Inferring a consensus problem list using penalized multistage models for ordered data

Philip S. Boonstra*
Department of Biostatistics, University of Michigan, USA
and
John C. Krauss
Division of Hematology Oncology, University of Michigan, USA

April 22, 2020

Abstract

A patient's medical problem list describes his or her current health status and aids in the coordination and transfer of care between providers. Because a problem list is generated once and then subsequently modified or updated, what is not usually observable is the provider-effect. That is, to what extent does a patient's problem in the electronic medical record actually reflect a consensus communication of that patient's current health status? To that end, we report on and analyze a unique interview-based design in which multiple medical providers independently generate problem lists for each of three patient case abstracts of varying clinical difficulty. Due to the uniqueness of both our data and the scientific objectives of our analysis, we apply and extend so-called multistage models for ordered lists and equip the models with variable selection penalties to induce sparsity. Each problem has a corresponding non-negative parameter estimate, interpreted as a relative log-odds ratio, with larger values suggesting greater importance and zero values suggesting unimportant problems. We use these fitted penalized models to quantify and report the extent of consensus. We conduct a simulation study to evaluate the performance of our methodology and then analyze the motivating problem list data. For the three case abstracts, the proportions of problems with model-estimated non-zero log-odds ratios were 10/28, 16/47, and 13/30. Physicians exhibited consensus on the highest ranked problems in the first and last case abstracts but agreement quickly deteriorated; in contrast, physicians broadly disagreed on the relevant problems for the middle – and most difficult – case abstract.
Keywords: conditional multinomial; L_0 penalty; ranked lists; variable selection

*1415 Washington Heights, Ann Arbor, Michigan, USA, 48109-2029; Tel: +1 734 615 1580; philb@umich.edu

1 Introduction

A patient's medical problem list is defined as the minimal number of diagnoses that describe that patient's current health status and risks to future health (Krauss et al., 2016). It serves as a "dynamic 'table of contents' " (Weed, 1968) for the patient, which is useful for coordination of care between providers and care environments (Krauss et al., 2016). All providers of care for a patient work from the same problem list and update it at each encounter, but little is known about how much consensus there is between each provider's individually generated problem list. There is clinical interest in having the problem list accurately reflect the patient's current health. In other words, to what extent does a patient's problem list in the electronic medical record reflect a consensus communication of that patient's current health status? The statistical methodology developed in this paper is directly motivated by the idiosyncrasies of this ranked data context, as elucidated below.

The data upon which our methodology is based were collected via a series of interviews of faculty physicians at the University of Michigan (Ann Arbor, MI) conducted by the second author (JCK, the interviewer) between May 2013 and July 2014. All faculty members in the Department of General Medicine and the Department of Family Medicine (approximately 150 in total) were electronically invited to participate, and thirty-eight consented. Each interview consisted of the participating physician reviewing three real, previously reported patient case abstracts (labeled A, B, and C) that have been specifically developed for physician training in clinical reasoning (Meyer et al., 2013). For each case, the physician was asked to write down what would be her problem list for that patient as if she were the provider of care. The first six interviews were used as training for the interviewer to standardize the process as well as to develop a written vocabulary of expected problems for that case. The subsequent 32 interviews comprised the study data. For any novel problems encountered in this second round of interviews that were, in the opinion of the interviewer, similar to an existing problem already in the vocabulary, the interviewer noted this similarity and asked whether the subject would consider these equivalent or not. If the subject said 'no', then the novel problem was left as is. The cases were presented in the same order (alphabetical by label) for all interviews, based on an assumption that the most complex clinical case would be B and the simplest clinical case would be C. The interview results are summarized in Figure 1. See Krauss et al. (2016) for more details on the study design and case abstracts. The data obtained in this study – 32 de-novo problem lists generated for the same patient at a single point in time – do not naturally occur in a medical chart. Therefore, this study provides a unique opportunity to measure physician agreement and the degree to which a newly generated problem list is consistently generated. In other words, to what extent can a physician expect the accompanying problem list she receives with a patient to be the same problem list she herself would generate for that patient?

Similar questions arise in other diverse ranked data contexts, including election polling (Gormley and Murphy, 2008; Gormley et al., 2008), sorting genomic features (DeConde et al., 2006; Boulesteix and Slawski, 2009; Lin and Ding, 2009; Lin, 2010; Li et al., 2017, 2018), identifying bovine feeding preferences (Nombekela et al., 1994), handicapping horse races (Plackett, 1975; Benter et al., 2008), ranking basketball teams (Deng et al., 2014), or

indexing search engine results (Webber et al., 2010). However, once the data are in hand, the subsequent analysis typically converges on a common goal, namely that of measuring agreement between the rankers.

So called 'multistage models', which are essentially a sequence of conditional multinomial distributions, can be used for aggregating and modeling a set of ordered lists such as what we analyze here (Plackett, 1975; Luce, 1959; Benter et al., 2008; Gormley and Murphy, 2008; Mollica and Tardella, 2017). However, multiple idiosyncracies, both with respect to the underlying nature of the data and with respect to our scientific objectives, require novel extensions to this multistage, model-based approach. There are three such proposed in this manuscript. First, we adjust multistage models to handle so-called ragged lists, which can have different lengths because the ranker chooses to stop ranking. The length of each list becomes informative in this case, and we model the fatigue process of rankers. Second, we equip the likelihood with a modified L_0-type variable selection penalty to induce sparsity among the maximum penalized likelihood estimates. Sparsity is particularly desirable here because many aggregators will rank *all* items, requiring a post-hoc determination as to whether and where to truncate the consensus list. Although such penalties have been widely used, to our knowledge they have not yet been applied to models for ranked data, and thus this work represents a novel amalgamation of classical statistical models for ordered data with modern penalized regression techniques warranted by the motivating context. Third, we provide a computational framework in the `R` statistical environment for fitting these penalized models, including a coordinate ascent algorithm and tuning parameter selection based upon information theoretic criterion to select the appropriate amount of penalization. The remainder of this manuscript provides technical background (Section 2), describes in detail each of our proposed extensions (Section 3), presents a simulation study evaluating our methodology against eight possible comparators (Section 4), and then finally illustrates their application to the problem list data of interest (Section 5) and the NBA team rankings data from Deng et al. (2014) (Section 6). We conclude with a discussion in Section 7.

2 Technical background

Assume that each ranker is ordering items from a set of items, where each item is unambiguously mapped to an integer label $\{1, \ldots, v\}$. As noted in the introduction, a 'ranker' could be anyone or anything from a person to a search engine to a cow to a case-control study; however, in our motivating context and therefore our methodology, the ranker is sentient and free to stop ranking at any point. Lists from multiple rankers are available, and we model the process of constructing these lists. Such models usually require that the data be formulated as either *ranked* or *ordered* lists (Marden, 1996). Both data types convey equivalent information, and both take the set of all permutations of the v integers as their support. However, whereas a ranked list gives the ranks of the v items, an ordered list permutes the v items themselves based upon their ranking. Specifically, the sth entry of a ranked list is the rank assigned to the item having integer label s (lower numbers indicate higher ranks), and the sth entry of an ordered list is the integer label of the item that is ranked sth (items

appearing early in the list are ranked higher). The data in this paper are formulated as ordered lists, but we will refer to items that are ordered first as being 'highly ranked'.

The orderings may be incomplete. For example, top-ranked lists of genes based upon phenotypic association do not include every single gene but are always truncated, e.g. to the top 25 genes (DeConde et al., 2006). The New York Times Hardcover Nonfiction Best Seller List (`https://www.nytimes.com/books/best-sellers/hardcover-nonfiction`, accessed 15-Mar-19) publishes a weekly list of the 15 best-selling hardcover non-fiction books, and extant but unpublished are the number 16, 17, etc. best-sellers. When such lists are also top-weighted, meaning that disagreement between two lists at higher ranks is more important than at lower ranks, Webber et al. (2010) call them 'indefinite'. Not considering the top-weighted characteristic, such lists are called 'top-k lists' by Dwork et al. (2001) when they have been uniformly truncated to the first k items and 'partial lists' by Deng et al. (2014) when the point of truncation differs from list to list. Importantly, partial lists could be longer but have been artificially truncated, beyond the purview of the ranker and external to the ordering process. Distinct from these are what we call *ragged* lists, which we define as an ordered list arising from a ranker who is free to stop ranking. Subtly, a ragged list may be complete, if that ranker chooses to order all items. If there are unranked items, one can infer that the un-ranked items are below the ranked items. The problem list data we study here are ragged, since each physician was free to select as many or as few problems as desired. Although some existing rank aggregation methodologies can analyze ragged lists, to our knowledge there are none explicitly designed to model the stopping process nor induce sparsity in the aggregated list.

Figure 1 plots the frequency with which each item (problem) was ranked by a physician for each of the three patient abstracts. Focusing on case A, thirty physicians ranked DIABETES MELLITUS somewhere in their constructed problem list for this patient, whereas eight problems were ranked by just one physician (not necessarily the same person). 23/32 physicians ranked OSTEOARTHRITIS somewhere on their list, but only in one physician's list was it in the top 4. In contrast, 26/32 physicians ranked PNEUMONIA first on their list. Less overall agreement was observed on case B, with no problems being ranked first by more than eight physicians, and 18 problems appearing on exactly one list.

We now introduce some notation. For $i = 1, \ldots, n$, the ith ranker's ordered list of ℓ_i items is denoted by $\boldsymbol{x}_i = \{x_{i1}, x_{i2}, \ldots, x_{i\ell_i}\}$, with $x_{is} \in \{1, \ldots, v\}$ and $s = 1, \ldots, \ell_i$ indexing each stage. If the lists are complete, then $\ell_i \equiv v$ for all lists; if they are partial, then $\ell_i \equiv \ell < v$ for all i, where ℓ is artificially chosen and external to the modeling process; if they are ragged, then $\ell_i \leq v$ for each i, with potentially different values of ℓ_i for each i. We describe two broad approaches for analyzing ordered lists.

2.1 Pairwise similarities

One approach for quantifying agreement is to measure a distance or similarity between any pair of lists $\boldsymbol{x}_{i_1}$ and $\boldsymbol{x}_{i_2}$. For complete ordered lists, Kendall's τ (Kendall, 1948) or Spearman's ρ (Spearman, 1904) could be used. Lin and Ding (2009) proposed the Cross

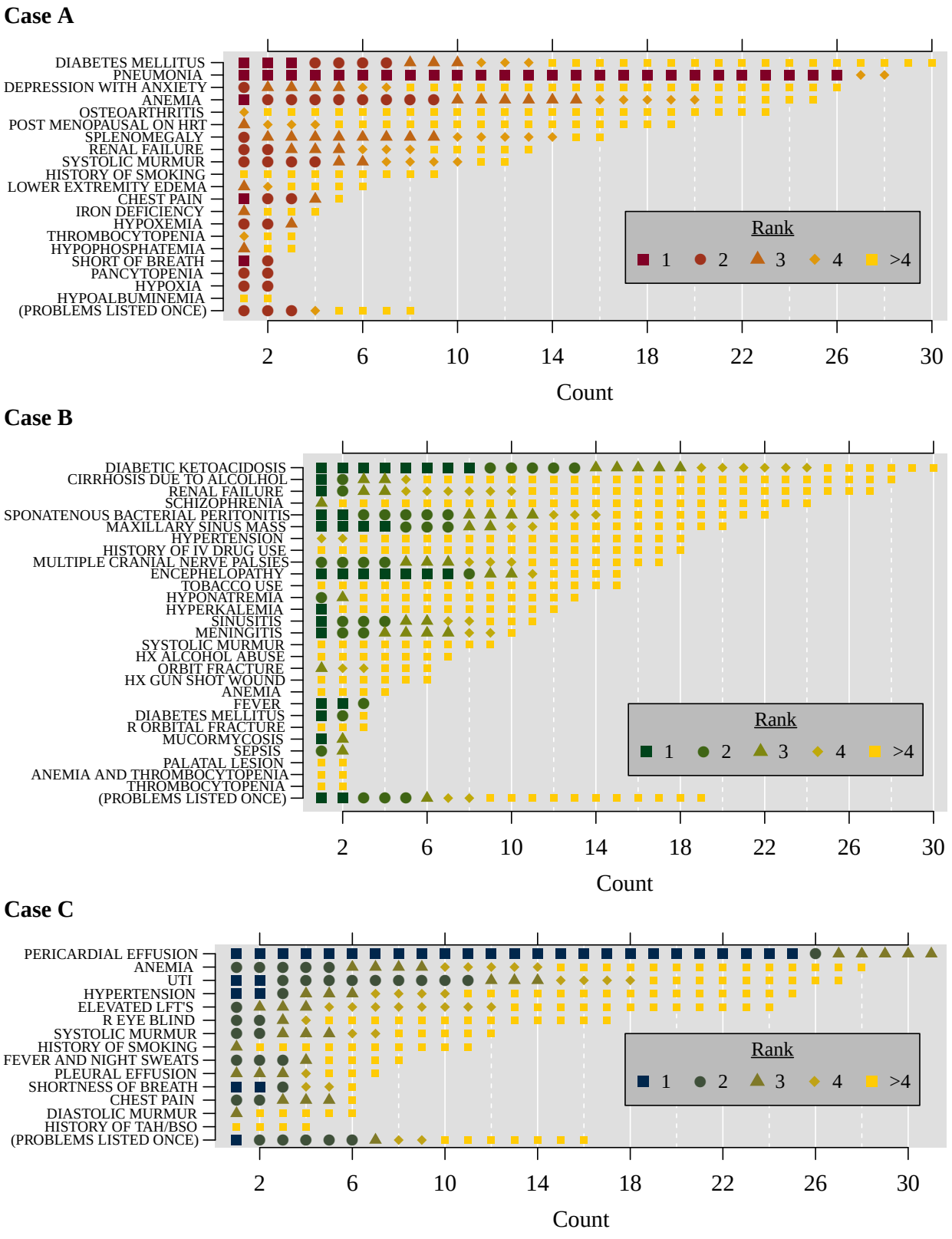

Figure 1: Counts of the frequency that each problem was listed on any of the $n = 32$ generated lists for each of three case abstracts, with shading and shape used to indicate the rank of that problem. For brevity, only those problems listed by at least two physicians are shown.

Entropy Monte Carlo (CEMC) algorithms that approximate an aggregated list that is, on average, closest to all observed lists with respect to one of these correlations. As we discuss below, these pairwise similarities may not always be appropriate for ordered lists. The rank-biased overlap (RBO, Webber et al., 2010) is a more recent example specifically designed for ordered lists. Given a user-specified parameter $\psi \in (0, 1)$, the RBO between two lists $\boldsymbol{x}_{i_1}$ and $\boldsymbol{x}_{i_2}$ is

$$\text{RBO}_\psi(\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}) = \frac{1-\psi}{\psi} \sum_{d=1}^{\infty} \psi^d |\boldsymbol{x}_{i_1,1:d} \cap \boldsymbol{x}_{i_2,1:d}|/d,$$

where the expression $|\boldsymbol{x}_{i_1,1:d} \cap \boldsymbol{x}_{i_2,1:d}|/d$ denotes the size of the intersection of the first d elements divided by d. This proportion of the first d elements of each list that are shared is the so-called agreement at depth d. Agreements across all possible depths are then infinitely averaged using a convergent series of weights $\{\psi^d\}_{d=1}$. Values of this similarity measure fall in the interval $[0, 1]$, where 1 indicates perfect overlap at all depths, and 0 indicates no overlap anywhere. The RBO assumes that each list is long enough so as to be effectively infinite. The exact value can only be calculated by examining an infinite value of depths. However, by truncating the calculation to some finite depth and determining the smallest and largest possible added value beyond this depth, a window within which the true RBO must lie can be created, the width of which decreases as the depth of truncation increases, due to the infinite series being convergent.

Krauss et al. (2016) proposed a length-dependent version of the RBO (LDRBO), specifically for measuring the similarity between two finite ragged lists:

$$\text{LDRBO}_\psi(\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}) = \frac{\sum_{d=1}^{\max\{\ell_{i_1}, \ell_{i_2}\}} \psi^d |\boldsymbol{x}_{i_1,1:d} \cap \boldsymbol{x}_{i_2,1:d}|/d}{\sum_{d=1}^{\max\{\ell_{i_1}, \ell_{i_2}\}} \psi^d}.$$

LDRBO measures average agreement, like RBO. It differs in that the maximum depth evaluated is always the longer of the two lists. Also contrasting with RBO, ψ can be set to 1 for LDRBO, in which case LDRBO simplifies to the average agreement across all depths. LDRBO and RBO will become similar as $\min\{\ell_{i_1}, \ell_{i_2}\}$ increases. It is also noteworthy that rank-based similarity measures such as (LD)RBO yield qualitatively different interpretations than standard correlation measures like Spearman's ρ, even for very simple cases. For example, $\boldsymbol{x}_{i_1} = \{1, 2, 3, 4\}$ and $\boldsymbol{x}_{i_2} = \{4, 3, 2, 1\}$ have perfect negative correlation ($\rho = -1$); in contrast, with $\psi = 1$, LDRBO evaluates to a middling value of $(0/1 + 0/2 + 2/3 + 4/4)/4 \approx 0.42$. LDRBO can only be zero between lists having no intersection, such as $\boldsymbol{x}_{i_1} = \{1, 2, 3, 4\}$ and $\boldsymbol{x}_{i_3} = \{5, 6, 7, 8\}$, which, coincidentally, have a perfect positive correlation ($\rho = 1$). Thus, in the context of ordered lists, (LD)RBO better reflects the intuition that the exemplar pair $\{\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}\}$ share more in common than $\{\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_3}\}$.

Using the motivating problem list data and setting $\psi = 1$, Krauss et al. (2016) used numerical methods to identify a theoretical 'consensus problem list' having the largest median value of LDRBO across all 32 physicians' constructed lists. For Case A in Figure 1, the estimated consensus problem list was {PNEUMONIA, DIABETES MELLITUS, ANEMIA, SPLENOMEGALY, DEPRESSION WITH ANXIETY, OSTEOARTHRITIS, RENAL FAILURE, HYPOXIA}, having a median LDRBO of 0.683.

The findings of Krauss et al. (2016) not withstanding, insofar as the objective is to measure consensus and calculate an aggregated consensus list, similarity-based approaches such as the LDRBO may fall short. For example, the above methods can be used to calculate a consensus problem list for any group of lists, no matter how disparate the data are. Further, there is no obvious mathematical rationale to suggest that maximizing the median pairwise LDRBO – as opposed to the mean, minimum, or maximum LDRBO – results in the 'right' consensus list, nor that $\psi = 1$ is the right choice. Finally, even with these relatively small datasets, there are practical computational challenges to this approach: with 28 unique problems in Figure 1, there are $28! \approx 10^{29.5}$ permutations of length 28 to search across as possible consensus lists, plus all candidate lists less than length 28. Krauss et al. (2016) used an approximate 'branch and bound' algorithm to substantially limit the scope of the search.

3 Model-based approaches

These reasons provide compelling rationale to consider instead model-based approaches for the analysis of our problem list data and ordered lists in general.

One model-based approach assumes that the ranking process is a Markov process. Treating the set of v items as the state space of a Markov chain, one can calculate the stationary distribution of the corresponding transition matrix, where larger probabilities correspond to higher ranked items (Lin, 2010; Dwork et al., 2001; DeConde et al., 2006). Depending on how the transition matrix is determined, there are several such approaches, which Lin (2010) label MC1, MC2, and MC3. Readers are referred to these references for more details. These approaches are amenable to the analysis of ragged lists, but, without further pruning, the resulting consensus list is always an ordering of all items listed by at least one ranker.

A second approach, called Mallows model, posits that each ranked list differs from some unknown, consensus ranking, i.e. a parameter vector to be estimated, according to a probabilistic model based upon a distance, i.e. Spearman's ρ or Kendall's τ, between the two list (Mallows, 1957). Besides the consensus ranking, the other parameter to be estimated in a Mallows-type model is the dispersion $\phi \in [0, 1]$, where $\phi = 0$ means that all rankers are exactly recapitulating the consensus ranking, and $\phi = 1$ means that rankers are choosing items uniformly at random, without regard to the consensus ranking. Fligner and Verducci (1988) and Li et al. (2019) have both extended the Mallows model, allowing ϕ to depend upon the stage s, such that $\phi(s)$ can be small when s at earlier stages, where agreement is more important, and $\phi(s)$ can be large at later stages, where agreement is usually less important.

A third approach, and that which we extend in this paper, is the multistage model, which explicitly formulates the list-generating process (Plackett, 1975; Luce, 1959). The ith ranker generates an ordered list of length v from among a pre-specified, fixed-length set of items, starting with his/her/its most-preferred item. Define $\mathcal{O}_{is}$ to be the set of items yet to be

ranked just before the sth stage:

$$\mathcal{O}_{is} = \begin{Bmatrix} \{1, \ldots, v\}, & s = 1 \\ \{k : k \notin \{x_{is'}\}_{s'<s}\}, & s > 1 \end{Bmatrix}, \tag{1}$$

and let $1_{[X]}$ be 1 when the statement X is true and 0 otherwise. The Plackett-Luce (PL) probability that item $k \in \{1, \ldots, v\}$, is ordered sth is $\Pr(x_{is} = k|\mathcal{O}_{is}) = 1_{[k\in\mathcal{O}_{is}]} \exp(\theta_k) / \sum_{j\in\mathcal{O}_{is}} \exp(\theta_j)$, i.e. proportional to $\exp(\theta_k)$ until it gets ordered, and zero afterwards. There are v parameters, $\Theta = \{\theta_1, \theta_2, \ldots, \theta_v\}$. Of these, $v - 1$ are identified, and without loss of generality, we may assume that $\min_j\{\theta_j\} \equiv 0$. See Section 5.6, Marden (1996) for an overview of classical multistage models. In contrast to the Mallows model, there is no explicit consensus list in this model; however, the set of the numeric weights θ_ks gives, both in an absolute and relative sense, the order of preference across all items.

Analogous to the extended Mallows model proposed by Li et al. (2019), Benter added a dampening effect to PL models to allow for the relative preference between items to depend on the stage (Benter et al., 2008; Gormley and Murphy, 2008). Let a dampening function $\delta(s)$ map the set of integers $s \in \{1, \ldots, v - 1\}$ to the interval $(0, 1]$, with $\delta(1) \equiv 1$ for identifiability. When $\delta(s)$ is small, the distinction between items decreases, and so, assuming that preferences are always strongest at early stages, it is reasonable to constrain $\delta(s)$ to be non-increasing with s. At the final stage, $\delta(v)$ may take any value, since there is no choice remaining. When $\delta(\cdot)$ is limited to the set of non-increasing functions, this dampening function serves analogously to the ψ parameter in the (LD)RBO measures and the ϕ-function in the extended Mallows models, namely to reflect that agreement at higher ranks is relatively more important than at lower ranks.

Thus, the Benter-Plackett-Luce (BPL) model for the probability of selecting item k at the sth stage conditional on the choices from the previous $s - 1$ stages is $\Pr(x_{is} = k|\mathcal{O}_{is}) = 1_{[k\in\mathcal{O}_{is}]} \exp(\theta_k\delta(s)) / \sum_{j\in\mathcal{O}_{is}} \exp(\theta_j\delta(s))$, for $k = 1, \ldots, v$ and $s = 1, \ldots, \ell_i$. To be estimated are the $v-1$ identified parameters in Θ plus the number of parameters in the chosen functional form of $\delta(\cdot)$, which we discuss in Section 3.1 below. At stage s, the log-odds of ordering item k_1 over k_2, conditional on neither having been yet ordered, are $\delta(s)[\theta_{k_1} - \theta_{k_2}]$. We now propose two novel extensions based upon the BPL model – one to the model itself and one to its estimation – tailored to the objectives of the problem list analysis.

3.1 Ranker Fatigue

In some contexts, a ranker's list is a purposefully incomplete ordering of a subset of all possible items. In our case study, physicians stopped listing problems upon having decided that the already-listed problems adequately described the case abstracts. It is sensible therefore to model not just the ordering process but also the terminating process. This contrasts with standard PL/BPL models, which assume that $\ell_i \equiv v$. Notationally, this can be indicated by artificially extending the length of each ragged list $\boldsymbol{x}_i$ by one and filling in this additional item with 0, i.e. $x_{i\ell_i} \equiv 0$; this is not an actual item but rather indicates the list's termination.

Now the probability of selecting item $k = 0, \ldots, v$ in the sth stage, $s = 1, \ldots, \ell_i$, conditional on the previous $s-1$ stages is written as

$$\Pr(x_{is} = k|\mathcal{O}_{is}) = \frac{1_{[k\in\mathcal{O}_{is}]} \exp(\delta(s)\theta_k) + 1_{[k=0\cap s>1]} \exp(\theta_0)}{\sum_{j\in\mathcal{O}_{is}} \exp(\delta(s)\theta_j) + 1_{[s>1]} \exp(\theta_0)}. \tag{2}$$

Like the standard BPL model, this assumes that there are a finite number of v items to be ranked; however, it is not assumed that all rankers will rank all items. Rather, a new parameter θ_0 measures the 'fatigue' of ranker i beyond the first stage, and $\Theta = \{\theta_0, \theta_1, \ldots, \theta_v\}$ is length $v+1$. The number of identified elements, not counting the dampening function $\delta(s)$, is one less than the length of Θ, and we set $\min_{j:j>0}\{\theta_j\} \equiv 0$ to identify the model. At stage $s > 1$, ranker i will stop ordering items with probability $\exp(\theta_0)/(\sum_{j\in\mathcal{O}_{is}} \exp(\delta(s)\theta_j) + \exp(\theta_0))$. The probability of stopping increases with s as well as with the total weight of the items previously ordered. Let $\boldsymbol{\beta} = \{\Theta, \delta(\cdot)\}$ denote all parameters in the model. The log-likelihood of list $\boldsymbol{x}_i$ is the logarithm of its joint density:

$$\begin{aligned}\log f_i(\boldsymbol{\beta}) &= \sum_{s=1}^{\ell_i} \log \Pr(x_{is}|\mathcal{O}_{is}) \\ &= \theta_0 + \sum_{s=1}^{\ell_i - 1} \delta(s)\theta_{x_{is}} - \sum_{s=1}^{\ell_i} \log\Big(\sum_{j\in\mathcal{O}_{is}} \exp(\delta(s)\theta_j) + 1_{[s>1]} \exp(\theta_0)\Big),\end{aligned} \tag{3}$$

where $\mathcal{O}_{is}$ is as defined in Equation (1). This is the model we will use in our simulations study and analysis of the problem list data. We discuss the choice of dampening function $\delta(\cdot)$ in Section 3.2 and consider strategies for inducing sparsity in the fitted model in Section 3.3.

Remark 1 With the added fatigue parameter, the assumed minimum list length for any ranker is $\ell_i = 2$, corresponding to exactly one actual item ranked ($x_{i1} \in \{1, \ldots, v\}$) followed by a decision to stop ($x_{i2} = 0$). This is the rationale for having $1_{[k=0\cap s>1]}$ (as opposed to $1_{[k=0]}$) in the numerator of Equation (2), which gives that $\Pr(x_{i1} = 0|\mathcal{O}_{i1}) \equiv 0$ for any $\boldsymbol{\beta}$.

Remark 2 A reviewer observed that to be able to call equation (3) a regular likelihood, we must make an implicit conditional independence assumption, namely that at each stage, a ranker's probability is conditionally independent of all previous probabilities given $\mathcal{O}_{is}$. We do so here, similar to previous uses of the BPL model, e.g. Gormley and Murphy (2008).

3.2 Parameterization of $\delta(\cdot)$

In their choice of $\delta(\cdot)$ for the analysis of Irish presidential poll data, Gormley and Murphy (2008) placed no restrictions on $\delta(\cdot)$ apart from requiring $0 \leq \delta(s) \leq 1$ for all s, resulting in $v-2$ parameters to be estimated. The context of our analysis suggests that, at a minimum, $\delta(\cdot)$ should be non-increasing in its argument to reflect that strength of preference is non-increasing with stage, i.e. rank. For this reason, and also being cognizant of the statistical

cost of estimating many additional parameters, we constructed a two-parameter dampening function: $\delta(s) = \delta_2\delta_1^{s-1} + (1-\delta_2)^{2s-1}$, with scalar parameters $\delta_1, \delta_2 \in [0,1]$. This family contains dampening functions ranging from constant strength of preference ($\delta_1 = \delta_2 = 1$), strength of preference decreasing to a non-zero asymptote ($\delta_1 = 1; \delta_2 < 1$), or strength of preference decreasing to total lack of preference at lower ranks ($\delta_1 < 1$).

3.3 Estimating a Consensus Ordering

A standard maximum likelihood estimate (MLE) approach for estimating $\boldsymbol{\beta} = \{\Theta, \delta_1, \delta_2\}$ would calculate $\hat{\boldsymbol{\beta}}_{\text{MLE}} = \arg\max_{\boldsymbol{\beta}} \sum_{i=1}^{n} \log f_i(\boldsymbol{\beta})$ subject to the constraint that $\min_{j:j>0}\{\hat{\theta}_j\} \equiv 0$, so as to identify the model. However, even with this constraint, some of the parameters will still only be weakly identified, e.g. those corresponding to items appearing on only one observation's list, and their estimates will be close to zero. An ideal model estimation process would adaptively recognize these weakly identified parameters and set them all exactly equal to zero. Note that this is a different type of variable selection problem than is typical: setting $\hat{\theta}_k$ equal to zero in a BPL-type model does not remove the item from the fitted model but rather minimizes its relative weight. No item that has been ranked at least once can ever be removed entirely from the fitted model, i.e. by forcing $\hat{\theta}_k = -\infty$ or $\exp(\hat{\theta}_k) = 0$, without resulting in a zero-valued likelihood function. Rather, this variable selection problem is one of identifying the set of items whose corresponding parameter estimates should be smallest and co-equal. Keeping in mind the scientific objective of constructing a consensus ordered list, a natural definition is then the set of non-zero $\hat{\theta}_k$'s, sorted in decreasing order. If the data are disparate enough to suggest that rankers are effectively ordering items at random, then the consensus list may be small or even the empty set, i.e. no consensus.

A common technique for dimension reduction in a maximum likelihood framework is to subtract from the log-likelihood function a penalty function on the item weights, $g(\Theta, \lambda)$. For a given value of λ, we would then calculate the penalized MLE (PMLE), defined as $\hat{\boldsymbol{\beta}}(\lambda) = \arg\max_{\boldsymbol{\beta}} \{\sum_{i=1}^{n} \log f_i(\boldsymbol{\beta}) - g(\boldsymbol{\beta}, \lambda)\}$. Assuming the model is not to be penalized for estimating θ_0, the simplest possible BPL model would be $\theta_k \equiv 0$ and $\delta_1 = \delta_2 = 1$, and a LASSO-type penalty (Tibshirani, 1996) applied to a BPL model would take the form $g(\boldsymbol{\beta}, \lambda) = \lambda\left(\sum_k \theta_k + |\log \delta_1| + |\log \delta_2|\right)$ (note that if every θ_k wasn't non-negative by design, we would need $|\theta_k|$ instead of θ_k). Relative to standard maximum likelihood estimation, this penalty would shrink each θ_k down towards zero and δ_1 and δ_2 up towards 1, more so for larger values of λ; some elements may be shrunk entirely. This latter characteristic makes the LASSO a variable selection penalty. As noted in the first paragraph of this section, variable selection is a crucial feature in our context, but it is less evident that shrinkage of the item weights is required or even desirable. Because each θ_k is relatively defined, if a parameter estimate $\hat{\theta}_k$ gets set to zero, any larger parameter estimates will also need to be decreased in order to maintain the same implied probabilities. For example, consider a BPL model with three items, where the current parameter estimates are $\{\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3\} = \{0, \log(2.9), \log(1.1)\}$. The estimated probability of selecting item 2 at stage 1 is $2.9/(1+2.9+1.1) = 0.58$. If $\hat{\theta}_3$ is to be set to zero to reflect that items 1 and 3 are equally least important, then the corresponding estimate of $\hat{\theta}_2$ must also be changed to approximately $\log(2.76)$ in order to maintain this

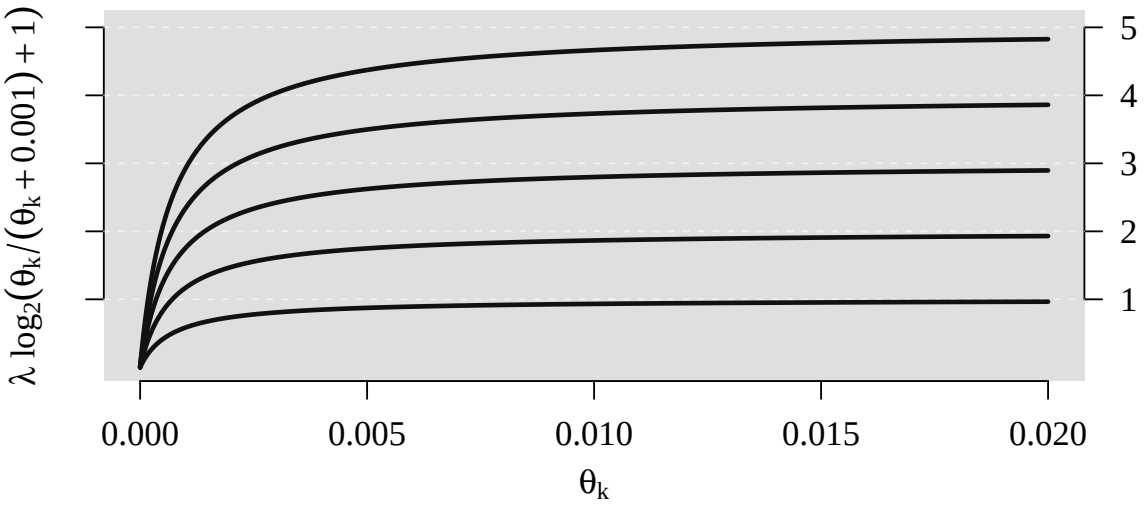

Figure 2: Seamless L_0 penalty under the default choice of $\tau = 0.001$ and different values of the penalty parameter (and asymptote) λ

probability: $2.76/(1+2.76+1) \approx 0.58$. That is, in order to change $\hat{\theta}_3$ from $\log(1.1)$ to 0 while maintaining the relative importance of $\hat{\theta}_2$, the latter must also be decreased. A LASSO-type penalty would induce *additional* shrinkage, beyond what was just described, and therefore may result in underfitting the model, i.e. not describing enough variability.

Variable selection without this additional shrinkage can be achieved with the L_0 penalty: $g(\boldsymbol{\beta}, \lambda) = \lambda \left(\sum_{k=1}^{v} 1_{[\theta_k \neq 0]} + 1_{[\delta_1 \neq 1]} + 1_{[\delta_2 \neq 1]} \right)$. This penalizes the log-likelihood for each additional parameter estimate that takes on a "non-simple" value by an amount λ, but the actual estimate does not further affect the penalty, i.e. there is no shrinkage. A computationally driven modification is called for here because $\lambda \left(\sum_{k=1}^{v} 1_{[\theta_k \neq 0]} + 1_{[\delta_1 \neq 1]} + 1_{[\delta_2 \neq 1]} \right)$ is a multivariate discontinuous function and therefore numerically difficult to use within a penalized likelihood framework. Dicker et al. (2013) created a continuous version, called the seamless L_0 penalty. Applied to our scenario, it is given by

$$g(\boldsymbol{\beta}, \lambda, \tau) = \lambda \sum_{k=1}^{v} \log_2 \left(\frac{\theta_k}{\theta_k + \tau} + 1 \right) + \lambda \log_2 \left(\frac{|\log \delta_1|}{|\log \delta_1| + \tau} + 1 \right) + \lambda \log_2 \left(\frac{|\log \delta_2|}{|\log \delta_2| + \tau} + 1 \right), \tag{4}$$

where $\tau > 0$ is an additional fixed constant parameter. In contrast to the discrete-valued L_0 penalty that is always equal either to 0 (for each $\theta_k = 0$ and $\delta_1, \delta_2 = 1$) or λ (for each $\theta_k > 0$ and $\delta_1, \delta_2 < 1$), the seamless L_0 penalty continuously transitions from 0 to λ, as illustrated in Figure 2. It becomes increasingly similar to the discontinuous L_0 penalty as τ is closer to 0.

3.4 Computational Implementation

We describe here our computational approach for fitting penalized BPL models using seamless L_0 penalties. All code was written in the `R` statistical environment (R Core Team, 2018; Wickham, 2017; Neuwirth, 2014; Li et al., 2018; Schimek et al., 2015) and is freely available via github (`https://github.com/psboonstra/RankModeling`). When g is an L_0-type penalty, maximizing $\sum_{i=1}^n \log f_i(\boldsymbol{\beta}) - g(\boldsymbol{\beta}, \lambda)$ is a non-convex optimization problem that is both computationally difficult and which admits the possibility of identifying local optima. These are the main challenges our algorithm must overcome.

As is typical in penalized estimation, we calculate the solution path for $\boldsymbol{\beta}$ under a grid of candidate values for λ. We apply a numerical coordinate ascent algorithm that iteratively cycles through all elements of $\boldsymbol{\beta}$ on a univariate basis, changing a given parameter estimate from its current value if doing so increases the penalized log-likelihood. After satisfying a specified convergence criterion to an estimate of $\boldsymbol{\beta}$ given the smallest value of λ, we use these values as a warm start for the next largest value of λ in the grid and so forth. The algorithm returns the entire solution path for $\boldsymbol{\beta}$ as a function of λ.

In more detail, suppose the current estimated value of $\boldsymbol{\beta} = \{\Theta, \delta_1, \delta_2\}$ at iteration m of the algorithm is denoted by $\hat{\Theta}^{(m)} = \{\hat{\theta}_0^{(m)}, \hat{\theta}_1^{(m)}, \ldots, \hat{\theta}_v^{(m)}\}$, $\hat{\delta}_1^{(m)}$, and $\hat{\delta}_2^{(m)}$. Given these values and λ, we calculate the penalized log-likelihood values when incrementing one parameter estimate by each value in a proposal sequence $\Gamma = \{\gamma_{-t}, \gamma_{-t+1}, \ldots, \gamma_{-1}, \gamma_0 \equiv 0, \gamma_1, \ldots, \gamma_{t-1}, \gamma_t\}$, where $\gamma_{-j} = -\gamma_j$ for all $j = 1, \ldots, t$. The inclusion of $\gamma_0 \equiv 0$ means that one proposal is to not change any values. If the parameter to be updated corresponds to an item, i.e. $\theta_1, \ldots, \theta_v$, then $\gamma_{\tilde{0}} = -\hat{\theta}_j^{(m)}$ is also added to Γ, so that every iteration includes proposing to set each item's parameter estimate to zero. Any proposals that would violate identifiability or model constraints, i.e. $\theta_k < 0$ or $\delta_1, \delta_2 \notin [0, 1]$, are truncated at the boundary of the constraint.

This step results in up to $2t + 2$ penalized log-likelihood calculations, and we identify $t_{\max} \in \{-t, -t+1, \ldots, 1, 0, \tilde{0}, 1, \ldots, t-1, t\}$, which is the index of Γ yielding the largest penalized log-likelihood (note that $\tilde{0}$ does not exist when updating θ_0, δ_1, or δ_2). We then set $\hat{\theta}_k^{(m+1)} \leftarrow \hat{\theta}_k^{(m)} + \gamma_{t_{\max}}$ (or $\hat{\delta}_j^{(m+1)} \leftarrow \hat{\delta}_j^{(m)} + \gamma_{t_{\max}}$) and repeat the step for another parameter. Each cycle consists of proceeding through a random permutation of all elements of $\hat{\Theta}^{(m)}$, $\hat{\delta}_1^{(m)}$, and $\hat{\delta}_2^{(m)}$, and the process restarts until a certain minimum number of consecutive cycles change all parameter estimates by less than some convergence criterion ϵ. We discuss the choice of Γ and all other required inputs at the end of this section.

The relative relationship between the parameters warrants also considering multivariate proposals to speed convergence and discourage the algorithm from getting stuck in local optima. We incorporated such proposals in our implementation. One proposal adds a single negative constant randomly taken from $\{\gamma_{-t}, \gamma_{-t+1}, \ldots, \gamma_{-1}\}$ to all $\hat{\theta}_k$'s, shifting them *towards*, but never less than, zero. A second multivariable proposal adds a single positive constant value randomly taken from $\{\gamma_1, \gamma_2, \ldots, \gamma_t\}$ to all non-zero $\hat{\theta}_k$'s as well as one randomly selected zero-valued $\hat{\theta}_k$, if there is more than one such zero-valued $\hat{\theta}_k$. A third proposal considers the current estimated item weights $\hat{\Theta}^{(m)}$ in increasing order and, with probability $1/16 = 0.0625$, exchanges the index of each pair of neighboring parameter estimates. Note

the proposal swaps parameter estimates based on their values, not their labels. For example, if $\{\hat{\theta}_1^{(m)}, \hat{\theta}_2^{(m)}, \hat{\theta}_3^{(m)}\} = \{0, \log(2.9), \log(1.1)\}$, the proposal would swap $\hat{\theta}_1^{(m)}$ and $\hat{\theta}_3^{(m)}$, i.e. $\{\log(1.1), \log(2.9), 0\}$. This probability of swapping is a tuning parameter, the specific value of which was arbitrarily selected. We also considered a fourth multivariate proposal for the dampening function when $\hat{\delta}_2^{(m)} < 1$. The proposal is $\tilde{\delta}_1 = \hat{\delta}_2^{(m)}\hat{\delta}_1^{(m)} + (1 - \hat{\delta}_2^{(m)})^3$ and $\tilde{\delta}_2 = 0$. The rationale is that the proposed dampening function is identical to the current dampening function at the first two (and most important) stages, but with a less complex formulation, since $\tilde{\delta}_2 = 0$. Each of these proposals is ad-hoc and, except for the last, stochastically made, but we emphasize that they are only ever accepted if doing so results in a larger penalized log-likelihood.

3.5 Default values

Our algorithm for approximating the maximized penalized log-likelihood requires choosing input values, most important being the grid of values of λ, the constant τ in equation (4), the proposal sequence Γ, and the convergence criterion ϵ. In our analyses, we used the default values described below, so that at a minimum a user need only provide the data, comprising a set of ordered lists.

The default choice of convergence criterion is $\epsilon_{\text{def}} = 0.001$, which also means that the number of significant digits retained by the algorithm, generally equal to $\lceil \log_{10}(1/\epsilon) \rceil$, is by default 3. For a default value of the proposal sequence Γ, used in both the univariate and multivariate proposals, the algorithm calculates the evenly spaced sequence of t values between $\log(\epsilon)$ and 0 and exponentiates it, setting the positive half, $\gamma_{1,\text{def}}, \gamma_{2,\text{def}}, \ldots, \gamma_{t,\text{def}}$, equal to the result (with the lower half being the symmetric values $-\gamma_{t,\text{def}}, \ldots, -\gamma_{1,\text{def}}$). The default choice of t, when not provided, is $t_{\text{def}} = \lceil \log_{10}(1/\epsilon) \rceil$, yielding $\Gamma_{\text{def}} = \{-1, -0.032, -0.001, 0, 0.001, 0.032, 1\}$. The default choice of τ is $\tau_{\text{def}} = \epsilon$. Finally, a default grid of λs is calculated with an initial run of the algorithm that identifies the smallest λ that yields the most parsimonious possible model, say, $\lambda_{\max}$, and then calculates the 200 evenly spaced values (on the log-scale) between $10^{-5}\lambda_{\max}$ and $\lambda_{\max}$.

Our implementation also allows for the user to specify multiple sets of initial parameter values, $\boldsymbol{\beta}^{(0)}$, or to request multiple randomly generated sets of initial values. The algorithm is independently run for each set of initial values, and the result of each separate run is reported. This allows for a straightforward assessment of the impact of starting values on the final converged parameter estimates. We used three sets of initial values in both our simulation study and our data analyses.

3.6 Model Selection

We consider two information criteria to select $\lambda > 0$. The small-sample Akaike Information Criterion (AIC, Akaike, 1973; Hurvich and Tsai, 1989) and the Bayesian Information Criterion (BIC, Schwarz et al., 1978) both resemble a "model fit + model complexity" tradeoff. Letting $\tilde{p}_\lambda = 1 + \sum_{k=1}^{v} 1_{[\hat{\theta}_k \neq 0]} + 1_{[\hat{\delta}_1 \neq 1]} + 1_{[\hat{\delta}_2 \neq 1]}$ denote the number of parameters in a fitted

model under a given λ (the constant 1 is for θ_0) and $\hat{\boldsymbol{\beta}}_\lambda$ denote all BPL parameter estimates under a given λ, they are both given by $-2\sum_{i=1}^{n}\log f_i(\hat{\boldsymbol{\beta}}_\lambda)+2h(\tilde{p}_\lambda)$, where $h(\tilde{p}_\lambda)=\frac{\tilde{p}_\lambda n}{(n-\tilde{p}_\lambda-1)_+}$ for the small-sample AIC, where $(\cdot)_+$ is the positive-part function, and $h(\tilde{p}_\lambda)=\log(n)\tilde{p}_\lambda/2$ for the BIC.

4 Simulation study

To evaluate the finite sample performance of the penalized BPL models, including our proposed estimation procedure, we conducted a simulation study. Our two main objectives were to (i) compare the ranking performance of our penalized BPL models against possible comparators and (ii) evaluate the ability of our penalized BPL models to estimate the true, unknown item weights and distinguish between non-trivial and trivial items. These are distinct objectives because some rank aggregation methods only result in a fully ranked list of all items, whereas we are additionally interested in demonstrating that our penalized BPL model results in better estimation of the underlying weights than the unpenalized counterpart and that unimportant items can be identified as such.

We generated 1000 simulated datasets for each of 36 scenarios: 12 models described in Table 1 and three dataset sizes (30, 100, or 500 rankers). Collectively, these scenarios cover a range of characteristics: number of items (v), number of rankers (n), degree of raggedness (θ_0), and the typical size and variation of item weights (θ_ks). For each simulated dataset and each scenario, we fit an unpenalized BPL model ($\lambda=0$) and penalized BPL models using AIC and BIC. For comparison, we evaluated our previous pairwise-LDRBO maximization (Krauss et al., 2016); the sample arithmetic mean ('AMean'), geometric mean ('GMean'), and median of the corresponding ranked lists; MC1, MC2, and MC3 (Lin, 2010; Schimek et al., 2015); the CEMC algorithms that identify the aggregated list that is, on average, closest to all observed lists (Lin and Ding, 2009; Schimek et al., 2015) using Spearman's ρ (labeled CEMC$_\rho$) or Kendall's τ (CEMC$_\tau$); and both the standard Mallows model (MM) and extended Mallows model (EMM) as implemented in Li et al. (2019). For calculating the sample means and medians, we assumed that each unranked item had rank equal to the average of the unused ranks. So, for example, when a list ranks seven items out of a possible $v=15$, then the eight unranked items each get assigned a rank of 11.5. This filling-in of missing ranks was only used for calculation of the sample mean and median ranks.

A challenge in comparing multiple different ranking approaches is that each returns a qualitatively different entity. That is, the BPL models estimate a set of item weights, the MC models estimate a stationary probability distribution taking on values in the unit simplex, the means and median summarize the ranks, and the remaining methods – LDRBO CEMC$_\rho$, CEMC$_\tau$, MM, and EMM – give an integer-valued ordering from most to least preferred. To compare these disparate results, we calculated a metric that we called the 'ordered root-mean-squared error (RMSE)'. Let $\theta_{(k)}$, $k=1,\ldots,v$, denote the true value of the kth largest item weight, and let $\theta_{(\hat{k})}$ denote the true value of the item weight the method ranks kth. Then, the *ordered RMSE* is defined by $\sqrt{\sum_{k=1}^{v}(\theta_{(k)}-\theta_{(\hat{k})})^2/v}$. The true values of each item

weight are always used but are out of order, and methods underperform according to the extent they mis-rank items with substantially different true weights. The advantage of this metric is that it is applicable across all ranking approaches. The disadvantage is that it is not equipped to handle tied ranks, which can occur in the penalized BPL results and the sample means or medians of the ranks. For the penalized BPL models, we break ties based upon the ordering of items in the solution path at $\lambda = 0$. For the sample means and medians, we break ties based upon the frequency that an item was ranked first; remaining ties were randomly resolved. The smallest possible ordered RMSE is 0, when all items have been ordered according to their true item weight.

As a subsequent finite-sample assessment focusing exclusively on the BPL models, we calculated additional quantitative measures of the ability of these fitted models to estimate the unknown item weights, which none of the comparator methods do. Specifically, let $\hat{\theta}_k$ be the estimate of the kth unknown item weight θ_k (note that these are no longer the order statistics). Then, we calculated the standard RMSE, defined as $\sqrt{\sum_{k=1}^{v}(\hat{\theta}_k - \theta_k)^2/v}$. This differs from the ordered RMSE described in the previous paragraph in that the ordered RMSE does not require a method to estimate the item weight itself, whereas the standard RMSE does. We also calculated the true positive rate (TPR), defined as $\left(\sum_{k=1}^{v} 1_{[\hat{\theta}_k>0]} \times 1_{[\theta_k>0]}\right) / \sum_{k=1}^{v} 1_{[\theta_k>0]}$, and the true negative rate (TNR), defined as $\left(\sum_{k=1}^{v} 1_{[\hat{\theta}_k=0]} \times 1_{[\theta_k=0]}\right) / \sum_{k=1}^{v} 1_{[\theta_k=0]}$. Finally, we calculated Youden's index, defined as TPR + TNR - 1, and the running time for each method.

4.1 Results

Table 2 gives the average values of the ordered RMSE multiplied by 1000 and then rounded to the nearest integer for readability. All values within 5% of each rowwise minimum are in **bold**. The penalized BPL models, labeled 'λ_{AIC}' and 'λ_{BIC}', are not optimal in all scenarios but generally compare favorably to the remaining methods. The best overall method, AMean, was bold in all rows, followed closely by MC3. The worst method was MC1, which was bold in just two rows.

The ordered RMSE metric does not directly characterize how well the item weights were estimated. To that end, the results in Table 3 make a direct comparison of the unpenalized and penalized BPL models in their ability to estimate the item weights. In contrast to Table 2, one or both penalized versions are nearly always preferred with regard to the standard RMSE metric. The penalized BPL models typically have a better TNR, whereas the unpenalized BPL model has a better TPR. On balance, the discriminatory ability of the penalized BPL models are better, as evidenced by higher values of Youden's index. Finally, the penalized BPL models require about twice as much running time.

Table 1: Description of twelve true BPL models for generating ragged, ordered lists used in the simulation study. The last column gives the first, second, and third quartiles from the distribution of actual list length under these models.

Label	θ_0	$\{\theta_k : \theta_k > 0, k > 0\}$	$\#\{\theta_k : \theta_k = 0, k > 0\}$	$\#\{\theta_k : k > 0\}$	$\ell_i : \{Q_1, Q_2, Q_3\}$
1	-1.0	$\{0.09 - 0.01(k-1)\}_{k=1}^{5}$	5	10	{6,9,9}
2	0.5	$\{0.09 - 0.01(k-1)\}_{k=1}^{5}$	5	10	{2,4,6}
3	-1.0	$\{1.5 - 0.1(k-1)\}_{k=1}^{9}$	1	10	{8,9,9}
4	0.5	$\{1.5 - 0.1(k-1)\}_{k=1}^{9}$	1	10	{3,6,8}
5	-1.0	$\{1.5\}_{k=1}^{5}$	5	10	{7,9,9}
6	0.5	$\{1.5\}_{k=1}^{5}$	5	10	{3,5,8}
7	-1.0	$\{0.09 - 0.01(k-1)\}_{k=1}^{9}$	11	20	{11,17,19}
8	0.5	$\{0.09 - 0.01(k-1)\}_{k=1}^{9}$	11	20	{4,7,12}
9	-1.0	$\{1.5 - 0.1(k-1)\}_{k=1}^{15}$	5	20	{13,18,19}
10	0.5	$\{1.5 - 0.1(k-1)\}_{k=1}^{15}$	5	20	{5,10,14}
11	-1.0	$\{1.5\}_{k=1}^{8}$	12	20	{13,18,19}
12	0.5	$\{1.5\}_{k=1}^{8}$	12	20	{5,9,14}

Table 2: Mean values of the *ordered RMSE* (multiplied by 1000 and then rounded to the nearest integer) for the unpenalized and penalized versions of the proposed BPL model against eleven comparators across 36 scenarios (12 generating models from Table 1 × three sample size configurations). All values within 5% of each rowwise minimum are in **bold**. For λ_{AIC} and λ_{BIC}, the tied zero-valued estimated items weights are resolved based upon the $\lambda = 0$ value in the solution path.

		BPL			LDRBO	AMean	GMean	Median	MC1	MC2	MC3	CEMC$_\rho$	CEMC$_\tau$	EMM	MM
Label	n	$\lambda = 0$	λ_{AIC}	λ_{BIC}		Rank	Rank	Rank							
1	30	**48**	**48**	**48**	**48**	**48**	**48**	**48**	51	**48**	**48**	**48**	**48**	**48**	**48**
1	100	**45**	**45**	**46**	**47**	**45**	**45**	**46**	51	**46**	**45**	**46**	**45**	**46**	**45**
1	500	**39**	**40**	**40**	43	**39**	**40**	43	51	49	**39**	**40**	**40**	**39**	**40**
2	30	**49**	**49**	**49**	**48**	**48**	**48**	**49**	51	**49**	**48**	**49**	**48**	**48**	**48**
2	100	**45**	**46**	**47**	**47**	**46**	**46**	**47**	51	51	**46**	**47**	**46**	**46**	**46**
2	500	**40**	**42**	43	43	**40**	**40**	43	51	51	**40**	43	**40**	**41**	**41**
3	30	**242**	247	247	347	**232**	244	256	571	247	**233**	287	249	267	246
3	100	**130**	**131**	**133**	267	**130**	137	155	571	147	**130**	179	145	149	141
3	500	**56**	**55**	**55**	145	**55**	58	86	571	74	**56**	75	65	60	64
4	30	268	274	272	333	**250**	265	278	569	305	**252**	359	268	291	265
4	100	**147**	**148**	154	282	**145**	**150**	169	572	209	**145**	247	154	167	155
4	500	**62**	**62**	**63**	166	**63**	**65**	89	571	132	**63**	107	70	68	69
5	30	15	19	18	275	**13**	18	34	1041	22	14	44	24	53	25
5	100	**0**	**0**	**0**	38	**0**	**0**	**0**	1043	**0**	**0**	**0**	**0**	**0**	**0**
5	500	1	**0**	**0**	**0**	**0**	**0**	**0**	1044	**0**	**0**	**0**	**0**	**0**	**0**
6	30	38	31	32	324	**25**	40	60	1037	78	30	135	50	87	44
6	100	4	**0**	**0**	58	**0**	**0**	**0**	1043	**0**	**0**	3	**0**	**0**	**0**
6	500	4	**0**	**0**	**0**	**0**	**0**	**0**	1044	**0**	**0**	**0**	**0**	**0**	**0**
7	30	**41**	**42**	**42**	**41**	**41**	**41**	**41**	**43**	**41**	**41**	**41**	**41**	**41**	**41**
7	100	**40**	**40**	**40**	**40**	**40**	**40**	**40**	43	**40**	**40**	**40**	**40**	**40**	**40**
7	500	**37**	**37**	**38**	**38**	**36**	**37**	**38**	43	42	**36**	**37**	**37**	**37**	**37**
8	30	**41**	**41**	**41**	**41**	**41**	**41**	**41**	**43**	**42**	**41**	**41**	**41**	**41**	**41**
8	100	**41**	**41**	**41**	**41**	**40**	**40**	**41**	43	43	**40**	**41**	**40**	**40**	**40**
8	500	**38**	39	39	**38**	**37**	**37**	**38**	43	43	**37**	**38**	**37**	**37**	**37**
9	30	**333**	338	**335**	423	**320**	**325**	352	716	**327**	**320**	366	**333**	345	**330**
9	100	**192**	**193**	**193**	306	**192**	**196**	218	716	**201**	**192**	234	207	213	201
9	500	**84**	**84**	**85**	182	**86**	**87**	107	716	107	**85**	116	98	102	94
10	30	357	368	366	420	**337**	**338**	367	714	392	**337**	425	**347**	376	356
10	100	237	**210**	220	325	**204**	**207**	232	716	337	**204**	290	216	234	220
10	500	149	**95**	100	199	**93**	**94**	120	716	330	**92**	154	104	113	101
11	30	114	117	112	426	**105**	110	185	1031	121	**105**	218	157	208	150
11	100	0	**0**	**0**	95	**0**	**0**	1	1032	0	**0**	1	3	4	0
11	500	**0**	**0**	**0**	**0**	**0**	**0**	**0**	1032	**0**	**0**	**0**	**0**	**0**	**0**
12	30	198	197	179	486	**147**	166	231	1024	249	**148**	376	197	277	203
12	100	43	**0**	1	177	**0**	1	3	1032	42	**0**	28	4	8	1
12	500	40	**0**	**0**	1	**0**	**0**	**0**	1031	4	**0**	**0**	**0**	**0**	**0**

Table 3: Five operating characteristics comparing the unpenalized and penalized versions of the BPL model across 36 scenarios (12 generating models from Table 1 × three sample size configurations). RMSE , TPR, TNR, and Youden are all multiplied by 100 and rounded to the nearest integer. Each value of an operating characteristic that is within 5% of the better of the two is in **bold**.

		RMSE×100			TPR×100			TNR×100			Youden×100			Run Time (sec.)		
Label	n	$\lambda=0$	λ_{AIC}	λ_{BIC}	$\lambda=0$	λ_{AIC}	λ_{BIC}	$\lambda=0$	λ_{AIC}	λ_{BIC}	$\lambda=0$	λ_{AIC}	λ_{BIC}	$\lambda=0$	λ_{AIC}	λ_{BIC}
1	30	1267	246	**205**	**91**	13	9	14	**89**	**92**	**5**	2	1	**4**	20	20
1	100	459	137	**91**	**87**	20	10	20	88	**94**	7	**8**	4	**5**	26	26
1	500	156	76	**58**	**86**	29	8	27	89	**97**	13	**18**	5	**10**	72	72
2	30	1184	214	**167**	**87**	12	9	15	**91**	**94**	2	**3**	2	**3**	17	17
2	100	492	147	**100**	**82**	19	10	24	87	**92**	**6**	**7**	2	**3**	22	22
2	500	212	81	**63**	**79**	25	10	35	89	**95**	13	**15**	5	**5**	61	61
3	30	**880**	970	1000	**98**	36	32	82	**100**	**100**	**80**	36	31	**5**	20	20
3	100	**358**	390	594	**100**	93	69	**99**	**100**	**100**	**99**	93	69	**7**	26	26
3	500	157	**139**	**138**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**24**	75	75
4	30	**907**	**941**	972	**96**	30	26	75	**100**	**100**	**71**	30	26	**4**	18	18
4	100	**349**	412	673	**99**	89	57	**98**	**100**	**100**	**97**	89	57	**6**	25	25
4	500	155	**143**	**144**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**18**	60	60
5	30	725	**375**	**392**	**100**	**99**	**99**	23	**97**	**96**	23	**96**	**94**	**4**	21	21
5	100	356	**198**	**189**	**100**	**100**	**100**	22	92	**98**	22	92	**98**	**7**	31	31
5	500	156	99	**88**	**100**	**100**	**100**	22	89	**100**	22	89	**100**	**20**	89	89
6	30	823	**431**	**442**	**100**	**96**	**96**	25	**96**	**95**	25	**92**	**91**	**4**	19	19
6	100	412	**225**	**220**	**100**	**100**	**100**	25	91	**98**	25	91	**98**	**5**	27	27
6	500	185	103	**89**	**100**	**100**	**100**	26	90	**99**	26	90	**99**	**13**	75	75
7	30	4283	1137	**1045**	**94**	17	11	7	85	**90**	1	**1**	1	**16**	108	108
7	100	997	240	**112**	**95**	21	8	7	84	**94**	2	**5**	2	**24**	127	127
7	500	326	115	**57**	**91**	26	7	14	83	**95**	5	**9**	2	**60**	336	336
8	30	3094	433	**300**	**91**	14	7	11	88	**93**	**2**	2	1	**13**	96	96
8	100	948	214	**116**	**91**	17	7	11	86	**93**	**2**	2	1	**12**	74	74
8	500	550	85	**72**	**85**	12	7	20	**91**	**94**	**5**	3	2	**17**	183	183
9	30	1595	**803**	**839**	**97**	33	33	17	**98**	**98**	14	**31**	**31**	**17**	94	94
9	100	471	**284**	346	**98**	70	56	20	**97**	**100**	18	**66**	55	**27**	137	137
9	500	190	**125**	150	**99**	91	80	23	92	**100**	22	**82**	**80**	**94**	479	479
10	30	1597	761	**720**	**96**	32	29	20	**97**	**98**	16	**29**	27	**14**	82	82
10	100	719	**442**	464	**96**	77	61	30	79	**91**	26	**56**	52	**11**	76	76
10	500	520	290	**270**	**96**	**95**	87	41	61	**81**	36	55	**68**	**19**	217	217
11	30	1238	**446**	**450**	**100**	92	**96**	12	**97**	**93**	12	**90**	**89**	**16**	86	86
11	100	591	244	**205**	**100**	**100**	**100**	14	89	**97**	14	89	**97**	**21**	105	105
11	500	352	195	**122**	**100**	**100**	**100**	24	66	**98**	24	66	**98**	**46**	274	274
12	30	1356	**616**	**599**	**100**	87	91	17	**96**	**92**	17	**82**	**82**	**12**	72	72
12	100	827	380	**322**	**100**	**100**	**100**	26	80	**91**	26	80	**91**	**9**	73	73
12	500	688	324	**248**	**100**	**100**	**100**	33	57	**86**	33	57	**86**	**15**	195	195

5 Data analysis: Problem lists

We now analyze the motivating problem list data. Tables 4–6 give the parameter estimates for all models for cases A–C, respectively. The BIC-based results are given for comparison, and we focus on the AIC-based fitted models. Figures S1–S3 in the Supplement give the full solution paths from our algorithm, with the AIC and BIC solutions noted. Tables 4–6 also include the consensus problem list from Krauss et al. (2016) and the other comparator methods evaluated in the simulation study. To alternatively characterize the extent of physician consensus, Figure 3 plots the probability of the most preferred item at each stage according to the AIC-estimated BPL model fit, conditional on all prior stages having also selected the most preferred item. Each such modal list continues until the item "0" is selected.

Of the 28 unique problems listed for case A, 10 were estimated to have non-zero weights

according to the AIC-selected model; these 10 problems are the consensus problem list according to the model. The BIC-selected model included 12 problems. Using AIC, the estimate of δ_1 was 1 and the estimate of δ_2 was 0.62, suggesting that relative preferences quickly decrease and level off at about 2/3 their starting values. For example, at stage 3, the dampening function evaluates to $\delta(3) = 0.62 + 0.38^5 \approx 0.63$, and the relative weight of, say, ANEMIA at this stage (supposing it has not yet been ranked) decreases to $5.30 \times 0.63 = 3.3$. The BPL models' ranks agree with the length-8 consensus problem list reported in Krauss et al. (2016) as well as with AMean and GMean on the three most important problems, and they nearly agree with Median, MC2, MC3, CEMC$_\rho$, CECM$_\tau$, and MM. Among the comparator methods, the BPL models agreed most closely with AMean. There was disagreement with between some methods at lower ranks, however. The BPL models did not put one problem from Krauss et al. – HYPOXIA – anywhere in their consensus list. The fatigue parameter θ_0 was estimated to be 2.57 in the AIC-selected model. This value does not translate into an expected list length, which is a multidimensional function of all elements of $\boldsymbol{\beta}$. Thus, we simulated many lists from the fitted model to characterize the distribution of list lengths. The first, second, and third quartiles of the length of these simulated lists was $(4, 6, 9)$, compared to values of $(5, 8, 9)$ for the observed case A data. From Figure 3, the most preferred item at stage 1 (PNEUMONIA) is estimated to be selected with probability about 0.57, decreasing to about 0.20 for subsequent stages. This seems to disagree with the empiric proportion of physicians who ranked PNEUMONIA first, which was $26/32 \approx 0.82$. This model misspecification is likely due to the fact that two physicians ranked it 4th and four others never ranked it. The probabilities sometimes *increase* with stage due to the effect of the dampening function.

Table 4: Parameter estimates from fitted penalized Benter-Plackett-Luce (BPL) models applied to **case A** data, ordered by estimated values of θ_k from using $\lambda = \lambda_{\text{AIC}}$. The final row, $\tilde{p}_\lambda$, gives the number of non-zero parameters in the estimated model. For comparison, the remaining columns give the ranked list of problems according to the listed alternative algorithm or model; comparator entries in **bold** indicate ranks that were either (i) discrepant with λ_{AIC} by not more than 1 position or (ii) ranks that were larger than the total number of non-zero estimated weights according to λ_{AIC} and which λ_{AIC} estimated to be zero. The parenthetical numbers in the AMean, GMean, and Median rank columns correspond to the summary rank of that problem, and the parenthetical numbers in the MC columns correspond to the estimated transition probability for the stationary distribution ($\times 100$).

	BPL			LDRBO	AMean	GMean	Median	MC1	MC2	MC3	CEMC_ρ	CEMC_τ	EMM	MM
Problem / Parameter	$\lambda = 0$	λ_{AIC}	λ_{BIC}		Rank	Rank	Rank							
θ_0	3.95	2.57	2.78											
PNEUMONIA	9.54	6.86	7.32	**1**	**1(3.3)**	**1(1.6)**	**1(1.0)**	**1(5.6)**	**1(19.8)**	**1(15.2)**	**1**	**1**	**1**	**1**
DIABETES MELLITUS	7.94	5.31	5.74	**2**	**2(5.5)**	**2(4.4)**	**3(5.0)**	5(3.9)	**3(10.5)**	**3(8.8)**	**3**	**3**	4	**3**
ANEMIA	7.93	5.30	5.74	**3**	**3(6.5)**	**3(4.6)**	**2(4.0)**	**2(4.5)**	**2(14.1)**	**2(9.1)**	**2**	**2**	**2**	**2**
DEPRESSION WITH ANXIETY	7.15	4.55	4.98	**5**	**4(8.2)**	**4(6.9)**	**4(7.0)**	9(3.7)	**4(7.8)**	**4(6.3)**	7	**4**	**5**	**5**
OSTEOARTHRITIS	6.64	4.07	4.49	**6**	**5(10.1)**	**6(9.1)**	**5(8.0)**	12(3.6)	**5(5.8)**	**6(4.9)**	8	**6**	**6**	8
SPLENOMEGALY	6.42	3.79	4.21	4	**6(10.8)**	**5(8.0)**	**7(12.5)**	3(4.1)	**7(3.9)**	**5(5.5)**	4	**5**	3	4
POST MENOPAUSAL ON HRT	6.25	3.68	4.09		**7(11.4)**	9(9.8)	**6(9.5)**	15(3.5)	**6(4.7)**	9(4.4)	13	9	**8**	9
RENAL FAILURE	5.90	3.29	3.70	**7**	**8(12.3)**	**8(9.6)**	**8(16.8)**	**7(3.8)**	**8(2.0)**	**8(4.5)**	5	**7**	**9**	**7**
SYSTOLIC MURMUR	5.84	3.22	3.64		**9(12.4)**	7(9.3)	**10(17.0)**	4(3.9)	**9(2.0)**	7(4.5)	6	**8**	7	6
HISTORY OF SMOKING	4.76	2.26	2.65		**10(15.1)**	**11(13.9)**	**9(17.0)**	16(3.4)	21(1.5)	**11(2.7)**	25	**10**	**10**	**10**
CHEST PAIN	4.17	0.00	2.00		**11(15.7)**	**10(13.1)**	**12(18.0)**	8(3.8)	**14(1.6)**	**10(2.7)**	**15**	**12**	**12**	**11**
LOWER EXTREMITY EDEMA	4.10	0.00	2.00		**12(15.8)**	**12(14.4)**	**11(17.8)**	**14(3.5)**	**20(1.5)**	**12(2.5)**	9	**11**	**11**	**12**
IRON DEFICIENCY	3.37	0.00	0.00		**13(16.6)**	**15(15.7)**	**13(18.0)**	**22(3.3)**	**15(1.6)**	**14(2.1)**	**11**	**14**	**13**	**14**
HYPOXEMIA	3.04	0.00	0.00		**14(16.7)**	**13(14.9)**	**17(18.2)**	6(3.8)	**17(1.6)**	**13(2.2)**	**24**	**13**	**20**	**13**
HYPOPHOSPHATEMIA	2.87	0.00	0.00		**16(17.0)**	**19(16.4)**	**15(18.0)**	**23(3.3)**	**11(1.6)**	**17(1.9)**	10	**16**	**16**	**16**
THROMBOCYTOPENIA	2.83	0.00	0.00		**15(17.0)**	**18(16.4)**	**14(18.0)**	**21(3.3)**	**24(1.5)**	**16(2.0)**	**16**	**15**	**15**	**15**
SHORT OF BREATH	2.68	0.00	0.00		**17(17.2)**	**14(15.5)**	**18(18.5)**	10(3.6)	**19(1.6)**	**15(2.0)**	**27**	**17**	**19**	**18**
HYPOXIA	2.36	0.00	0.00	8	**18(17.2)**	**16(15.8)**	**21(18.5)**	**19(3.4)**	**12(1.6)**	**19(1.9)**	**28**	**18**	**23**	**17**
PANCYTOPENIA	2.30	0.00	0.00		**19(17.3)**	**17(15.9)**	**23(18.5)**	**11(3.6)**	**13(1.6)**	**18(1.9)**	**19**	**20**	**24**	**19**
HYPOALBUMINEMIA	2.14	0.00	0.00		**20(17.4)**	**20(16.9)**	**16(18.2)**	**25(3.2)**	**28(1.5)**	**20(1.8)**	**20**	**19**	**14**	**20**
FEVER	1.15	0.00	0.00		**21(17.7)**	**21(16.9)**	**19(18.5)**	**13(3.5)**	**22(1.5)**	**21(1.7)**	**14**	**22**	**21**	**21**
CONGESTIVE HEART FAILURE	1.10	0.00	0.00		**23(17.8)**	**22(17.0)**	**27(18.5)**	**18(3.4)**	10(1.6)	**23(1.7)**	**22**	**27**	**27**	**24**
PULMONARY EDEMA	1.07	0.00	0.00		**25(17.8)**	**23(17.0)**	**25(18.5)**	**17(3.4)**	**18(1.6)**	**24(1.6)**	**17**	**25**	**28**	**26**
PULMONARY EMBOLISM	0.95	0.00	0.00		**22(17.7)**	**24(17.3)**	**24(18.5)**	**20(3.3)**	**16(1.6)**	**22(1.7)**	**23**	**21**	**25**	**22**
TACHYCARDIA	0.90	0.00	0.00		**24(17.8)**	**26(17.5)**	**28(18.5)**	**27(3.0)**	**23(1.5)**	**25(1.6)**	**12**	**28**	**22**	**23**
DEPRESSION	0.86	0.00	0.00		**26(17.9)**	**25(17.5)**	**20(18.5)**	**24(3.2)**	**26(1.5)**	**26(1.6)**	**21**	**23**	**17**	**27**
HIGH HAPTOGLOBIN	0.86	0.00	0.00		**27(17.9)**	**28(17.7)**	**22(18.5)**	**28(2.7)**	**25(1.5)**	**28(1.6)**	**18**	**24**	**26**	**25**
ANXIETY	0.00	0.00	0.00		**28(17.9)**	**27(17.6)**	**26(18.5)**	**26(3.0)**	**27(1.5)**	**27(1.6)**	**26**	**26**	**18**	**28**
δ_1	0.99	1.00	1.00											
δ_2	0.61	0.62	0.61											
λ	0	3.38	1.19											
$\tilde{p}_\lambda$	30	12	14											

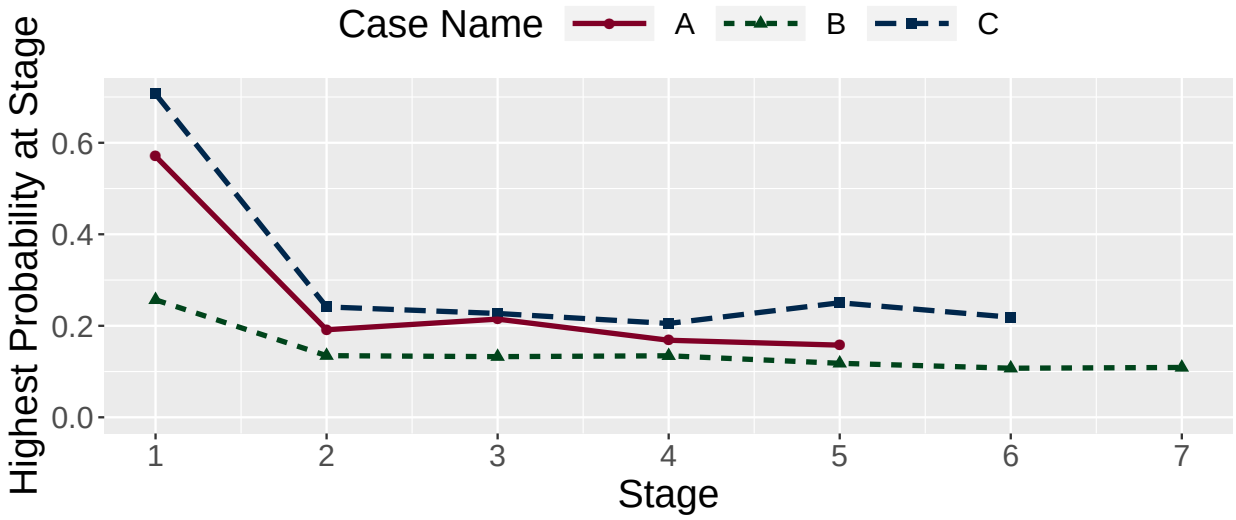

Figure 3: The probability of selectingn the most preferred item at each stage according to the AIC-estimated BPL model fit in Tables 4–6, conditional on all prior stages having also selected the most preferred item. Each such modal list continues until the item "0" is selected.

Case B, given in Table S7, was the most challenging, consistent with the a priori expectation in the protocol design. There were 47 unique problems appearing in at least one of the 32 lists, and 14 unique problems were ranked highest on at least one list. DIABETIC KETOACIDOSIS had the largest log-odds ratio of 4.56 (AIC) or 4.84 (BIC), both approximately 0.9 larger than the next highest ranked problem, RENAL FAILURE. Beyond the first rank, the difference in log-odds ratios between consecutive problems was even smaller, e.g. 0.16 between ranks 2 and 3, 0.13 between ranks 3 and 4, and 0.28 between ranks 4 and 5, reflecting uncertainty on the part of the physicians regarding which items to rank where. The AIC-selected consensus problem list, i.e. those items with strictly positive log-odds ratios, had length 16, similar to the length of the LDRBO-based consensus list. However, there was significant reordering of the problems: the top four problems of the AIC-based list were ranked 2nd, 5th, 4th, and 8th, respectively, on the LDRBO list. Further, three problems in the AIC-selected consensus list were not in the LDRBO-based list. The AIC-based list agreed to a large extent with AMean, GMean, MC3 and CECM$_\tau$. The fatigue parameter θ_0 was estimated to be 2.79, which together with the remaining parameter estimates, yields expected quartiles for the list length of $(5, 10, 14)$, compared to observed quartiles of $(8, 10, 12.5)$. In agreement with these findings, Figure 3 gives that the most preferred item at stage 1 (DIABETIC KETOACIDOSIS) is estimated to selected with probability about 0.26, compared to an observed proportion of $8/32 = 0.25$.

Remark 3 The LDRBO-based consensus list for Case B, given in Table S7 differs from that reported in Krauss et al. (2016): the 9th and 10th ranked items are swapped, and the 15th ranked item is different. For this paper we wrote a new algorithm for optimizing consensus, and it identified a list having a slightly larger (i.e. better) median pairwise LDRBO with the 32 physician lists: 0.584 here versus 0.581 reported in Krauss, et al.

Table 5: Parameter estimates from fitted penalized Benter-Plackett-Luce (BPL) models applied to **case B** data, ordered by estimated values of θ_k from using $\lambda = \lambda_{\text{AIC}}$. The final row, $\tilde{p}_\lambda$, gives the number of non-zero parameters in the estimated model. For comparison, the remaining columns give the ranked list of problems according to the listed alternative algorithm or model; comparator entries in **bold** indicate ranks that were either (i) discrepant with λ_{AIC} by not more than 1 position or (ii) ranks that were larger than the total number of non-zero estimated weights according to λ_{AIC} and which λ_{AIC} estimated to be zero. The parenthetical numbers in the AMean, GMean, and Median rank columns correspond to the summary rank of that problem, and the parenthetical numbers in the MC columns correspond to the estimated transition probability for the stationary distribution ($\times 100$).

	BPL			LDRBO	Mean	GMean	Median	MC1	MC2	MC3	CEMC$_\rho$	CEMC$_\tau$	EMM	MM
Problem / Parameter	$\lambda = 0$	λ_{AIC}	λ_{BIC}		Rank	Rank	Rank							
θ_0	3.50	2.79	3.07											
DIABETIC KETOACIDOSIS	5.27	4.56	4.84	**2**	**1(4.7)**	**1(3.0)**	**1(3.0)**	**1(3.6)**	**1(12.7)**	**1(8.85)**	**1**	**1**	**1**	**1**
RENAL FAILURE	4.39	3.68	3.96	5	**2(9.0)**	**2(6.5)**	4(6.0)	**3(2.6)**	5(6.4)	**2(5.98)**	4	**2**	5	4
SPONATENOUS BACTERIAL PERITONITIS	4.23	3.52	3.80	**4**	**4(11.5)**	**3(6.5)**	**2(5.0)**	**2(2.7)**	**2(8.5)**	**3(5.59)**	**2**	**3**	**3**	**2**
CIRRHOSIS DUE TO ALCOLHOL	4.10	3.39	3.67	8	**3(9.9)**	**5(7.7)**	**5(8.0)**	11(2.3)	6(5.1)	**4(5.16)**	7	**5**	7	7
MAXILLARY SINUS MASS	3.83	3.11	3.39	1	**5(13.3)**	**4(7.2)**	3(5.5)	**5(2.5)**	3(8.5)	**5(4.92)**	3	**4**	**4**	3
SCHIZOPHRENIA	3.60	2.89	3.17	11	**6(13.9)**	8(11.4)	**7(11.0)**	9(2.3)	**7(4.3)**	8(3.84)	10	**7**	8	8
ENCEPHELOPATHY	3.50	2.79	3.07	3	**8(16.7)**	**6(8.6)**	10(26.5)	**8(2.3)**	10(1.9)	**7(4.08)**	**6**	**6**	2	5
MULTIPLE CRANIAL NERVE PALSIES	3.46	2.75	3.03	6	**7(15.9)**	**7(10.0)**	6(8.5)	4(2.5)	4(7.0)	6(4.12)	5	**9**	6	6
HYPERTENSION	3.14	2.43	2.71	12	**9(18.2)**	11(15.3)	**9(15.5)**	13(2.2)	**8(3.2)**	**9(2.90)**	12	**10**	13	13
HISTORY OF IV DRUG USE	3.14	2.42	2.71	14	**10(18.5)**	13(16.0)	8(14.0)	22(2.1)	**9(3.1)**	12(2.80)	14	14	14	14
HYPONATREMIA	3.00	2.29	2.57	7	**11(19.9)**	**12(15.5)**	15(28.0)	**12(2.3)**	13(1.2)	**11(2.81)**	8	**12**	**12**	**12**
HYPERKALEMIA	2.89	2.17	2.46	10	14(20.7)	14(16.4)	**12(27.8)**	10(2.3)	14(1.2)	14(2.59)	9	**13**	**11**	**11**
TOBACCO USE	2.87	2.15	2.44		**13(20.5)**	15(18.0)	11(26.5)	15(2.2)	15(1.2)	15(2.43)	16	16	15	16
SINUSITIS	2.73	2.03	2.31	9	12(20.5)	9(13.6)	**13(28.0)**	6(2.4)	11(1.3)	10(2.88)	44	8	10	9
MENINGITIS	2.74	2.03	2.31		**15(21.0)**	10(14.0)	**14(28.0)**	7(2.3)	12(1.2)	13(2.79)	13	11	9	10
SYSTOLIC MURMUR	2.42	1.71	2.00		**16(23.5)**	**16(21.1)**	**16(28.0)**	21(2.1)	**16(1.2)**	**16(1.94)**	11	**15**	**16**	**15**
HX ALCOHOL ABUSE	2.05	0.00	1.62		**18(25.2)**	**20(23.4)**	**18(28.5)**	**26(2.1)**	**20(1.1)**	**18(1.62)**	15	**17**	**17**	**18**
ORBIT FRACTURE	1.93	0.00	1.50	13	**17(25.0)**	**17(21.4)**	**36(29.0)**	**18(2.2)**	**18(1.1)**	**17(1.78)**	**17**	**19**	**19**	**17**
HX GUN SHOT WOUND	1.84	0.00	1.42		**19(26.1)**	**22(24.8)**	**17(28.5)**	**29(2.1)**	**27(1.0)**	**19(1.46)**	**46**	**18**	**18**	**19**
ANEMIA	1.49	0.00	0.00		**20(26.8)**	**24(25.2)**	**42(29.0)**	**30(2.1)**	**17(1.1)**	**22(1.37)**	**19**	**23**	**20**	**21**
DIABETES MELLITUS	1.14	0.00	0.00		**22(27.1)**	**19(23.1)**	**22(29.0)**	**24(2.1)**	**41(1.0)**	**21(1.41)**	**45**	**20**	**35**	**24**
FEVER	1.14	0.00	0.00	15	**21(26.9)**	**18(22.0)**	**19(29.0)**	**19(2.2)**	**19(1.1)**	**20(1.45)**	**38**	**22**	**27**	**20**
R ORBITAL FRACTURE	1.12	0.00	0.00		**23(27.5)**	**25(26.2)**	**41(29.0)**	**28(2.1)**	**28(1.0)**	**25(1.25)**	**39**	**21**	**22**	**22**
THROMBOCYTOPENIA	0.77	0.00	0.00		**28(28.2)**	**34(27.6)**	**37(29.0)**	**43(1.8)**	**25(1.0)**	**28(1.12)**	**30**	**24**	**21**	**27**
MUCORMYCOSIS	0.75	0.00	0.00		**24(27.7)**	**21(24.6)**	**23(29.0)**	14(2.2)	**35(1.0)**	**23(1.27)**	**24**	**25**	**28**	**25**
SEPSIS	0.75	0.00	0.00		**25(27.7)**	**23(25.1)**	**24(29.0)**	16(2.2)	**40(1.0)**	**24(1.26)**	**37**	**28**	**40**	**23**
PALATAL LESION	0.75	0.00	0.00		**26(27.9)**	**27(26.5)**	**28(29.0)**	**31(2.0)**	**23(1.0)**	**26(1.19)**	**21**	**26**	**26**	**26**
ANEMIA AND THROMBOCYTOPENIA	0.73	0.00	0.00		**27(28.1)**	**32(27.2)**	**30(29.0)**	**36(1.9)**	**22(1.0)**	**27(1.14)**	**35**	**27**	**23**	**29**
TACHYCARDIA	0.05	0.00	0.00		**42(28.8)**	**44(28.5)**	**44(29.0)**	**46(1.6)**	**26(1.0)**	**44(1.02)**	**18**	**45**	**24**	**32**
HYPERSOSMOLAR COMA	0.04	0.00	0.00		**29(28.5)**	**29(27.0)**	**38(29.0)**	**27(2.1)**	**21(1.0)**	**30(1.09)**	**36**	**41**	**41**	**28**
PLASMA PROTEIN DISORDER	0.03	0.00	0.00		**37(28.7)**	**40(28.2)**	**33(29.0)**	**40(1.8)**	**34(1.0)**	**40(1.05)**	**22**	**44**	**29**	**33**
POTENTIAL CVA	0.02	0.00	0.00		**30(28.6)**	**26(26.5)**	**20(29.0)**	**17(2.2)**	**36(1.0)**	**29(1.10)**	**43**	**36**	**42**	**30**
HEPTO-RENAL SYNDROME	0.01	0.00	0.00		**31(28.6)**	**33(27.4)**	**25(29.0)**	**32(2.0)**	**24(1.0)**	**34(1.07)**	**40**	**29**	**37**	**31**
LEUKOCYTOSIS	0.01	0.00	0.00		**33(28.7)**	**35(27.6)**	**47(29.0)**	**34(2.0)**	**29(1.0)**	**35(1.07)**	**33**	**32**	**38**	**34**
CEREBROVASCULAR ACCIDENT	0.01	0.00	0.00		**38(28.8)**	**39(28.1)**	**27(29.0)**	**38(1.8)**	**47(1.0)**	**39(1.05)**	**20**	**37**	**32**	**37**
DEHYDRATION	0.01	0.00	0.00		**36(28.7)**	**36(27.7)**	**29(29.0)**	**33(2.0)**	**45(1.0)**	**36(1.06)**	**23**	**31**	**33**	**35**
ALCOHOLISM	0.01	0.00	0.00		**41(28.8)**	**41(28.2)**	**26(29.0)**	**39(1.8)**	**39(1.0)**	**41(1.04)**	**28**	**40**	**43**	**40**
ALCOHOLIC CIRRHOSIS WITH SBP	0.00	0.00	0.00		**32(28.7)**	**30(27.1)**	**45(29.0)**	**20(2.1)**	**30(1.0)**	**33(1.08)**	**26**	**30**	**39**	**36**
SMOKING	0.00	0.00	0.00		**39(28.8)**	**38(28.0)**	**39(29.0)**	**37(1.9)**	**44(1.0)**	**38(1.05)**	**25**	**43**	**34**	**39**
ASCITES	0.00	0.00	0.00		**34(28.7)**	**31(27.1)**	**31(29.0)**	**25(2.1)**	**42(1.0)**	**31(1.09)**	**42**	**38**	**36**	**38**
ALCOHOLIC CIRRHOSIS WITH ASCITES	0.00	0.00	0.00		**35(28.7)**	**28(26.6)**	**21(29.0)**	**23(2.1)**	**46(1.0)**	**32(1.09)**	**32**	**33**	**47**	**41**
RENAL FAILURE WITH HYPERKALEMIA	0.00	0.00	0.00		**40(28.8)**	**37(27.9)**	**43(29.0)**	**35(2.0)**	**32(1.0)**	**37(1.06)**	**27**	**34**	**45**	**42**
POLYSUBSTANCE ABUSE	0.00	0.00	0.00		**44(28.8)**	**43(28.3)**	**40(29.0)**	**41(1.8)**	**33(1.0)**	**42(1.04)**	**41**	**39**	**46**	**47**
PROTEIN CALORIE MALNUTRITION	0.00	0.00	0.00		**43(28.8)**	**42(28.3)**	**34(29.0)**	**42(1.8)**	**43(1.0)**	**43(1.03)**	**29**	**46**	**31**	**44**
MALNUTRITION	0.00	0.00	0.00		**45(28.9)**	**45(28.5)**	**35(29.0)**	**44(1.7)**	**31(1.0)**	**46(1.01)**	**34**	**35**	**25**	**43**
HEPATOMEGALY	0.00	0.00	0.00		**46(28.9)**	**46(28.6)**	**46(29.0)**	**45(1.6)**	**38(1.0)**	**45(1.01)**	**31**	**42**	**44**	**46**
HX MEDICAL NONCOMPLIANCE	0.00	0.00	0.00		**47(29.0)**	**47(28.8)**	**32(29.0)**	**47(1.5)**	**37(1.0)**	**47(0.99)**	**47**	**47**	**30**	**45**
δ_1	1.00	1.00	1.00											
δ_2	1.00	1.00	1.00											
λ	0	5.03	2.18											
$\tilde{p}_\lambda$	40	17	20											

Table 6: Parameter estimates from fitted penalized Benter-Plackett-Luce (BPL) models applied to **case C** data, ordered by estimated values of θ_k from using $\lambda = \lambda_{\text{AIC}}$. The final row, $\tilde{p}_\lambda$, gives the number of non-zero parameters in the estimated model. For comparison, the remaining columns give the ranked list of problems according to the listed alternative algorithm or model; comparator entries in **bold** indicate ranks that either were (i) discrepant with λ_{AIC} by not more than 1 position or (ii) larger than the total number of non-zero estimated weights according to λ_{AIC} and which λ_{AIC} estimated to be zero. The parenthetical numbers in the AMean, GMean, and Median rank columns correspond to the summary rank of that problem, and the parenthetical numbers in the MC columns correspond to the estimated transition probability for the stationary distribution ($\times 100$).

	BPL			LDRBO	Mean	GMean	Median	MC1	MC2	MC3	CEMC$_\rho$	CEMC$_\tau$	EMM	MM
Problem / Parameter	$\lambda = 0$	λ_{AIC}	λ_{BIC}		Rank	Rank	Rank							
θ_0	3.47	3.28	3.45											
PERICARDIAL EFFUSION	7.46	7.24	7.41	**1**	**1(1.8)**	**1(1.3)**	**1(1.0)**	**1(7.7)**	**1(18.7)**	**1(16.8)**	**1**	**1**	**1**	**1**
UTI	5.38	5.17	5.34	**2**	**2(5.8)**	**2(4.1)**	**2(4.0)**	**2(4.3)**	**2(13.6)**	**2(9.1)**	**2**	**2**	**2**	**2**
ANEMIA	4.99	4.78	4.95	**3**	**3(6.5)**	**3(5.2)**	**3(5.0)**	**3(3.9)**	**3(10.3)**	**3(7.8)**	**3**	**3**	**3**	**3**
ELEVATED LFT'S	4.57	4.36	4.53	**4**	**5(8.4)**	**5(6.6)**	**4(5.0)**	7(3.5)	**4(8.1)**	**5(6.3)**	**4**	**4**	**4**	**4**
HYPERTENSION	4.54	4.33	4.50	**6**	**4(8.1)**	**4(6.2)**	**5(6.0)**	**6(3.6)**	**5(6.5)**	**4(6.5)**	**5**	**5**	**5**	**5**
R EYE BLIND	3.56	3.34	3.52	**7**	**6(12.6)**	**7(10.6)**	**6(11.0)**	10(3.3)	**6(4.1)**	**7(3.9)**	9	**6**	**7**	**6**
SYSTOLIC MURMUR	3.32	3.10	3.28		**7(13.3)**	**6(10.4)**	**7(17.5)**	12(3.3)	10(1.7)	**6(4.0)**	**6**	**7**	**6**	**7**
HISTORY OF SMOKING	3.00	2.79	2.97	5	**8(14.8)**	11(13.4)	**8(18.0)**	**9(3.4)**	12(1.7)	10(3.0)	13	**8**	**8**	10
FEVER AND NIGHT SWEATS	2.74	2.52	2.70		**9(15.2)**	**8(12.4)**	14(18.5)	16(3.2)	**8(1.7)**	**8(3.2)**	**10**	11	12	12
PLEURAL EFFUSION	2.56	2.34	2.52		**10(15.6)**	12(13.5)	13(18.5)	15(3.2)	13(1.7)	12(2.9)	**11**	**10**	**9**	**9**
SHORTNESS OF BREATH	2.52	2.30	2.48		**11(15.7)**	9(12.6)	9(18.5)	4(3.8)	14(1.7)	9(3.0)	7	9	**10**	8
CHEST PAIN	2.41	2.19	2.38		**12(15.8)**	10(13.2)	**12(18.5)**	5(3.6)	7(1.7)	**11(3.0)**	8	**12**	**11**	**11**
DIASTOLIC MURMUR	2.23	2.02	2.20		**13(16.5)**	**13(15.0)**	10(18.5)	17(3.2)	9(1.7)	**13(2.4)**	26	**14**	**13**	**13**
HISTORY OF TAH/BSO	1.69	0.00	1.66		**14(17.5)**	**14(16.8)**	11(18.5)	**21(3.0)**	11(1.7)	**14(2.0)**	**16**	**13**	**14**	**14**
AORTIC DISSECTION	0.19	0.00	0.00		**15(18.5)**	**15(17.3)**	**15(18.8)**	8(3.4)	**16(1.6)**	**15(1.7)**	**30**	**15**	**19**	**16**
MYOCARDIAL INFARCTION	0.07	0.00	0.00		**18(18.5)**	**24(18.2)**	**16(18.8)**	**24(2.8)**	**21(1.6)**	**22(1.6)**	12	**22**	**16**	**15**
CARDIOMYOPATHY	0.04	0.00	0.00		**16(18.5)**	**16(17.7)**	**17(18.8)**	13(3.2)	**15(1.6)**	**16(1.7)**	**19**	**23**	**26**	**17**
INCREASED JVP	0.04	0.00	0.00		**20(18.5)**	**22(18.1)**	**21(18.8)**	**22(3.0)**	**30(1.6)**	**23(1.6)**	**28**	**18**	**17**	**18**
EKG CHANGES, OLD MI	0.03	0.00	0.00		**24(18.6)**	**25(18.2)**	**23(18.8)**	**25(2.8)**	**20(1.6)**	**25(1.6)**	**23**	**19**	**23**	**21**
THROMBOCYTOSIS	0.03	0.00	0.00		**27(18.7)**	**28(18.5)**	**20(18.8)**	**30(2.5)**	**27(1.6)**	**29(1.6)**	**24**	**27**	**15**	**26**
CONGESTIVE HEART FAILURE	0.03	0.00	0.00		**17(18.5)**	**17(17.7)**	**24(18.8)**	**18(3.1)**	**19(1.6)**	**17(1.7)**	**18**	**21**	**21**	**19**
RENAL INSUFFICIENCY	0.02	0.00	0.00		**28(18.7)**	**29(18.5)**	**26(18.8)**	**29(2.5)**	**29(1.6)**	**28(1.6)**	**25**	**28**	**18**	**25**
CARDIOMEGALY	0.02	0.00	0.00		**19(18.5)**	**18(17.7)**	**19(18.8)**	11(3.3)	**25(1.6)**	**18(1.7)**	**15**	**16**	**28**	**20**
HYPERTENSIVE HEART DISEASE	0.01	0.00	0.00		**21(18.5)**	**19(17.7)**	**29(19.0)**	**19(3.1)**	**28(1.6)**	**20(1.6)**	**20**	**17**	**22**	**24**
VALVULAR HEART DISEASE	0.01	0.00	0.00		**22(18.5)**	**20(17.7)**	**30(19.0)**	**14(3.2)**	**26(1.6)**	**19(1.6)**	**29**	**24**	**24**	**22**
IRON DEFICIENCY	0.01	0.00	0.00		**23(18.6)**	**21(18.0)**	**22(18.8)**	**20(3.1)**	**17(1.6)**	**21(1.6)**	**17**	**20**	**29**	**23**
EKG CHANGES	0.00	0.00	0.00		**25(18.6)**	**23(18.1)**	**28(19.0)**	**23(2.9)**	**22(1.6)**	**24(1.6)**	**21**	**26**	**27**	**27**
PULMONARY EMBOLISM	0.00	0.00	0.00		**29(18.7)**	**27(18.4)**	**27(19.0)**	**26(2.7)**	**18(1.6)**	**27(1.6)**	**14**	**29**	**25**	**30**
ASCVD	0.00	0.00	0.00		**26(18.7)**	**26(18.4)**	**18(18.8)**	**27(2.7)**	**24(1.6)**	**26(1.6)**	**22**	**30**	**30**	**28**
R SIDED HEART FAILURE	0.00	0.00	0.00		**30(18.7)**	**30(18.5)**	**25(18.8)**	**28(2.5)**	**23(1.6)**	**30(1.6)**	**27**	**25**	**20**	**29**
δ_1	1.00	1.00	1.00											
δ_2	0.86	0.86	0.86											
λ	0	3.04	0.01											
$\tilde{p}_\lambda$	29	15	16											

Finally, the results for case C are given in Table 6. Thirty unique problems were listed across all lists. The largest log-odds ratio was attributed to PERICARDIAL EFFUSION (7.24, 7.41 respectively for AIC, BIC). There was a significant gap between the next ranked item, urinary tract infection (UTI), and the difference in log-odds ratios was $7.24 - 5.17 \approx 2.07$, meaning that the model-estimated odds of ranking PERICARDIAL EFFUSION over UTI at stage 1 are $\exp\{7.24 - 5.17\} \approx 8$. In total, the consensus problem list was length 13 (AIC) or 14 (BIC), compared to an LDRBO-based length of 7. There was perfect agreement with the LDRBO-based list on the first four problems, with the only discrepancy occurring on HISTORY OF SMOKING, ranked 8th in the AIC-selected list and 5th in the LDRBO-based list. There was widespread agreement with most other comparator methods. The estimate for θ_0 was 3.28, and the set of parameter estimates yielded a simulation-based estimate of the expected quartiles for list length of $(4, 6, 9)$, which are similar to the observed quartiles of $(6, 7, 9)$. From Figure 3, PERICARDIAL EFFUSION had a model-estimated 0.71 probability of selection at stage 1, with the most preferred items at subsequent stages being selected with probability between 0.2 and 0.3.

6 Data analysis: NBA team rankings

We briefly present here a secondary analysis of a dataset first reported in Deng et al. (2014). After the 2011 NBA preseason, six professional news agencies ranked all 30 teams in the league (we do not analyze the 28 student surveys reported in that paper). These six lists are complete rather than ragged, and the value of the fatigue parameter maximizing the likelihood is thus $\theta_0 = \infty$, meaning it can be dropped from the model. We applied the same set of methods to these data, the results of which are reported in Table S1 of the Supplement. There was widespread overall agreement between all methods. However, the AIC-selected list was substantially more parsimonious, which is due to the small-sample correction: it will not estimate more parameters, i.e. item weights, than there are observations, i.e. rankers, of which there are just 6 in these data.

7 Discussion

A challenging, but not unique, feature of the problem list data is that each list may have a different length, making difficult the implementation of multistage models that assume a uniform list length. Moreover, it is useful to have an aggregated consensus list that has excluded unimportant items. With these objectives in mind, we have extended classical, multistage models and amalgamated them with modern penalized likelihood ideas. As seen in our second data example, these penalized BPL models apply equally well to the analysis of non-ragged data.

We have already mentioned some advantages a modeling approach has over the approach taken by Krauss et al. (2016), which calculated a hypothetical problem list maximizing pairwise similarity with the observed problem lists. One additional, yet-unmentioned advantage

over that approach and some of the others we've considered in this paper is that the penalized BPL models do not only order the items but also give an explicit numerical assessment of their relative importance by way of an estimated relative log-odds ratio. For example, in case B, we can conclude that there are a substantial number of problems for which the physicians were conflicted about: the difference in log-odds ratios between the 6th and 15th ranked problems, SCHIZOPHRENIA and SINUSITIS, respectively, was just $2.89 - 2.03 = 0.86$, and consequently any differences in log-odds ratios between these ranks was even smaller. This may be why the penalized BPL consensus lists differed from the LDRBO-based list. In each of our problem list analyses, the existing methods that our penalized BPL models agreed with most often, i.e. MC3 and CEMC$_\tau$, were also the methods that performed well in our simulation study.

Tables 4–6 and Figure 3 may seem inconsistent: the set of non-zero items in the tables is somewhat longer than the length of the modal lists plotted in Figure 3. However, these results describe different dimensions of consensus. The tables describe overall physician agreement on the sets of relevant problems for each case abstract, whereas the figure characterizes the model-estimated probability of the list that is most likely to be constructed by an individual physician. Our results suggest that, for cases A and C, a physician should not expect to construct a list that matches that of her colleague beyond the highest ranked item; collectively, however, the physicians are in agreement on the first five or so items. In contrast, for case B, there was generally no consensus.

One important design-based challenge to our analysis is with regard to the defining, naming, and grouping of problems. As described in the introduction, physicians were free to describe problems in their own words during the interview. If the physician named any clinically similar problems that had already been listed, either by her or another physician, the interviewer verbally observed this and offered that she could change her similar-sounding problem to match the already existing one; however, she was not forced to do so. This is why case B has HISTORY OF ALCOHOL ABUSE, ALCOHOLISM, ALCOHOLIC CIRRHOSIS WITH ASCITES, and ALCOHOLIC CIRRHOSIS WITH SBP all listed as separate problems. We also implicitly assumed that the number of possible items, v, for each case was exactly the number of unique items listed by all physicians, but it is possible that, if more interviews were to be conducted, additional unique problems would be introduced to the vocabulary. One must therefore assume that our sample size was sufficiently large to include, at a minimum, those problems that would fall in the consensus list.

Acknowledgments

Supported by the National Institutes of Health (UL1TR002240)

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. *Second International Symposium on Information Theory* pages 267–281.

Benter, W. et al. (2008). Computer-based horse race handicapping and wagering systems: A report. In Hausch, D. B., Lo, V. S. Y., and Ziemba, W. T., editors, *Efficiency of racetrack betting markets*, pages 183–198. World Scientific Publishing.

Boulesteix, A.-L. and Slawski, M. (2009). Stability and aggregation of ranked gene lists. *Briefings in bioinformatics* **10,** 556–568.

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.

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.

Dicker, L., Huang, B., and Lin, X. (2013). Variable selection and estimation with the seamless-l_0 penalty. *Statistica Sinica* **23,** 929–962.

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.

Fligner, M. A. and Verducci, J. S. (1988). Multistage ranking models. *Journal of the American Statistical association* **83,** 892–901.

Gormley, I. C. and Murphy, T. B. (2008). Exploring voting blocs within the irish electorate: A mixture modeling approach. *Journal of the American Statistical Association* **103,** 1014–1027.

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

Hurvich, C. M. and Tsai, C.-L. (1989). Regression and time series model selection in small samples. *Biometrika* **76,** 297–307.

Kendall, M. G. (1948). *Rank correlation methods.* Griffin, Oxford, England.

Krauss, J. C., Boonstra, P. S., Vantsevich, A. V., and Friedman, C. P. (2016). Is the problem list in the eye of the beholder? an exploration of consistency across physicians. *Journal of the American Medical Informatics Association* **23,** 859–865.

Li, H., Xu, M., Liu, J. S., and Fan, X. (2018). *ExtMallows: An Extended Mallows Model and Its Hierarchical Version for Ranked Data Aggregation.* R package version 0.1.0.

Li, H., Xu, M., Liu, J. S., and Fan, X. (2019). An extended mallows model for ranked data aggregation. *Journal of the American Statistical Association* .

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* .

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* .

Lin, S. (2010). Space oriented rank-based data integration. *Statistical Applications in Genetics and Molecular Biology* **9,**.

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.

Luce, R. D. (1959). *Individual Choice Behavior a Theoretical Analysis*. John Wiley and Sons, New York.

Mallows, C. L. (1957). Non-null ranking models. *Biometrika* **44,** 114–130.

Marden, J. I. (1996). *Analyzing and modeling rank data*. Chapman & Hall, London.

Meyer, A. N., Payne, V. L., Meeks, D. W., Rao, R., and Singh, H. (2013). Physicians' diagnostic accuracy, confidence, and resource requests: A vignette study. *JAMA Internal Medicine* **173,** 1952–1958.

Mollica, C. and Tardella, L. (2017). Bayesian plackett–luce mixture models for partially ranked data. *Psychometrika* **82,** 442–458.

Neuwirth, E. (2014). *RColorBrewer: ColorBrewer Palettes*. R package version 1.1-2.

Nombekela, S. W., Murphy, M. R., Gonyou, H. W., and Marden, J. I. (1994). Dietary preferences in early lactation cows as affected by primary tastes and some common feed flavors. *Journal of Dairy Science* **77,** 2393–2399.

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

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Schimek, M., Budinska, E., Kugler, K., Svendova, V., Ding, J., and Lin, S. (2015). Topklists: a comprehensive r package for statistical inference, stochastic aggregation, and visualization of multiple omics ranked lists. *Statistical Applications in Genetics and Molecular Biology* pages 311–316.

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

Spearman, C. (1904). The proof and measurement of association between two things. *The American Journal of Psychology* **15,** 72–101.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* **58,** 267–288.

Webber, W., Moffat, A., and Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)* **28,** 20.

Weed, L. L. (1968). Special article: Medical records that guide and teach. *New England Journal of Medicine* **278,** 593–600.

Wickham, H. (2017). *tidyverse: Easily Install and Load the 'Tidyverse'*. R package version 1.2.1.

Supplement

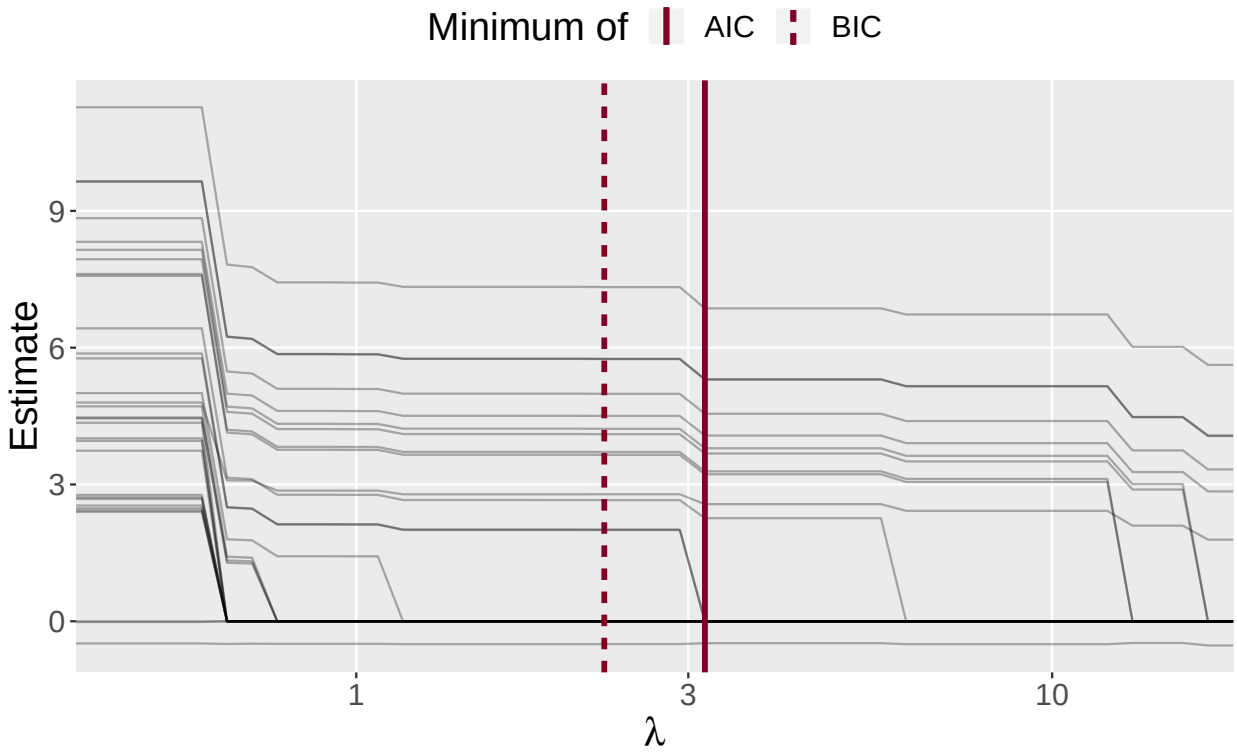

Figure S1: Solution path for **Case A**, with the choice of λ minimizing the AIC and BIC noted

Table S7: Parameter estimates from fitted penalized Benter-Plackett-Luce (BPL) models applied to the NBA data from Deng et al. (2014), ordered by estimated values of θ_k from using $\lambda = \lambda_{\text{AIC}}$. The final row, $\tilde{p}_\lambda$, gives the number of non-zero parameters in the estimated model. For comparison, the remaining columns give the ranked list of problems according to the listed alternative algorithm or model; comparator entries in **bold** indicate ranks that were either (i) discrepant with λ_{AIC} by not more than 1 position or (ii) ranks that were larger than the total number of non-zero estimated weights according to λ_{AIC} and which λ_{AIC} estimated to be zero.

	BPL			LDRBO	Mean	GMean	Median	MC1	MC2	MC3	CEMC$_\rho$	CEMC$_\tau$	EMM	MM
Team	$\lambda = 0$	λ_{AIC}	λ_{BIC}		Rank	Rank	Rank							
HEAT	24.06	16.78	24.22	**1**	**1(1.2)**	**1(1.1)**	**1(1.0)**	**1(16.13)**	**1(18.69)**	**1(17.91)**	**1**	**1**	**1**	**1**
THUNDER	21.65	14.36	21.80	**3**	**3(2.7)**	**3(2.6)**	**3(3.0)**	**3(10.92)**	**2(11.93)**	**3(11.59)**	**2**	**3**	**2**	**2**
MAVERICKS	21.31	14.03	21.46	**2**	**2(2.7)**	**2(2.4)**	**2(2.5)**	**2(12.66)**	**3(11.93)**	**2(11.78)**	**3**	**2**	**3**	**3**
BULLS	20.33	13.05	20.49	**4**	**4(3.5)**	**4(3.4)**	**4(4.0)**	**4(10.92)**	**4(8.08)**	**4(9.35)**	**4**	**4**	**4**	**4**
LAKERS	13.23	0.00	13.24	**6**	**6(6.7)**	**6(6.6)**	**6(6.5)**	**5(4.38)**	**6(5.41)**	**6(5.00)**	**6**	**6**	**6**	**6**
CLIPPERS	13.03	0.00	13.03	**5**	**5(6.7)**	**5(6.4)**	**5(6.0)**	**6(4.11)**	**5(5.93)**	**5(5.04)**	**5**	**5**	**5**	**5**
SPURS	12.33	0.00	12.34	**7**	**7(8.0)**	**7(7.7)**	**7(7.5)**	**7(4.11)**	**8(3.89)**	**7(4.09)**	**7**	**8**	**8**	**7**
CELTICS	12.20	0.00	12.21	**8**	**8(8.2)**	**8(7.8)**	**9(9.0)**	**8(4.11)**	**9(3.77)**	**8(4.01)**	**9**	**10**	**10**	**9**
KNICKS	11.87	0.00	11.88	**9**	**9(8.7)**	**9(8.4)**	**8(8.5)**	**10(3.66)**	**7(3.89)**	**9(3.72)**	**8**	**7**	**7**	**8**
GRIZZLIES	11.83	0.00	11.84	**10**	**10(9.0)**	**10(8.9)**	**10(9.0)**	**9(3.84)**	**10(3.18)**	**10(3.48)**	**10**	**9**	**9**	**10**
MAGIC	9.90	0.00	9.90	**17**	**12(12.7)**	**12(12.5)**	**12(11.5)**	**13(2.08)**	**12(2.21)**	**12(2.14)**	**11**	**13**	**12**	**12**
PACERS	9.88	0.00	9.89	**14**	**13(13.3)**	**13(13.3)**	**13(13.5)**	**12(2.18)**	**13(1.87)**	**13(1.97)**	**12**	**12**	**13**	**13**
NUGGETS	9.74	0.00	9.75	**11**	**11(10.8)**	**11(10.0)**	**11(9.5)**	**11(2.76)**	**11(2.71)**	**11(2.87)**	**26**	**11**	**11**	**11**
TRAILBLAZERS	9.53	0.00	9.53	**13**	**14(14.0)**	**14(13.9)**	**14(14.0)**	**14(2.01)**	**14(1.68)**	**14(1.83)**	**13**	**14**	**14**	**14**
76ERS	9.12	0.00	9.13	**15**	**15(14.8)**	**16(14.8)**	**15(15.0)**	**15(1.92)**	**16(1.52)**	**15(1.67)**	**14**	**15**	**15**	**15**
HAWKS	8.19	0.00	8.19	**12**	**16(15.0)**	**15(14.7)**	**16(15.0)**	**16(1.84)**	**15(1.55)**	**16(1.67)**	**15**	**18**	**18**	**16**
ROCKETS	8.11	0.00	8.11	**16**	**17(17.0)**	**17(16.9)**	**17(16.5)**	**18(1.36)**	**17(1.39)**	**17(1.35)**	**17**	**16**	**16**	**17**
BUCKS	7.49	0.00	7.50	**18**	**18(17.7)**	**18(17.6)**	**18(17.0)**	**17(1.37)**	**18(1.27)**	**18(1.28)**	**16**	**17**	**17**	**18**
SUNS	6.25	0.00	6.25	**19**	**19(20.3)**	**19(20.3)**	**19(20.5)**	**19(1.02)**	**20(1.02)**	**19(1.01)**	**20**	**20**	**20**	**19**
WARRIORS	6.01	0.00	6.02	**22**	**20(21.2)**	**21(21.1)**	**20(21.5)**	**22(0.90)**	**19(1.06)**	**21(0.94)**	**19**	**22**	**19**	**20**
NETS	5.60	0.00	5.60	**21**	**21(21.2)**	**20(21.0)**	**21(21.5)**	**20(1.00)**	**21(0.98)**	**20(0.96)**	**18**	**19**	**21**	**21**
TIMBERWOLVES	5.40	0.00	5.40	**23**	**22(22.3)**	**22(22.3)**	**22(22.5)**	**23(0.87)**	**22(0.87)**	**22(0.86)**	**21**	**21**	**22**	**22**
PISTONS	4.37	0.00	4.37	**25**	**25(24.3)**	**25(24.3)**	**26(25.0)**	**25(0.78)**	**24(0.73)**	**25(0.75)**	**22**	**24**	**23**	**25**
HORNETS	3.96	0.00	3.96	**20**	**23(23.2)**	**23(22.8)**	**23(24.0)**	**24(0.83)**	**23(0.75)**	**23(0.83)**	**25**	**25**	**25**	**23**
JAZZ	3.75	0.00	3.75	**26**	**24(23.5)**	**24(23.2)**	**24(24.5)**	**21(0.94)**	**25(0.73)**	**24(0.81)**	**24**	**23**	**24**	**24**
KINGS	3.70	0.00	3.70	**24**	**26(24.5)**	**26(24.3)**	**25(25.0)**	**26(0.77)**	**26(0.71)**	**26(0.74)**	**23**	**26**	**26**	**26**
WIZARDS	3.04	0.00	3.04	**28**	**27(27.0)**	**27(27.0)**	**27(27.0)**	**27(0.66)**	**27(0.60)**	**27(0.62)**	**27**	**27**	**27**	**27**
CAVALIERS	2.19	0.00	2.19	**27**	**29(27.7)**	**29(27.6)**	**29(28.0)**	**28(0.66)**	**29(0.56)**	**29(0.60)**	**29**	**29**	**29**	**29**
RAPTORS	2.10	0.00	2.10	**29**	**28(27.5)**	**28(27.4)**	**28(28.0)**	**29(0.64)**	**28(0.58)**	**28(0.60)**	**28**	**28**	**28**	**28**
BOBCATS	0.00	0.00	0.00	**30**	**30(29.8)**	**30(29.8)**	**30(30.0)**	**30(0.53)**	**30(0.51)**	**30(0.52)**	**30**	**30**	**30**	**30**
δ_1	1.00	1.00	1.00											
δ_2	1.00	1.00	1.00											
λ	0	29.009	0.000											
$\tilde{p}_\lambda$	29	4	29											

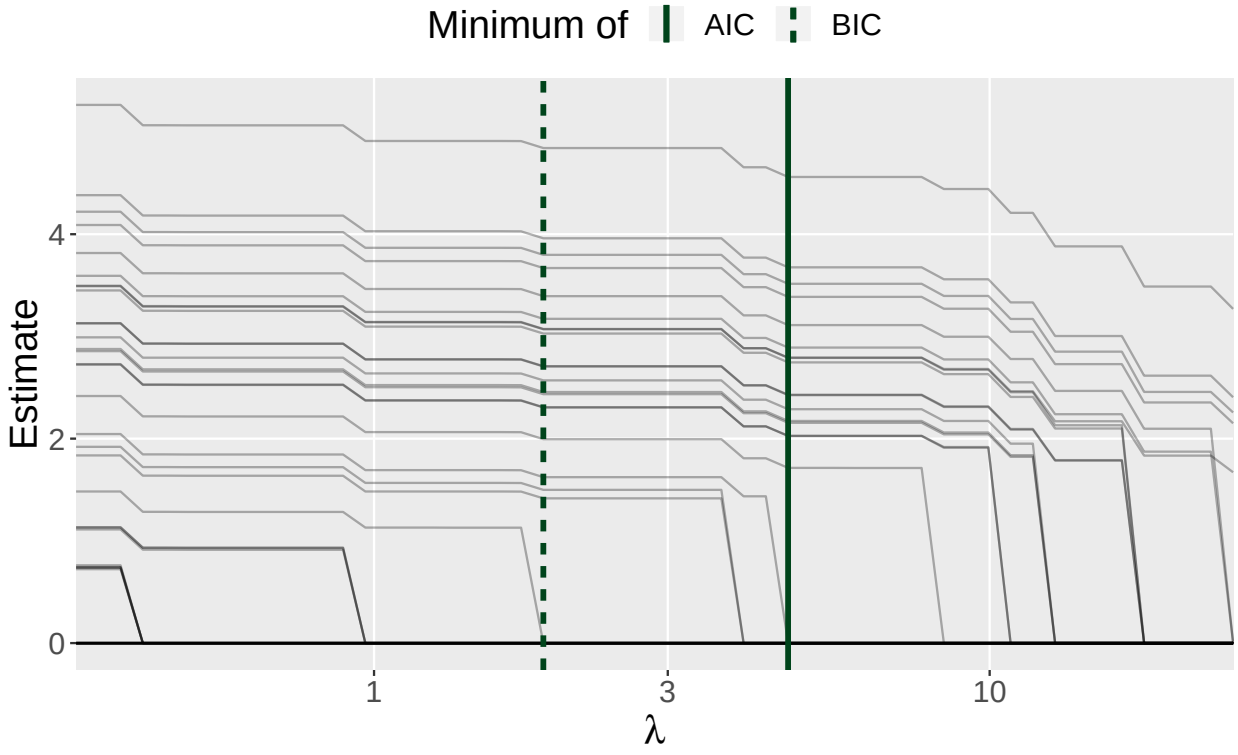

Figure S2: Solution path for **Case B**, with the choice of λ minimizing the AIC and BIC noted

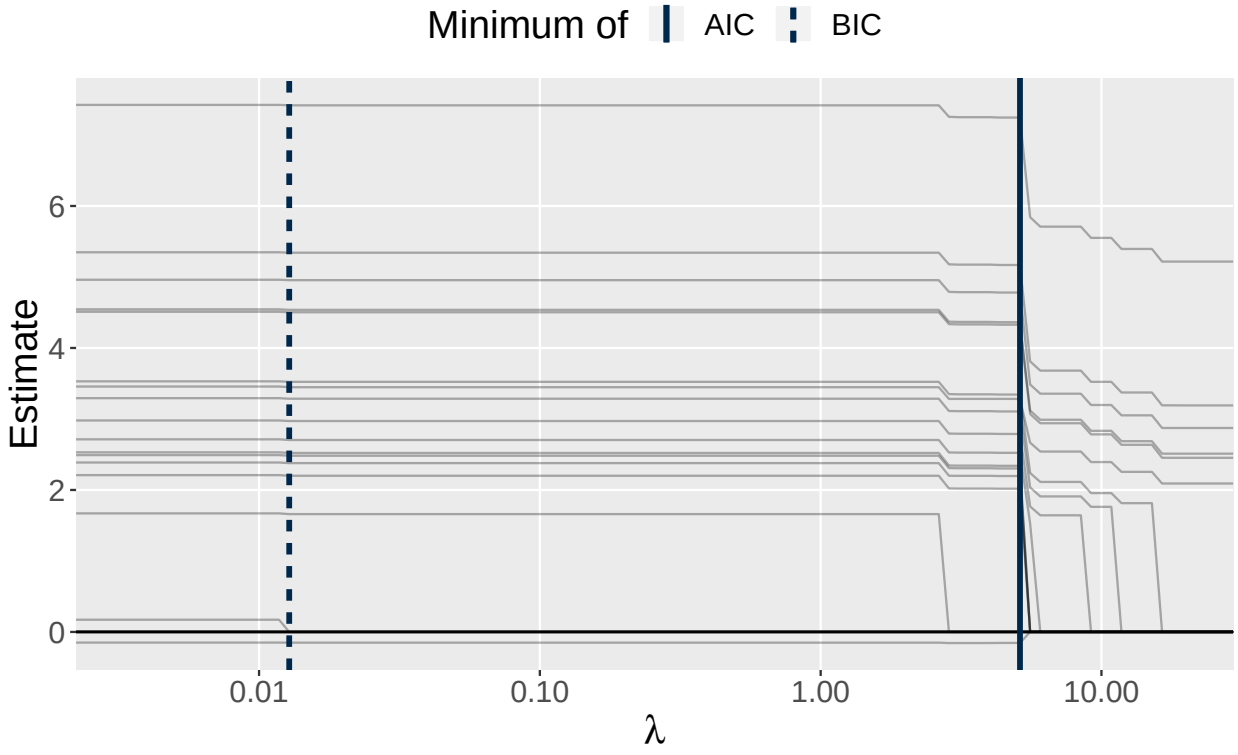

Figure S3: Solution path for **Case C**, with the choice of λ minimizing the AIC and BIC noted