tag{7}$$

where $n_{\tau_k,t}^{1|0} = \sum_{i \in U_R} I(\tau_k^{1|0}(i) = t)$.

Note that the relative rank order of all background entities, τ_k^0, follows the uniform distribution, that is,

$$P\big(\tau_k^0 = \tau \mid I\big) = \frac{1}{(n - n_I)!}. \tag{8}$$

Putting (8), (6), and (7) together, we have

$$\begin{aligned} &P(\tau_k \mid I, \gamma_k) \\ &\quad = P\big(\tau_k^0 \mid I\big) \times P\big(\tau_k^{1|0} \mid I, \gamma_k\big) \times P\big(\tau_k^1 \mid \tau_k^{1|0}; I\big) \\ &\quad = \{(n - n_I)! \times A_{\tau_k,I} \times (C(\gamma_k, n_I))^{n_I} \times (B_{\tau_k,I})^{\gamma_k}\}^{-1}, \end{aligned}$$

where

$$A_{\tau_k,I} \triangleq \prod_{t=1}^{n-n_I+1} \big(n_{\tau_k,t}^{1|0}!\big) \text{ and } B_{\tau_k,I} \triangleq \prod_{i \in U_R} \tau_k^{1|0}(i).$$

Thus, the conditional probability of $\{\tau_1, \ldots, \tau_m\}$ is

$$\begin{aligned} P(\tau_1, \ldots, \tau_m \mid I, \gamma) = [(n - n_I)!]^{-m} \times \prod_{k=1}^{m} \{A_{\tau_k,I} \\ \times (C(\gamma_k, n_I))^{n_I} \times (B_{\tau_k,I})^{\gamma_k}\}^{-1}. \end{aligned} \tag{9}$$

We give I an informative prior

$$\pi(I) \propto \exp^{-\frac{(n_I - n \cdot p)^2}{2\sigma^2}},$$

where p is the hyperparameter representing the expected percentage of relevant entities in U, and set σ^2 as a tunable hyperparameter (whose default value is $\sigma^2 = \frac{1}{\sqrt{m}}$). We let $\{\gamma_k\}_{1 \le k \le m}$ have an independent exponential prior, that is, $\pi(\gamma) = \prod_{1 \le k \le m} f(\gamma_k)$, where $f(\gamma_k) = \lambda e^{-\lambda \gamma_k}$, λ is the mean of the exponential distribution. In BARD, we use $\lambda = 1$ as the default setting, and allow the user to specify the value of λ based on their own judgment for a practical problem. We also tested using a uniform prior in hypercube $[0, 10]^m$ for γ, which resulted in a very similar performance to the exponential prior.

Given the above prior distributions, we get the joint posterior distribution of (I, γ):

$$\begin{aligned} &P(I, \gamma \mid \tau_1, \ldots, \tau_m) \\ &\quad \propto \pi(I)\pi(\gamma)P(\tau_1, \ldots, \tau_m \mid I, \gamma) \\ &\quad = \frac{\pi(I)}{[(n - n_I)!]^m} \cdot \prod_{k=1}^{m} \frac{f(\gamma_k)}{A_{\tau_k,I} \times (C(\gamma_k, n_I))^{n_I} \times (B_{\tau_k,I})^{\gamma_k}}, \end{aligned} \tag{10}$$

which induces the following conditional distributions:

$$\begin{aligned} &P(\gamma_k \mid \tau_1, \ldots, \tau_m; I, \gamma_{[-k]}) \\ &\quad = P(\gamma_k \mid \tau_k; I) \\ &\quad \propto e^{-\lambda\gamma_k} \times (C(\gamma_k, n_I))^{-n_I} \times (B_{\tau_k,I})^{-\gamma_k}, \end{aligned} \tag{11}$$

$$P(I_i \mid \tau_1, \ldots, \tau_m; I_{[-i]}, \gamma) \sim \text{Bernoulli}\left(\frac{q_i(\gamma)}{q_i(\gamma) + 1}\right), \tag{12}$$

where

$$q_i(\gamma) = \frac{\pi(I_{[I_i=1]})}{\pi(I_{[I_i=0]})} \cdot \prod_{k=1}^{m} \frac{P(\tau_k \mid I_{[I_i=1]}, \gamma_k)}{P(\tau_k \mid I_{[I_i=0]}, \gamma_k)}.$$

These distributions enable us to draw samples from $P(I, \gamma \mid \tau_1, \ldots, \tau_m)$ via Gibbs sampling. The posterior probabilities, $P(I_i \mid \tau_1, \ldots, \tau_m)$ and $P(\gamma_k \mid \tau_1, \ldots, \tau_m)$, can be obtained from the Monte Carlo samples and used to generate the aggregated rank list and reliability measures of base rankers. Since the conditional distribution shown in (11) is not a standard distribution, we use the random-walk Metropolis algorithm to draw samples from it (see Liu 2001 for a comprehensive review).

3.4 Extensions to Partial Lists and Supervised Rank Aggregation

Since a partial list can be viewed as an incomplete version of a full list, the aggregation of partial lists can be treated as a missing data problem and solved via data augmentation strategies (Tanner and Wong 1987). To be precise, we let $\{\tau_1^P, \ldots, \tau_m^P\}$ be the m partial lists of interest, and let $\{\tau_1^*, \ldots, \tau_m^*\}$ be their unobserved underlying full lists. We are interested in drawing samples from the following target distribution

$$P\left(I, \gamma \mid \tau_1^P, \ldots, \tau_m^P\right) \propto \pi(I)\pi(\gamma)P\left(\tau_1^P, \ldots, \tau_m^P \mid I, \gamma\right),$$

which can be achieved via Gibbs sampling based on the following conditional distributions:

$$P\left(\tau_1^*, \ldots, \tau_m^* \mid \tau_1^P, \ldots, \tau_m^P; I, \gamma\right) = \prod_{k=1}^{m} P\left(\tau_k^* \mid \tau_k^P; I, \gamma_k\right),$$

$$P(I, \gamma \mid \tau_1^*, \ldots, \tau_m^*) \propto \pi(I)\pi(\gamma)\prod_{k=1}^{m} P(\tau_k^* \mid I, \gamma_k).$$

Given that the distribution $P(I, \gamma \mid \tau_1^*, \ldots, \tau_m^*)$ has been analyzed in the previous section, we only need to focus on $P(\tau_1^*, \ldots, \tau_m^* \mid \tau_1^P, \ldots, \tau_m^P; I, \gamma)$, or more concretely, $P(\tau_k^* \mid \tau_k^P; I, \gamma_k)$ here. Let Ω_k be the set of full lists that are compatible with τ_k^P, we have

$$P\left(\tau_k^* \mid \tau_k^P; I, \gamma_k\right) \propto P(\tau_k^* \mid I, \gamma_k) \cdot I(\tau_k^* \in \Omega_k).$$

Again, we can use random walk Metropolis algorithm to draw samples from this distribution.

BARD can also be applied to the scenario where training data are available. Let $\{\tau_1, \ldots, \tau_m\}$ be the m lists (full or partial) of interest, and $i_1 \preccurlyeq i_2 \preccurlyeq \cdots \preccurlyeq i_s$ be the training information, which gives the true relative rank of s entities $\{i_1, i_2, \ldots, i_s\}$ in U. In BARD, a natural way to make use of the training information is to put constraints on I with respect to $i_1 \preccurlyeq i_2 \preccurlyeq \cdots \preccurlyeq i_s$, that is, if $I_{i_t} = 1$, then $I_{i_{t'}} = 1$ for all $t' \le t$. Incorporating the training data into the analysis may help BARD better estimate the quality parameters $\{\gamma_k\}_{1 \le k \le m}$ of the m base rankers, and thus, improve the final results.

4. MODEL DIAGNOSTICS AND REMEDIES

4.1 Detecting Violation of the Independence Assumption

Although we will show in Section 5 that BARD is reasonably robust to the violation of the "independent rankers" assumption, it is desirable to detect a severe violation of the assumption and further improve BARD based on this information. Standard correlation measures such as the Spearman and the Kendall correlations do not work here because any pair of informative rankings are unconditionally correlated since they are supposed to capture the same signal about entity relevancy. This type of correlation is not what we are interested in. Instead, we wish to detect groups of rankings that are "over-correlated" relative to their quality levels.

Consider all the ranks entity i received from all the rankers, $\{\tau_1(i), \ldots, \tau_m(i)\}$. It forms a natural distribution on the rank space $\{1, \ldots, n\}$, denoted as Q_i. If entity i has a strong positive/negative signal, a significant proportion of the rankers would give it a high/low rank, so that Q_i skews toward the left/right tail; if entity i belongs to the background, Q_i should be close to be uniform. To capture these key features of Q_i, we fit Q_i with a rescaled Beta distribution:

$$Q_i(t) \propto \text{dBeta}\left(\frac{t}{n+1}; \alpha_i, \beta_i\right) \cdot I(t \in \{1, 2, \ldots, n\}),$$

where $\text{dBeta}(x; \alpha, \beta)$ is the density of the Beta distribution with parameters (α, β). Assuming that $\{\frac{\tau_1(i)}{n+1}, \ldots, \frac{\tau_m(i)}{n+1}\}$ are iid draws from distribution $\text{Beta}(\alpha_i, \beta_i)$, we denote the estimated parameters as $(\hat{\alpha}_i, \hat{\beta}_i)$, and the fitted distribution as $Q(\hat{\alpha}_i, \hat{\beta}_i)$ ($\hat{Q}_i$ for short).

For any pair of base rankers τ_{j_1} and τ_{j_2}, without loss of generality, we assume that $\tau_{j_1}(i) \le \tau_{j_2}(i)$. Given the fitted Beta distribution $\hat{Q}_i$, we use the quantity below to measure excessive correlatedness of them at entity i:

$$V_{j_1 j_2}^{(i)} \triangleq \sum_{\tau_{j_1}(i) \le t \le \tau_{j_2}(i)} Q(t; \hat{\alpha}_i, \hat{\beta}_i).$$

Intuitively, $V_{j_1 j_2}^{(i)}$ corresponds to the probability that a random sample from $\hat{Q}_i$ falls into the interval $[\tau_{j1}(i), \tau_{j2}(i)]$. A smaller $V_{j_1 j_2}^{(i)}$ means a smaller probability that the two independent rankers agree with each other by chance at entity i, hence a stronger evidence of nonindependence. Note that $V_{j_1 j_2}^{(i)}$ accounts for not only the distance between $\tau_{j_1}(i)$ and $\tau_{j_2}(i)$, but also their relative probabilities based on $\hat{Q}_i$. We can estimate the p-value $P_{j_1 j_2}^{(i)} \triangleq P(V_{xy} < V_{j_1 j_2}^{(i)})$ using Monte Carlo simulation, and summarize the overall evidence for the pair of rankers by the *coordination coefficient*:

$$\zeta_{j_1 j_2} \triangleq -\frac{1}{n}\sum_{i=1}^{n} \log P_{j_1 j_2}^{(i)}.$$

A larger $\zeta_{j_1 j_2}$ means that rankers τ_{j_1} and τ_{j_2} are "over-correlated." Alternatively, we can use the method of posterior predictive checking (Rubin 1984) to generate the Bayesian *coordination coefficient*, which will be computationally more demanding.

Under the null hypothesis that the two rankers are independent, we have by the central limit theorem that $\zeta_{j_1 j_2}$ follows $N(1, \frac{1}{n})$ approximately, which can be used to set a threshold for $\zeta_{j_1 j_2}$ to claim that τ_{j_1} and τ_{j_2} are not independent. The procedure for discovering correlated rankings can be summarized as follows:

- For each entity $i \in U$, fit a rescaled Beta distribution $\hat{Q}_i$ for $\{\tau_1(i), \ldots, \tau_m(i)\}$;
- For each ranker pair τ_{j_1} and τ_{j_2}, calculate the coordination coefficient $\zeta_{j_1 j_2}$ based on $\{\hat{Q}_i\}_{i \in U}$;
- If $\zeta_{j_1 j_2}$ is larger than a threshold (e.g., significance level 0.05 with Bonfferoni correction), we say that τ_{j_1} and τ_{j_2} belong to a "block" of correlated rankers.

4.2 A Hierarchical Model for the Correlated Base Rankers

Once the underlying correlation structure among the rankers are detected, we can modify BARD to avoid the negative impact of the correlation. Assume that the correlated base rankers fall into M blocks $\{G_1, \ldots, G_M\}$, where the rankers within a block are highly correlated while the rankers from different blocks are conditionally independent given the entity membership I.

Let G_0 be all the other conditionally independent rankers. To simplify the problem, we assume that every base ranker provides a complete ranking list in this article. The more general scenario involving partial lists can be solved based on a similar principle.

Let κ_j be the representative ranker of group G_j, and let $\gamma_j > 0$ denote the quality measure of the ranker block G_j. We modify the BARD model into the following hierarchical form:

$$P(\kappa_j \mid I, \gamma_j) = P\big(\kappa_j^0 \mid I\big) P\big(\kappa_j^{1|0} \mid I, \gamma_j\big) P\big(\kappa_j^1 \mid I, \kappa_j^{1|0}\big),$$
$$P(\tau_k \mid \kappa_j, \beta_j) \propto \exp\left\{-\frac{\beta_j}{|G_j|} \cdot d(\tau_k, \kappa_j)\right\}.$$

where $\beta_j > 0$ measures the average magnitude of correlation between κ_j and base rankers in group G_j, $|G_j|$ is the number of rankers in group G_j, and $d(\tau_k, \kappa_j)$ is the Spearman's footrule distance or Kendall tau distance between τ_k and κ_j. The joint likelihood can be written as

$$P(\kappa_1, \ldots, \kappa_M; \tau_1, \ldots, \tau_m \mid I, \gamma_j) = \prod_{k \in G_0} P(\tau_k \mid I, \gamma_k) \cdot \prod_{j=1}^{M} \left[P(\kappa_j \mid I, \gamma_j) \prod_{k \in G_j} P(\tau_k \mid \kappa_j) \right].$$

In words, the model assumes that the base rankers within each block G_j are conditionally independent of each other given the common ranker κ_j.

Given the prior distribution

$$\pi(I, \gamma, \beta) = \pi(I) \prod_{j=1}^{M} \pi(\gamma_j)\pi(\beta_j),$$

the joint posterior distribution is

$$P(I, \gamma, \beta \mid \tau_1, \ldots, \tau_m) \propto \pi(I, \gamma, \beta) P(\tau_1, \ldots, \tau_M \mid I, \gamma, \beta).$$

An MCMC sampler for simulating from this distribution can be implemented based on the following conditional distributions:

$$\begin{aligned} &P(\kappa_j \mid I, \gamma_j, \beta_j, \{\tau_k\}_{k \in G_j}) \\ &\quad \propto P(\kappa_j \mid I, \gamma_j) \prod_{k \in G_j} P(\tau_k \mid \kappa_j) \\ &\quad = P\big(\kappa_j^0 \mid I\big) P\big(\kappa_j^{1|0} \mid I, \gamma_j\big) P\big(\kappa_j^1 \mid I, \kappa_j^{1|0}\big) \\ &\qquad \times \exp\left\{-\frac{\beta_j}{m_j} \sum_{k \in G_j} d(\tau_k, \kappa_j)\right\}, \\ &P(\beta_j \mid \kappa_j, \{\tau_k\}_{k \in G_j}) \\ &\quad \propto \pi(\beta_j) \exp\left\{-\frac{\beta_j}{m_j} \sum_{k \in G_j} d(\tau_k, \kappa_j)\right\}; \text{ and} \\ &P(I, \gamma \mid \kappa_1, \ldots, \kappa_M) \\ &\quad \propto \pi(I)\pi(\gamma) \prod_{j=1}^{M} P(\kappa_j \mid I, \gamma). \end{aligned}$$

A random walk Metropolis algorithm can be used to sample from $P(\kappa_j \mid I, \gamma_j, \beta_j, \{\tau_k\}_{k \in G_j})$. With a noninformative prior for β_j, $P(\beta_j \mid \kappa_j, \{\tau_k\}_{k \in G_j})$ becomes an exponential distribution. Sampling from distribution $P(I, \gamma \mid \kappa_1, \ldots, \kappa_M)$ can be achieved by the technique developed in Section 3. We use BARD_{HM} to denote this modification of BARD with hierarchical model.

We note here that a full Bayesian approach that simultaneously detects correlation block structures and infers parameters in the hierarchical model can be designed rather straightforwardly based on the methods described in this section. However, the computational cost for inferring the full hierarchical model is often too high, and the following two-step approximation strategy already works very effectively: (1) for each block G_j, generate its representative ranker κ_j from $\{\tau_k\}_{k \in G_j}$ by a simple method (e.g., AriM, MC_1, or CEMC); (2) apply ordinary BARD to $\{\kappa_j\}_{j=1}^m \cup \{\tau_k\}_{k \in G_0}$.

5. SIMULATION STUDIES

5.1 Simulation Under the BARD Model

Let $U = \{1, \ldots, n\}$, of which the first 10% are the relevant entities (i.e., $U_R = \{1, \ldots, [n/10]\}$). We generate the base rankers $\{\tau_k\}_{1 \le k \le m}$ via the following scheme:

$$\tau_k = \text{sort}(i \in U \text{ by } X_{k,i} \downarrow), \text{ where } X_{k,i} \sim N(0, 1) \cdot I(i \in U_B) + N(\mu_k, 1) \cdot I(i \in U_R).$$

We examine two scenarios: (*A*) $\mu_k = \mu$ for all k, and (*B*) $\mu_k = \mu \cdot I(k \le \frac{m}{2})$. In scenario *A*, the base rankers are equally reliable; in scenario *B*, however, only the first 50% base rankers are informative. The parameter μ controls the signal strength of the dataset (a larger μ means that we have more information to distinguish relevant entities from irrelevant ones). We generate both full lists and top-d lists ($d = 0.2 \cdot n$) for each scenario and test four cases: full lists from scenario *A* (denoted as A_F), top-d lists from scenario *A* (denoted as A_P), full lists from scenario *B* (denoted as B_F), and top-20 lists from scenario *B* (denoted as B_P).

We first evaluate the impact of signal strength μ on the performance of BARD. Fixing $n = 100$ and $m = 10$, we tried four different values of μ ($\mu = 0.5,\ 1.0,\ 1.5$, and 2.0) for each of the above four cases. Under each configuration, 1000 independent datasets were simulated. To each dataset, we applied three naive methods (AriM, GeoM, MedR), four Markov-chain-based methods (MC_1, MC_2, MC_3, MC_4), two optimization-based methods (CEMC_F and CEMC_K), and BARD with $\lambda = 1$ under three different choices of the hyperparameter p ($p_1 = 0.05$, $p_2 = 0.10$, and $p_3 = 0.15$), respectively. Additionally, we include BARD with the constraint of equal quality, that is, $\gamma_1 = \cdots = \gamma_m$ (denoted by BARD_C), in the comparison. For each method, its average coverage rate across the 1000 parallel experiments under different configurations is calculated to evaluate the performance. (The coverage rate of an aggregated list is defined as the percentage of true relevant entities covered by the top-10 entities.)

The results are summarized in Table 3, from which we can see that: (1) when qualities of base rankers were the same (i.e., scenario *A*), BARD_C slightly outperformed BARD and achieved a similar performance as CEMC, which is supposed to be "optimal" in this case; (2) when the quality of base rankers varied greatly (i.e., scenario *B*), BARD uniformly outperformed all other methods, and the benefit increased with the increase of the signal strength μ; (3) both BARD_C and BARD were robust to the choice of the hyperparameter p. Figure 3 displays boxplots of $\{\bar{\gamma}_k\}_k$ obtained by BARD from the 1000 parallel runs under different configurations, suggesting that BARD was capable of

Table 3. Average coverage rates of different rank aggregation methods

Configuration				Naive methods			MC-based methods				CEMC		BARD_C			BARD		
Case	m	n	μ	AriM	GeoM	MedR	MC_1	MC_2	MC_3	MC_4	$d = F$	$d = K$	p_1	p_2	p_3	p_1	p_2	p_3
A_F	10	100	0.5	0.48	0.47	0.43	0.11	0.45	0.48	0.28	0.46	0.47	0.46	0.44	0.43	0.41	0.40	0.40
A_F	10	100	1.0	0.83	0.82	0.77	0.15	0.79	0.84	0.27	0.78	0.80	0.81	0.80	0.78	0.75	0.74	0.73
A_F	10	100	1.5	0.98	0.98	0.94	0.32	0.95	0.98	0.25	0.89	0.90	0.95	0.96	0.96	0.94	0.95	0.94
A_F	10	100	2.0	1.00	1.00	1.00	0.71	0.99	1.00	0.25	0.93	0.93	0.98	1.00	0.99	0.98	1.00	0.99
A_P	10	100	0.5	0.10	0.13	0.14	0.16	0.15	0.08	0.05	0.38	0.41	0.40	0.41	0.38	0.36	0.37	0.37
A_P	10	100	1.0	0.14	0.22	0.22	0.29	0.27	0.18	0.08	0.72	0.75	0.74	0.73	0.72	0.67	0.68	0.69
A_P	10	100	1.5	0.25	0.37	0.36	0.55	0.54	0.44	0.20	0.91	0.91	0.92	0.92	0.92	0.89	0.89	0.90
A_P	10	100	2.0	0.44	0.58	0.54	0.80	0.80	0.73	0.38	0.96	0.96	0.97	0.99	0.99	0.96	0.98	0.98
B_F	10	100	0.5	0.26	0.26	0.24	0.10	0.26	0.26	0.19	0.25	0.25	0.25	0.25	0.24	0.24	0.24	0.24
B_F	10	100	1.0	0.45	0.49	0.42	0.11	0.48	0.46	0.26	0.45	0.45	0.47	0.46	0.44	0.51	0.51	0.50
B_F	10	100	1.5	0.63	0.70	0.61	0.12	0.70	0.63	0.29	0.63	0.62	0.67	0.65	0.63	0.79	0.79	0.78
B_F	10	100	2.0	0.74	0.84	0.74	0.12	0.84	0.75	0.29	0.74	0.73	0.81	0.78	0.74	0.93	0.94	0.93
B_P	10	100	0.5	0.11	0.13	0.13	0.12	0.13	0.08	0.06	0.23	0.24	0.24	0.24	0.24	0.22	0.23	0.23
B_P	10	100	1.0	0.13	0.17	0.17	0.17	0.17	0.09	0.06	0.43	0.45	0.44	0.43	0.41	0.45	0.45	0.45
B_P	10	100	1.5	0.16	0.21	0.23	0.24	0.23	0.15	0.08	0.63	0.65	0.65	0.63	0.61	0.71	0.72	0.71
B_P	10	100	2.0	0.19	0.26	0.28	0.29	0.28	0.20	0.10	0.80	0.80	0.80	0.78	0.75	0.88	0.88	0.88

Remark: (1) in CEMC, $d = F$ stands for CEMC_F, and $d = K$ stands for CEMC_K; (2) BARD_C stands for BARD with constraint that $\gamma_1 = \cdots = \gamma_m$; (3) for both BARD_C and BARD, we tried three values for hyperparameter p, that is, $p_1 = 0.05$, $p_2 = 0.10$, and $p_3 = 0.15$ with hyperparameter $\lambda = 1$.

efficiently estimating qualities of base rankers when the signal strength was reasonably large (e.g., $\delta \geq 1.0$). We also applied BARD and BARD_C with $\lambda = 2$ to each of the simulated dataset and obtained very consistent results, indicting that BARD is robust to the specification of hyperparameter λ.

We next check the impact of the data size (i.e., number of entities n and the number of rankers m) on the performance of BARD. We fixed the signal strength $\mu = 1.0$, and tried two alternative combinations: $(m, n) = (10, 200)$ and $(m, n) = (20, 100)$. The results are summarized into Table 4, from which we can see that most of methods tested were not sensitive to the increase of n, although an increase of m led to better performances for most methods. More importantly, BARD performed quite robustly to different choices of n and m compared to the other methods.

5.2 Robustness of BARD

An important assumption in our model is that the rankers in consideration work independently, which can often be violated in real problems. To test how well our method tolerates the violation of this assumption, we simulated 20 rankings $\{\tau_1, \ldots, \tau_{20}\}$ falling into three groups:

$$G_1 = \{\tau_1, \tau_2, \tau_3, \tau_4\},\ G_2 = \{\tau_5, \tau_6, \tau_7, \tau_8\},\ \text{and}$$
$$G_0 = \{\tau_9, \ldots, \tau_{20}\},$$

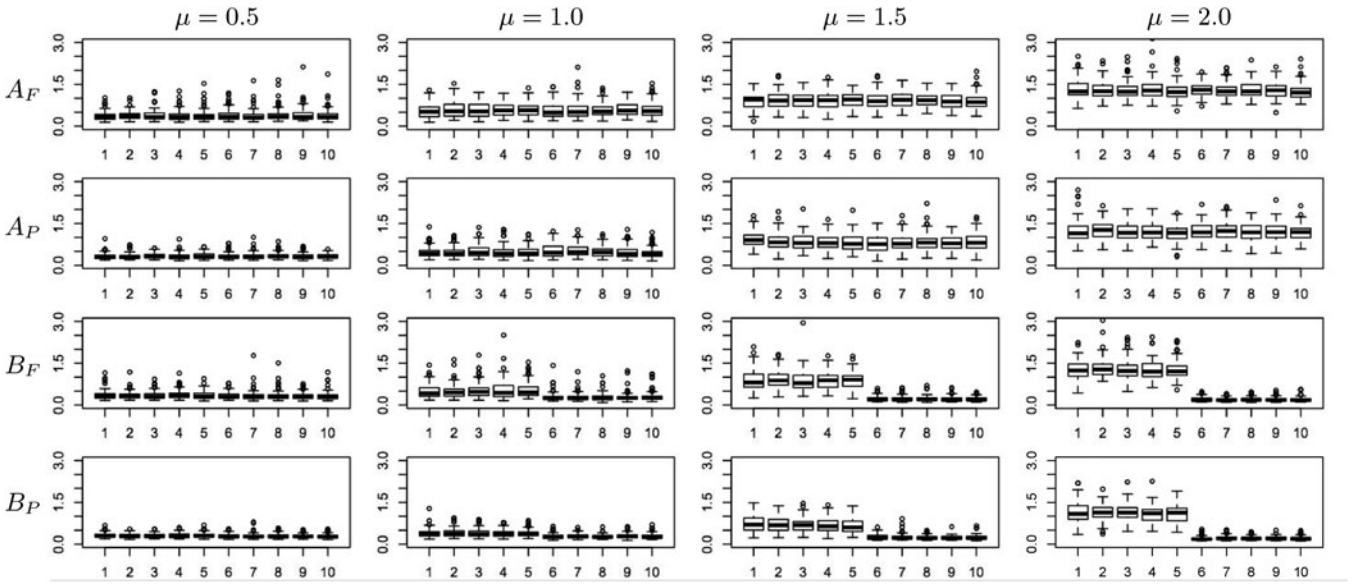


Figure 3. The boxplots of $\{\bar{\gamma}_k\}_k$ estimated by BARD from 1000 parallel runs under different configurations when $m = 10$ and $n = 100$ with hyperparameters $p = 0.1$ and $\lambda = 1$.

Table 4. Impact of data size to the performances of different methods

Configuration				Naive methods			MC-based methods				CEMC		BARD_C			BARD		
Case	m	n	μ	AriM	GeoM	MedR	MC_1	MC_2	MC_3	MC_4	$d = F$	$d = K$	p_1	p_2	p_3	p_1	p_2	p_3
A_F	10	200	1.0	0.84	0.83	0.77	0.18	0.79	0.84	0.33	0.83	0.83	0.81	0.80	0.78	0.77	0.76	0.75
A_F	10	100	1.0	0.84	0.82	0.78	0.15	0.79	0.84	0.26	0.83	0.84	0.82	0.79	0.78	0.74	0.73	0.73
A_F	20	100	1.0	0.96	0.96	0.91	0.10	0.95	0.96	0.32	0.96	0.96	0.94	0.94	0.93	0.89	0.91	0.89
A_P	10	200	1.0	0.13	0.22	0.21	0.35	0.32	0.22	0.09	0.74	0.75	0.73	0.73	0.72	0.68	0.69	0.69
A_P	10	100	1.0	0.13	0.20	0.21	0.23	0.23	0.14	0.06	0.74	0.74	0.73	0.73	0.72	0.68	0.69	0.69
A_P	20	100	1.0	0.17	0.24	0.22	0.64	0.58	0.52	0.20	0.90	0.92	0.90	0.89	0.87	0.84	0.84	0.83
B_F	10	200	1.0	0.46	0.51	0.42	0.11	0.50	0.47	0.27	0.44	0.46	0.49	0.48	0.46	0.55	0.55	0.54
B_F	10	100	1.0	0.46	0.51	0.42	0.11	0.50	0.47	0.29	0.46	0.46	0.49	0.46	0.45	0.51	0.50	0.49
B_F	20	100	1.0	0.63	0.67	0.57	0.10	0.67	0.64	0.33	0.63	0.64	0.63	0.60	0.54	0.69	0.68	0.66
B_P	10	200	1.0	0.13	0.18	0.18	0.17	0.15	0.08	0.05	0.43	0.44	0.45	0.44	0.41	0.47	0.48	0.46
B_P	10	100	1.0	0.13	0.17	0.17	0.15	0.13	0.08	0.05	0.43	0.44	0.43	0.43	0.41	0.46	0.47	0.45
B_P	20	100	1.0	0.15	0.17	0.18	0.43	0.39	0.34	0.15	0.61	0.61	0.59	0.56	0.42	0.62	0.61	0.57

where rankings in G_0 are independently generated, and rankings in G_1 and G_2 have very strong within group correlation. More precisely, we let $U = \{1, \ldots, 100\}$, let the relevant entities be $U_R = \{1, \ldots, 10\}$, and let the background entities be composed of two subsets: the "neutral" set $U_{B_1} = \{11, \ldots, 90\}$ and the "negative" set $U_{B_2} = \{91, \ldots, 100\}$. We define $\delta_i = I(i \in U_R) - I(i \in U_I)$, implying that $\delta_i = 1$ for $i \in U_R$, 0 for $i \in U_{B_1}$, and -1 for $i \in U_{B_2}$.

- A ranking τ_k in G_0 was simulated as $\tau_k = \text{sort}(i \in U \text{ by } X_{k,i} \downarrow)$, where $X_{k,i} \sim N(\delta_i \cdot \mu_k, 1)$ with $\mu_k \geq 0$. A larger μ_k means that τ_k can better separate relevant entities from background ones, thus a higher quality ranking;
- The rankings in G_1 and G_2 were generated via two steps: we first generated a common ranking

$$\kappa = \text{sort}(i \in U \text{ by } X_i \downarrow) \text{ where } X_i \sim N(\delta_i \cdot \mu, 1),$$

and then manipulated κ with random transpositions to generate a group of correlated rankings. Let $\mathcal{M}(\cdot)$ denote a random transposition operation. The aforementioned manipulation can be written as $\tau_k = \mathcal{M}^s(\tau)$ where s is number of such operations used. Note that a small s indicates a stronger correlation among the rankings.

Fixing $\mu = \mu_9 = \cdots = \mu_{12} = 0.5$, $\mu_{13} = \cdots = \mu_{16} = 1.0$, $\mu_{17} = \cdots = \mu_{20} = 1.5$, we simulated 1000 datasets for each of the following three configurations corresponding to $s = 20$, 60, and 100, respectively. Table 5 shows a typical dataset simulated with $s = 60$, from which we can see that the rankings within G_1 or G_2 are quite similar to each other for many entities. We applied BARD, BARD_{HM} as well as other methods to each of these simulated datasets. The results are summarized in Table 6 and Figure 4. From Table 6, we can see that: (1) BARD_{HM} uniformly outperformed all other methods; (2) BARD performed reasonably well even when correlations among the rankers within G_1 and G_2 are very strong (i.e., $s = 20$), and approached the performance of BARD_{HM} when the correlation was weaker (i.e., $s = 60$ or 100). These results are consistent with the information

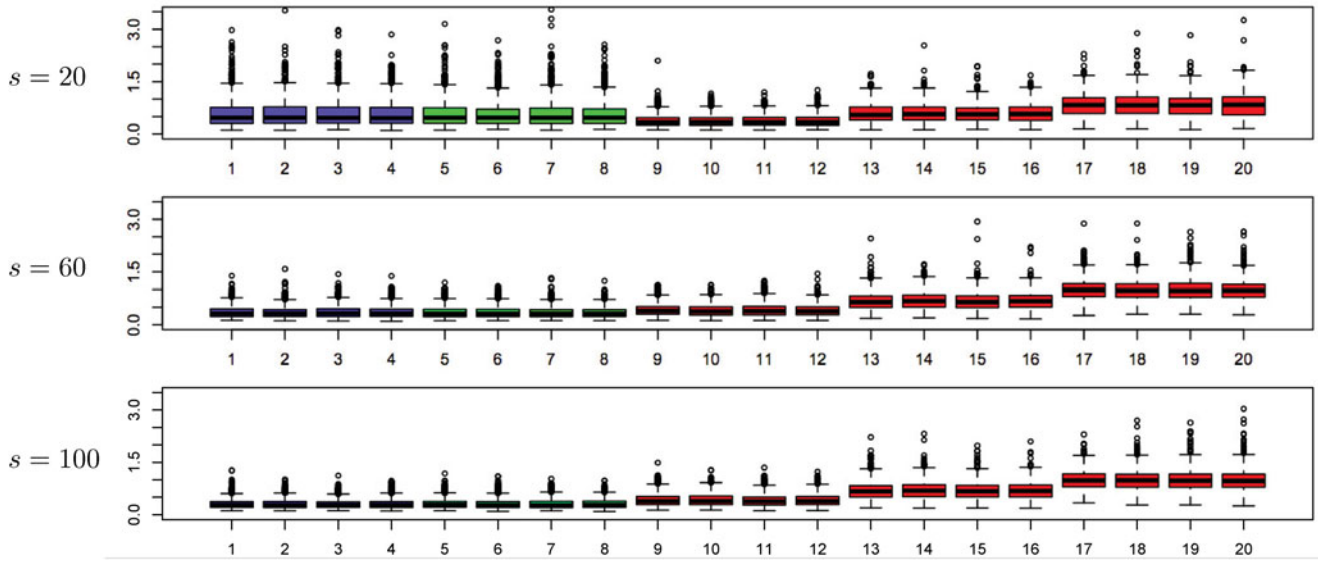


Figure 4. Boxplots of $\{\bar{\gamma}_k\}_k$ estimated by BARD from 1000 parallel runs when some base rankers are dependent of each other. The datasets are simulated from the mechanism described in Section 5.2, where the 20 rankers belongs to three blocks $G_1 = \{\tau_1, \ldots, \tau_4\}$, $G_2 = \{\tau_5, \ldots, \tau_8\}$ and $G_0 = \{\tau_9, \ldots, \tau_{20}\}$.

Table 5. A typical simulated dataset for testing the robustness of BARD

	G_1				G_2				G_0											
Entity	τ_1	τ_2	τ_3	τ_4	τ_5	τ_6	τ_7	τ_8	τ_9	τ_{10}	τ_{11}	τ_{12}	τ_{13}	τ_{14}	τ_{15}	τ_{16}	τ_{17}	τ_{18}	τ_{19}	τ_{20}
1	57	57	45	27	56	63	56	31	82	4	5	53	22	4	1	69	44	10	6	29
2	14	100	14	15	31	56	67	56	70	42	89	4	15	11	29	2	1	26	28	7
3	27	55	55	87	94	1	5	1	34	89	36	80	12	9	58	35	22	16	78	12
4	4	17	28	14	90	90	86	90	20	5	2	36	6	2	43	11	21	51	59	16
5	55	85	4	55	86	24	90	50	15	63	32	2	21	36	48	23	20	13	21	31
6	5	5	5	75	49	49	32	99	48	22	53	78	13	6	45	17	58	49	1	60
7	73	15	99	25	17	53	13	73	21	67	19	22	1	46	4	19	3	3	16	10
8	31	52	53	57	76	26	17	17	57	83	23	68	3	1	21	76	8	2	30	18
9	22	92	87	10	73	17	26	72	27	30	3	74	16	77	2	1	10	7	4	2
10	62	10	77	77	7	76	76	29	8	18	63	66	32	20	5	91	41	21	5	34
11	70	24	86	24	6	6	16	26	64	38	66	33	47	56	92	36	39	56	45	62
12	41	88	24	86	34	42	6	22	38	58	49	97	36	40	14	55	54	53	81	33
13	8	76	82	34	30	66	3	89	59	72	38	40	25	43	76	26	86	61	15	54
14	25	68	31	42	11	29	11	30	73	68	100	25	94	92	40	46	59	92	32	43
15	79	82	76	82	81	83	48	94	32	12	46	52	68	96	50	59	81	69	10	55
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
86	78	19	50	78	66	41	63	42	87	41	27	92	100	18	20	33	11	88	50	39
87	83	71	19	51	63	7	66	63	9	8	21	15	29	21	88	27	43	23	22	8
88	19	95	10	8	26	30	39	7	30	76	45	26	14	66	62	5	36	65	63	52
89	13	8	78	83	42	44	45	44	96	57	70	95	48	75	10	24	6	45	39	35
90	42	42	34	4	69	69	69	69	100	95	96	29	27	41	61	87	17	93	46	44
91	9	87	1	76	89	14	85	11	75	100	59	58	81	99	57	49	69	82	47	83
92	94	94	72	94	96	87	89	87	83	29	58	82	97	47	98	86	73	100	86	49
93	72	72	12	72	80	85	50	3	36	36	6	69	54	42	65	74	25	85	95	92
94	23	22	18	23	92	92	7	92	90	98	88	20	66	55	56	64	100	40	100	95
95	77	59	21	84	88	88	87	32	52	59	42	72	72	100	44	100	84	79	12	93
96	84	80	46	95	91	2	10	51	97	50	76	24	95	97	87	88	98	84	82	89
97	49	99	84	22	10	91	91	96	84	21	75	65	99	31	83	99	89	55	84	99
98	95	7	3	49	64	55	79	58	33	77	97	89	90	86	84	62	92	99	80	80
99	74	74	95	74	37	58	55	33	85	79	98	55	24	58	51	95	96	98	74	100
100	39	39	39	39	58	37	58	76	53	16	74	96	46	98	78	66	67	97	97	96

Remark: The 100 entities belongs to three subsets $U_R = \{1, 2, \ldots, 10\}$, $U_{B_1} = \{11, \ldots, 90\}$, and $U_{B_2} = \{91, \ldots, 100\}$. The entities in U_R have strong positive signal, the entities in U_{B_2} have strong negative signal, the entities in U_{B_1} do not have strong signal. The 20 rankings fall into three blocks $G_1 = \{\tau_1, \ldots, \tau_4\}$, $G_2 = \{\tau_5, \ldots, \tau_8\}$, and $G_0 = \{\tau_9, \ldots, \tau_{20}\}$. Rankings from different blocks are generated independently, the rankings in block G_0 are generated independently, while the rankings within G_1 or G_2 come from a common ranking with random manipulations. The quality of rankings in G_1 and G_2 is relatively low, while G_0 contains rankings at different quality levels.

provided by Figure 4, from which we can see that BARD tended to overestimate the quality of the rankers in G_1 and G_2 when the correlation within G_1 and G_2 is very strong (i.e., $s = 20$). Altogether, these results indicate that BARD_{HM} is efficient in dealing with correlated rankers, and BARD is reasonably robust to the model assumptions for rank aggregation.

5.3 Discovery of Highly Correlated Rankers

Here, we test the performance of the proposed coordination coefficient for detecting correlation structures among the rankers. Figure 5 shows the empirical distribution as well as the fitted Beta distribution of Q_i for three typical entities from Table 5 (entity 1, 11, and 91), suggesting that the Beta-distribution approximation does effectively capture the key feature of different types of entities. We calculated the Spearman correlation matrix, Kendall correlation matrix and coordination coefficient matrix for the dataset shown in Table 5, and displayed the results in Figure 6 (the second column of subfigure (a)). Similar results for other two datasets simulated under different correlation levels ($s = 20$ and 100) are also shown in Figure 6 (the first and third columns of subfigure (a)). We observe that the proposed

Table 6. BARD is robust to the assumption of "independent rankers"

	Naive methods			MC-based methods				CEMC		BARD			BARD_{HM}		
s	AriM	GeoM	MedR	MC_1	MC_2	MC_3	MC_4	$d = F$	$d = K$	p_1	p_2	p_3	p_1	p_2	p_3
20	0.74	0.75	0.66	0.11	0.73	0.74	0.32	0.70	0.71	0.72	0.70	0.65	0.87	0.85	0.84
60	0.78	0.81	0.72	0.10	0.80	0.79	0.32	0.75	0.75	0.83	0.84	0.82	0.87	0.86	0.85
100	0.78	0.81	0.72	0.10	0.81	0.78	0.32	0.75	0.75	0.85	0.85	0.83	0.85	0.85	0.85

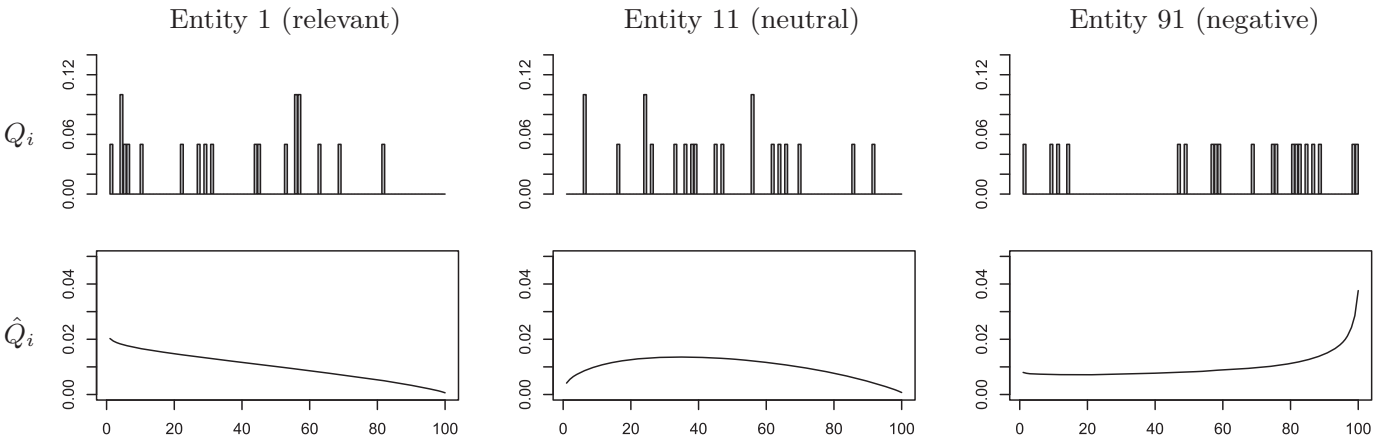


Figure 5. The natural distribution of $\{\tau_k(i)\}_{k=1}^m$ and the fitted Beta distribution $\hat{Q}_i$ for three typical entities ($i = 1$, 11, and 91) in Table 5.

method based on the coordination coefficient worked well in all cases, whereas the correlation coefficients were effective only when the dependence is extremely strong. To better evaluate the performance of the proposed method, we simulated 100 datasets for each of the three correlation levels (s =20, 60, and 100), and calculated the pair-wise discovery rates of the proposed method. The results are shown in Figure 6 (b), suggesting that the proposed method based on coordination coefficient is indeed effective to capture the correlation structure of base rankers.

6. REAL DATA APPLICATIONS

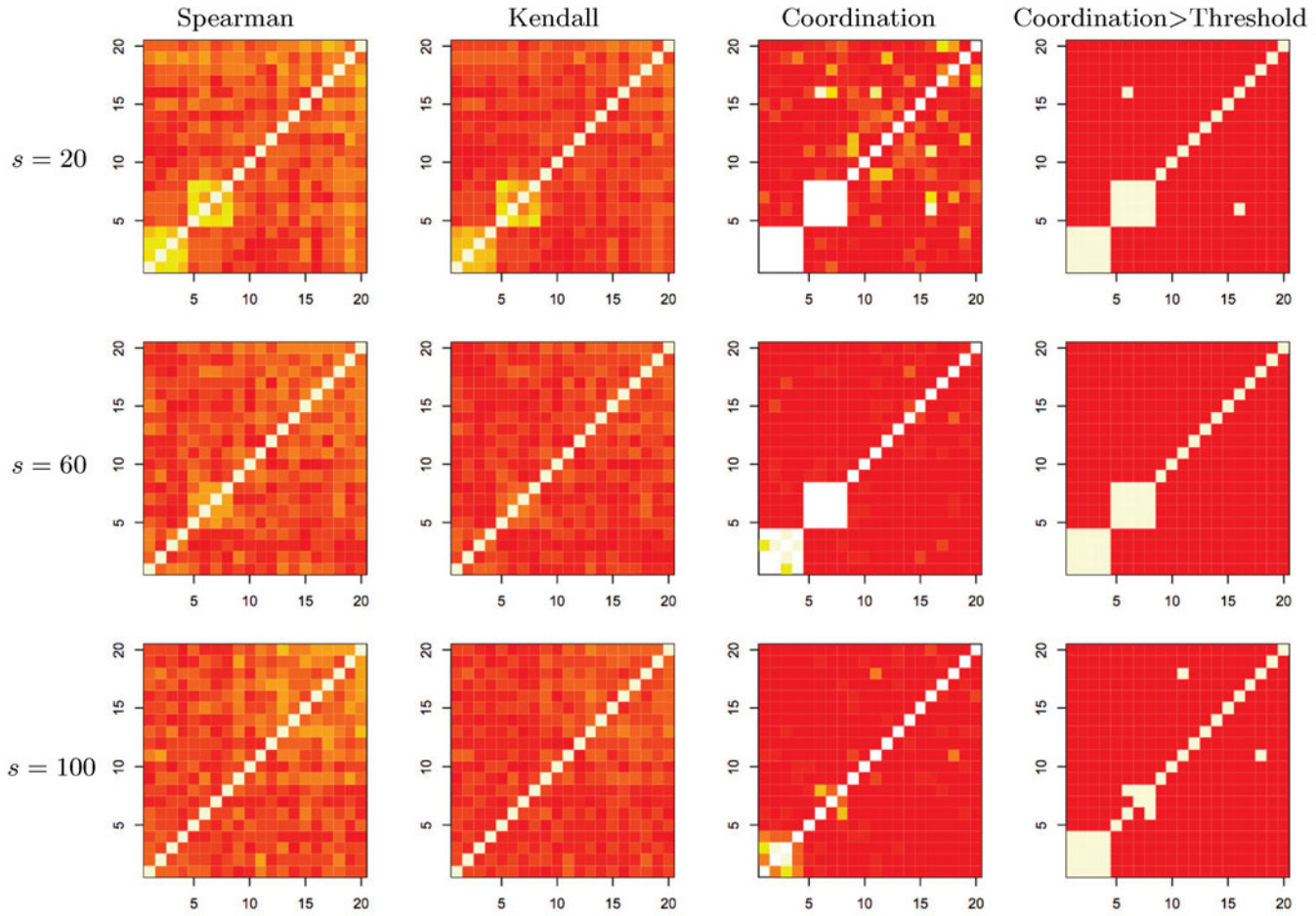
6.1 Aggregating Rankings of Cancer-Related Genes

We revisited the interesting work of DeConde et al. (2006) for combining results from five different microarray-based prostate cancer studies. The first six columns of Table 7 present the rankings of the top-25 ranked genes that were found to be up-regulated in prostate tumors compared to normal prostate tissues from five studies (Dhanasekaran et al. 2001; Luo et al. 2001;

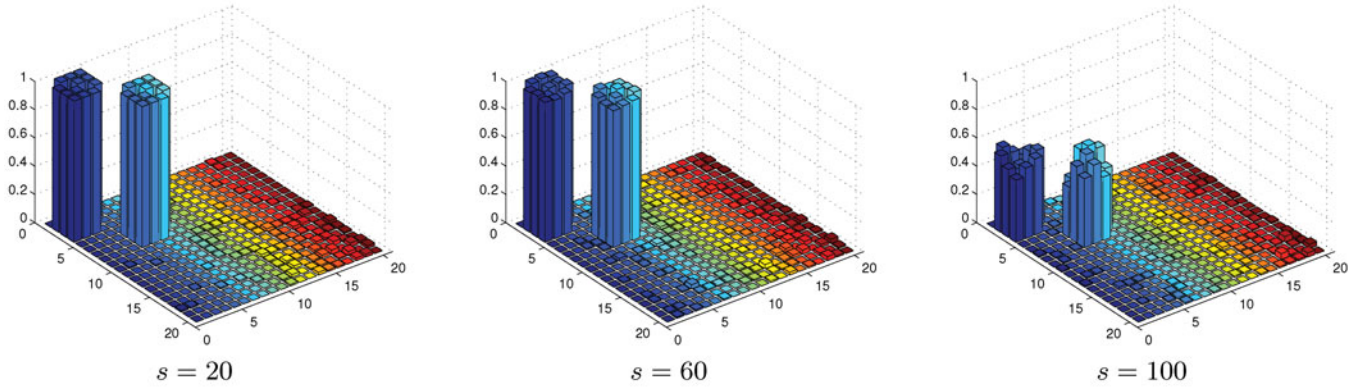
Table 7. Bayesian rank aggregation of top-25 genes from five prostate cancer studies

Individual top-25 genes from five prostate cancer studies						Top genes reported by BARD				Original ranks				
Rank	Luo(L)	Welsh(W)	Dhana(D)	True(T)	Singh(S)	Entity	$\rho_i^{(10)}$	$\rho_i^{(15)}$	$\rho_i^{(20)}$	L	W	D	T	S
1	HPN	HPN	OGT	AMACR	HPN	HPN	1.00	1.00	1.00	1	1	4	2	1
2	AMACR	AMACR	AMACR	HPN	SLC25A6	AMACR	1.00	1.00	1.00	2	2	2	1	–
3	CYP1B1	OACT2	FASN	NME2	EEF2	FASN	1.00	0.98	1.00	–	5	3	–	9
4	ATF5	GDF15	HPN	CBX3	SAT	OACT2	0.99	0.97	0.97	–	3	7	–	–
5	BRCA1	FASN	UAP1	GDF15	NME2	GDF15	0.95	0.97	0.95	–	4	13	5	17
6	LGALS3	ANK3	GUCY1A3	MTHFD2	LDHA	UAP1	0.97	0.97	0.94	–	8	5	–	25
7	MYC	KRT18	OACT2	MRPL3	CANX	OGT	0.96	0.97	0.94	–	–	1	–	–
8	PCDHGC3	UAP1	SLC19A1	SLC25A6	NACA	NME1	0.64	0.93	0.93	14	12	14	9	–
9	WT1	GRP58	KRT18	NME1	FASN	KRT18	0.97	0.97	0.92	–	7	9	–	11
10	TFF3	PPIB	EEF2	COX6C	SND1	STRA13	0.51	0.80	0.82	–	13	11	–	–
11	MARCKS	KRT7	STRA13	JTV1	KRT18	EEF2	0.47	0.89	0.81	–	–	10	14	3
12	OS-9	NME1	ALCAM	CCNG2	RPL15	PPIB	0.17	0.54	0.72	–	10	23	–	–
13	CCND2	STRA13	GDF15	AP3S1	TNFSF10	SLC19A1	0.28	0.65	0.59	–	25	8	–	–
14	NME1	DAPK1	NME1	EEF2	SERP1	CANX	0.03	0.20	0.47	–	16	–	–	7
15	DRRK1A	TMEM4	CALR	RAN	GRP58	GUCY1A3	0.32	0.66	0.46	–	–	6	–	–
16	TRAP1	CANX	SND1	PRKACA	ALCAM	GRP58	0.07	0.24	0.42	–	9	–	–	15
17	FMO5	TRA1	STAT6	RAD23B	GDF15	STAT6	0.01	0.12	0.34	–	–	17	–	–
18	ZHX2	PRSS8	TCEB3	PSAP	TMEM4	NME2	0.04	0.12	0.32	–	–	–	3	5
19	RPL36AL	EMTPD6	EIF4A1	CCT2	CCT2	TCEB3	0.00	0.08	0.31	–	–	18	–	–
20	ITPR3	PPP1CA	LMAN1	G3BP	SLC39A6	TMEM4	0.01	0.07	0.28	–	15	–	–	18
21	GCSH	ACADSB	MAOA	EPRS	RPL5	CALR	0.01	0.15	0.27	–	–	15	–	–
22	DDB2	PTPLB	ATP6V0B	CKAP1	RPS13	SND1	0.01	0.17	0.25	–	–	16	–	10
23	TFCP2	TMEM23	PPIB	LIG3	MTHFD2	EIF4A1	0.00	0.06	0.25	–	–	19	–	–
24	TRAM1	MRPL3	FMO5	SNX4	G3BP2	ANK3	0.06	0.10	0.24	–	6	–	–	–
25	YTHDF3	SLC19A1	SLC7A5	NSMAF	UAP1	MAOA	0.00	0.07	0.22	–	–	21	–	–

Remark: Totally, 89 distinct genes appear in the top-25 lists of the five studies, which are referred to as Luo, Welsh, Dhana, True, and Singh, respectively. And, $\rho_i^{(k)}$ stands for vector ρ obtained from BARD with hyperparameter $p = \frac{k}{89}$.



(a) Performance of different measurements for three typical data sets generated at different dependence levels



(b) Discovery rate of the coordination-coefficient based method from 100 simulated data sets

Figure 6. Performance of the coordination-coefficient based method for simulated data generated from the mechanism described in section 5.2, where the 20 rankings fall into three groups $G_1 = \{\tau_1, \tau_2, \tau_3, \tau_4\}$, $G_2 = \{\tau_5, \tau_6, \tau_7, \tau_8\}$, and $G_0 = \{\tau_9, \ldots, \tau_{20}\}$. The rankings in G_0 are independently generated; the rankings in G_1 and G_2 have strong within group correlation, since each ranking group are generated from a common ranking with s random transposition operations. A smaller s means a stronger within group correlation. We simulated 100 datasets for $s = 20$, 60, and 100 respectively, and applied the proposed method based on the coordination coefficient to each of the 300 simulated datasets. The pair-level discover rates are summarized into figure (b); detailed comparisons with Spearman and Kendall correlation measurements for three typical datasets are illustrated in figure (a). From the figure, we can see that the proposed method works reasonably well for all cases, while the Spearman or Kendall correlation coefficients are effective only when the dependence is extremely strong.

Welsh et al. 2001; Singh et al. 2002; True et al. 2006). These five studies relied on different technologies, and their results show that they are quite different in the genes selected to be included in the top-25 list. Lin and Ding (2009) analyzed this dataset, found that the gene list in Luo et al. (2001) is the least common compared to the other four studies, and downgraded its weight in their analysis.

Letting U be the 89 genes appeared in the five top-25 lists, and applying BARD with $\lambda = 1$ to this dataset, we obtain consistent results under different choices of the hyperparameter p

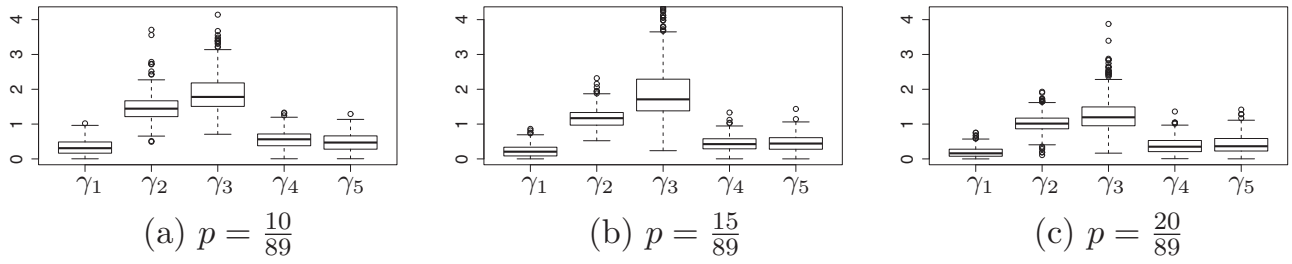


Figure 7. Posterior distributions of $\{\gamma_1, \ldots, \gamma_5\}$ obtained from BARD under different hyperparameter p for the dataset of cancer-related genes.

($p = \frac{10}{89}, \frac{15}{89}$, and $\frac{20}{89}$). As shown in Table 7, the top genes selected by BARD under different configurations reflect the consensus of the base rankers, and are robust to the choices of p. As illustrated by Figure 7, the gene lists from Welsh et al. (2001) and Dhanasekaran et al. (2001) are relatively reliable, while that from Luo et al. (2001) is of a lower quality. However, the Markov-chain-based methods (MC_1, MC_2, MC_3, MC_4, and MC_T) gave very poor results when applied to this dataset: in all these methods, the stationary distribution π of the transition matrix P degenerated to a point mass at gene OGT, that is, $\pi_i = 1$ if $i = \text{OGT}$ and $\pi_i = 0$ for all the other genes, indicating that except OGT, all other genes cannot be effectively distinguished.

6.2 Aggregating Rankings of NBA Teams

Ranking sports teams has attracted tremendous attention from both sports analysts and academics. Numerous ranking methods have been proposed for different sports, including NBA, NFL, MLB, NCAA football, etc. (see Langville and Meyer 2012 for a comprehensive review). The BARD method produces an aggregated ranking considering results of any number of ranking methods, which can be used to provide better predictions of game outcomes and to evaluate the effectiveness of different sports statistics in generating rankings.

As explained in the NBA team ranking example of Section 1 and shown in Table 1, we collected six professional rankings and 28 amateur rankings for the 30 NBA teams in the 2011–2012 season. Defining the 16 teams that entered the 2011–2012 playoffs (i.e., the first 16 teams in Table 1) as "relevant entities," we expect BARD to give higher γ_k's to both professional rankers and students who had paid more attentions to NBA games.

We applied BARD to the dataset with hyperparameter $p = \frac{16}{30}$ and $\lambda = 1$ to "predict" which teams can make their appearance in the playoffs. The results are summarized into Figure 8. From subfigure (a), we can see that BARD figures out the quality difference among different rankers successfully: boxplots of the γ_i's show a clear decreasing trend with the decrease of the

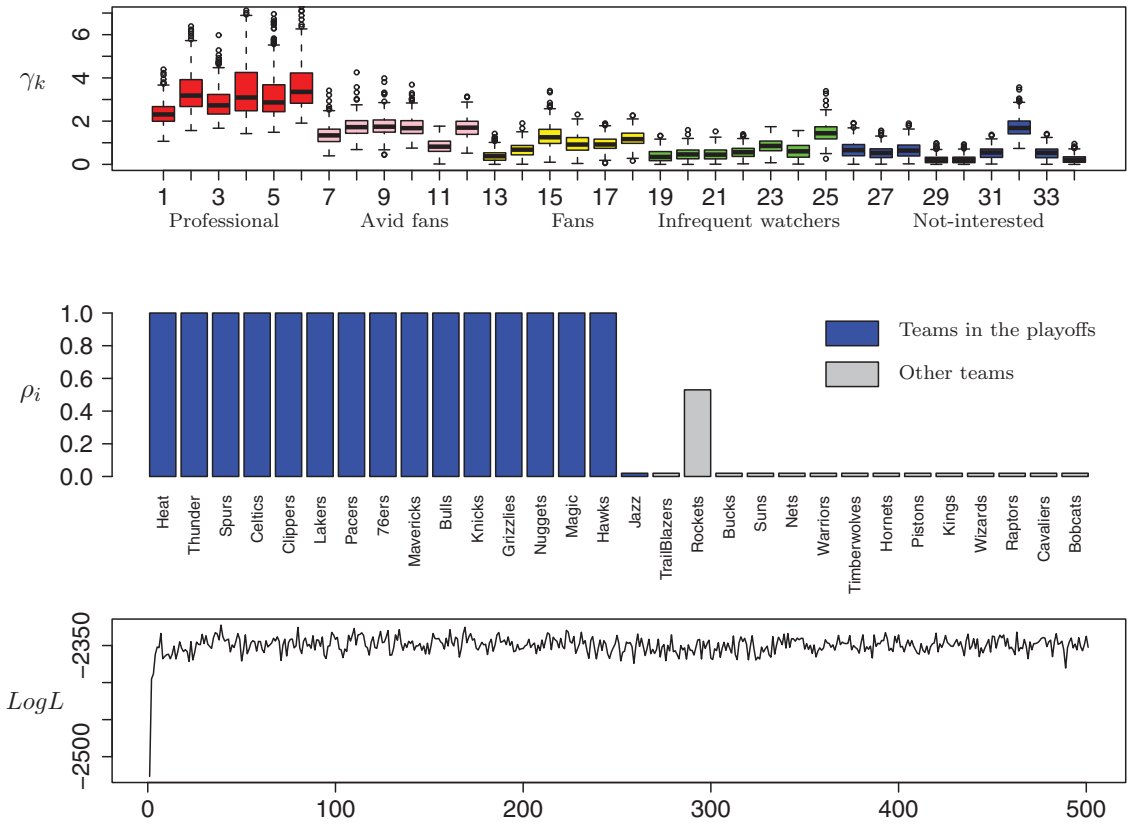


Figure 8. Results from BARD for aggregating 34 rankings of 30 NBA teams in season 2011–2012.

basketball knowledge level of the rankers. We also observed some interesting "outliers" from these boxplots. For example, ranker S_5 is an outlier in the group of *Avid fans*, which precisely reflects the fact that S_5 gave high ranks to Warriors and Wizards, two teams that failed to enter the playoffs. Similarly, the low quality values estimated for rankers S_7 and S_8 also reflect correctly the fact that both S_7 and S_8 gave high ranks to multiple weak teams.

Moreover, the aggregated ranking outperformed individual rankings in terms of being closer to the "truth," even though amateur rankings from the students contain considerable amount of noises. The aggregated ranking makes only one mistake: putting Rockets instead of Jazz into playoffs. Among the six professional rankings, however, only P_5 achieved the same result as the aggregated ranking; the other five rankings made two mistakes: P_1 missed Nuggets and Jazz, P_3 missed Magic and Jazz, P_2, P_4, and P_6 missed Hawkes and Jazz.

7. DISCUSSION

In this article, we propose the Bayesian rank aggregation (BARD) method for the rank aggregation problem. By giving each base ranker a specific quality parameter and estimating these parameters using the data, BARD measures the reliability of the base rankers in a quantitative way and makes use of this information to improve the aggregated rank list. Compared to the methods in the literature, BARD works significantly better when the equality of base rankers varies greatly. Both simulation studies and real data applications demonstrated the usefulness and superiority of BARD.

BARD assumes that: (1) the entities involved can potentially be divided into two subsets: relevant entities U_R and background entities U_B; (2) given the group indicators of entities $I = \{I_i\}_{i \in U}$, the rankers $\tau_1, \ldots, \tau_m$ are conditionally independent; (3) for each base ranker τ_k, internal relative rankings of entities within each subset (i.e., relevant or background) are assigned randomly, and the rank of a relevant entity among the background entities follows a power law distribution. To apply BARD to a practical problem, we need to check whether the above assumptions (especially, the first two) hold approximately. BARD is reasonably robust if some base rankers are moderately correlated. However, if correlations among certain rankers are very strong, BARD may report biased results. Section 4 describes an effective tool for a fast detection of strong ranker correlations and a Bayesian hierarchical model to account for the correlation structure. Our simulation results showed that our two-step strategy performed satisfactorily when facing strong ranker correlations.

BARD requires that all base rankers in consideration have a common objective. That is, at a conceptual level there is a common "true" set of relevant entities U_R, and all base rankers aim to rank relevant entities from U_R better than background ones. If the data collected in practice do not satisfy this requirement (e.g., the rankings from different base rankers have different goals, or are purely based on the opinion of each individual ranker), BARD may not be an appropriate tool to use.

In general, BARD is robust to different choices of hyperparameters, such as the expected percentage p of relevant entities in U when p comes from a proper range (e.g., $[0.01, 0.2]$), and the prior for ranker quality λ. In some practical problems, choices of p and λ are obvious. If not, we may need to try different choices from proper ranges and check how robust the results are before a conclusion can be made.

The framework of BARD supports us to deal with full rankings, partial rankings and rankings with ties. It is possible to further extend this framework to problems with more complicated structures. For example, if some covariates of the entities of interest are also observed, which can potentially influence rankings of some base rankers, it is desirable to link these covariates to the quality parameters of corresponding base rankers to achieve a better performance.

APPENDIX

Detailed information about the professional rankings of NBA teams used in Section 6.2.

Ranking	Provider	Link
P_1	FOXSports.com	*http://msn.foxsports.com/nba/powerRankings/2011-2012/PRE*
P_2	ESPN.com	*http://espn.go.com/nba/powerrankings/_/week/0*
P_3	SI.com	*http://sportsillustrated.cnn.com/2011/writers/britt_robson/12/20/preseason.power.rankings/index.html*
P_4	NBA.com	*http://www.nba.com/2011/news/powerrankings/12/21/preseason/index.html*
P_5	midwestsportsfans.com	*http://www.midwestsportsfans.com/2011/12/nba-power-rankings-preseason-edition/*
P_6	jsonline.com	*http://www.jsonline.com/sports/136175388.html*

[Received August 2012. Revised December 2013.]

REFERENCES

Ahmad, N., and Beg, M. M. S. (2002), "Fuzzy Logic Based Rank Aggregation Methods for the World Wide Web," in *Proceedings of the International Conference on Artificial Intelligence in Engineering and Technology*, Malaysia, pp. 363–368. [1023]

Aslam, J. A., and Montague, M. (2001), "Models for Metasearch," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 276–284. [1023,1026]

Beg, M. M. S. (2004), "Parallel Rank Aggregation for the World Wide Web," in *Proceedings of the 2004 International Conference on Intelligent Sensing and Information Processing*, Piscataway, NJ: IEEE Press, pp. 385–390. [1023]

Borda, J. C. (1781), "Mémoire sur les élections au scrutin," *Histoire del' Académie Royale des Sciences*, pp. 657–665. [1023]

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, article 15. [1023,1026,1035]

Dhanasekaran, S. M., Barrette, T. R., Ghosh, D., Shah, R., Varambally, S., Kurachi, K., Pienta, K. J., Rubin, M. S., and Chinnaiyan, A. M. (2001), "Delineation of Prognostic Biomarkers in Prostate Cancer," *Nature*, 412, 822–826. [1035,1037]

Diaconis, P. (1988), "Group Representation in Probability and Statistics," *Lecture Notes-Monograph Series*, 11, Hayward, CA: IMS. [1024]

Diaconis, P., and Graham, R. (1977), "Spearman's Footrule as a Measure of Disarrary," *Journal of the Royal Statistical Society*, Series B, 39, 261–268. [1024]

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*, pp. 613–622. [1023,1024,1026]

Fagin, R., Kumar, R., and Sivakumar, D. (2003), "Comparing Top k Lists," *SIAM Journal of Discrete Mathematics*, 17, 134–160. [1023]

Fagin, R., Lotem, A., and Naor, M. (2001), "Optimal Aggregation Algorithm for Middleware," in *Proceedings of the 20th ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems*, pp. 102–113. [1023]

Fligner, M. A., and Verducci, J. S. (1986), "Distance Based Ranking Models," *Journal of the Royal Statistical Society*, Series B, 48, 359–369. [1024]

Freund, Y., Iyer, R., Schapire, R. E., and Singer, Y. (2003), "An Efficient Boosting Algorithm for Combining Preferences," *Journal of Machine Learning Research*, 4, 933–969. [1023,1026]

Hull, D. A., Pedersen, J. O., and Schütze, H. (1996), "Method Combination for Document Filtering," in *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 279–287. [1023]

Lam, K. W., and Leung, C. H. (2004), "Rank Aggregation for Metasearch Engines," in *Proceedings of the 13th International Conference on World Wide Web*, pp. 384–385. [1023]

Langville, A. N., and Meyer, C. D. (2012), *Who's #1?: The Science of Rating and Ranking*, Princeton, NJ: Princeton University Press. [1037]

Lin, S. L., and Ding, J. (2009), "Integration of Ranked Lists Via Cross Entropy Monte Carlo With Applications to mRNA and microRNA Studies," *Biometrics*, 65, 9–18. [1023,1026,1036]

Liu, J. S. (2001), "Monte Carlo Strategies in Scientific Computing," in *Springer Series in Statistics*, eds. P. Bickel, et al., New York: Springer-Verlag. [1029]

Liu, Y., Liu, T., Qin, T., Ma, Z., and Li, H. (2007), "Supervised Rank Aggregation," in *Proceedings of the 16th International Conference on World Wide Web*, pp. 481–490. [1023,1026]

Luo, J., Duggan, D. J., Chen, Y., Sauvageot, J., Ewing, M., Bittner, M. L., Trent, J. M., and Isaacs, W. B. (2001), "Human Prostate Cancer and Benign Prostatic Hyperplasia: Molecular Dissection by Gene Expression Profiling," *Cancer Research*, 61, 4683–4688. [1035,1037]

Mallows, C. L. (1957), "Non-Null Ranking Models," *Biometrika*, 44, 114–130. [1024]

Manmatha, R., Rath, T., and Feng, F. (2001), "Modeling Score Distributions for Combining the Outputs of Search Engines," in *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 267–275. [1023]

Manmatha, R., and Sever, H. (2002), "A Formal Approach to Score Normalization for Meta-search," in *Proceedings of the 2nd International Conference on Human Language Technology Research*, pp. 98–103. [1023]

Meila, M., Phadnis, K., Patterson, A., and Bilmes, J. (2007), "Consensus Ranking Under the Exponential Model," in *Proceedings of the 23rd Annual Conference on Uncertainty in Artificial Intelligence* (UAI). [1024]

Montague, M., and Aslam, J. A. (2001), "Relevance Score Normalization for Meta-search," in *Proceedings of the 10th Conference on Information and Knowledge Management*, pp. 427–433. [1023]

Randa, M. E., and Straccia, U. (2003), "Web Metasearch: Rank vs. Score Based Rank Aggregation Methods," in *Proceedings of the 2003 ACM Symposium on Applied Computing*, pp. 841–846. [1023]

Rubin, D. B. (1984), "Bayesianly Justifiable and Relevant Frequency Calculations for the Applied Statistician," *The Annals of Statistics*, 12, 1151–1172. [1030]

Rubinstein, R. Y., and Kroese, D. P. (2004), *The Cross-Entropy Method. A UniPed Approach to Combinatorial Optimization, Monte-Carlo Simulation and Machine Learning*, New York: Springer. [1026]

Sese, J., and Morishita, S. (2001), "Rank Aggregation Method for Biological Databases," *Genome Informatics*, 12, 506–507. [1023]

Shaw, J. A., and Fox, E. A. (1994), "Combination of Multiple Searches," in *Proceedings of the 2nd Text Retrieval Conference*, pp. 243–252. [1023]

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D′, Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (2002), "Gene Expression Correlates of Clinical Prostate Cancer Behavior," *Cancer Cell*, 1, 203–209. [1036]

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. [1030]

True, L., Coleman, I., Hawley, S., Huang, A., Gifford, D., Coleman, R., Beer, T., Gelman, E., Datta, M., Mostaghel, E., Knudsen, B., Lange, P., Vessella, R., Lin, D., Hood, L., and Nelson, P. (2006), "A Molecular Correlate to the Gleason Grading System for Prostate Adenocarcinoma," *Proceedings of the National Academy of Sciences of the USA*, 103, 10, 991–10,996. [1036]

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*, pp. 443–452. [1023]

Vogt, C., and Cottrel, G. W. (1999), "Fusion Via a Linear Combination of Scores," *Information Retrieval*, 3, 151–173. [1023]

Welsh, J. B., Sapinoso, L. M., Su, A. I., Kern, S. G., Wang-Rodriguez, J., Moskaluk, C. A., Frierson, H. F. Jr, and Hampton, G. M. (2001), "Analysis of Gene Expression Identifies Candidate Markers and Pharmacological Targets in Prostate Cancer," *Cancer Research*, 61, 5974–5978. [1036]