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

March 14, 2019

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, among other things. 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. 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 deteriorates; in contrast, physicians broadly disagreed on the relevant problems for the middle and most difficult case abstract.

Keywords: conditional multinomial; L_0 penalty; variable selection; Akaike Information Criterion

*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 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) of 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. Or, 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 38 agreed and provided the protocol-approved informed consent. 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 data 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 (Li et al., 2017; Boulesteix and Slawski, 2009; Li et al., 2018), identifying bovine feeding preferences (Nombekela et al., 1994), handicapping horse races (Plackett, 1975; Benter et al., 2008), or

indexing search engine results (Webber et al., 2010). However, once the data are in hand, the subsequent analysis usually 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, are a common technique for aggregating and modeling a set of ordered lists such as what we seek to analyze here. However, multiple idiosyncracies, both with respect to the underlying nature of the data and with respect to our scientific objectives, require the development of novel extensions to this multistage, model-based approach. There are three distinctive such extensions that we describe in this manuscript. First, we equip multistage models to handle so-called ragged lists, which have different lengths. The length of each list becomes informative when the ranker chooses to stop ranking items, as in our case study, and we thus model the fatigue process of rankers. Second, we equip the likelihood with a modified L_0-type variable selection penalty so as to induce sparsity among the maximum penalized likelihood estimates. 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 rank data with modern penalized regression techniques warranted by the motivating context. Third, we provide a computational framework and code 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 describes in detail each of these extensions and illustrates in their application to the problem list data how they can offer a nuanced view of the data and the underlying homogeneity.

2 Technical Background

We assume that each ranker is ordering items from a list, 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 ranging 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 assumed to be formulated as ordered lists, but we will continue to refer to items that are ordered first as being 'highly ranked'.

The orderings may also 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/) 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 viewed as being more important than at lower ranks, Webber et al. (2010) call them 'indefinite'. Ignoring the top-weighted characteristic for now, we will more generally refer to these lists as partially ordered lists or simply partial lists. Importantly, partial lists could theoretically be made longer but have been artificially truncated, outside of the purview of the ranker and external to the ordering process. Distinct from these are what we call *ragged* lists, in which each list is incomplete due to each individual ranker's choice. Ragged lists do not exist beyond the last item ranked because the ranker has deliberately chosen to stop ordering the remaining, un-ranked items. One can, however, infer that the ranker places the un-ranked items at a lower rank than the ranked items.

The problem list data we study here are of the ragged type, since each physician was free to select and include as many or as as few problems as desired. 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, but eight problems were only ranked by a single 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 additional 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. Two broad approaches for analyzing ordered lists have been developed, which we now describe.

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. The rank-biased overlap (RBO, Webber et al., 2010) is a more recent example specifically designed for ordered lists. Given a tuning parameter $\psi \in (0, 1)$, the RBO between two lists $\boldsymbol{x}_{i_1}$ and $\boldsymbol{x}_{i_2}$ is

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

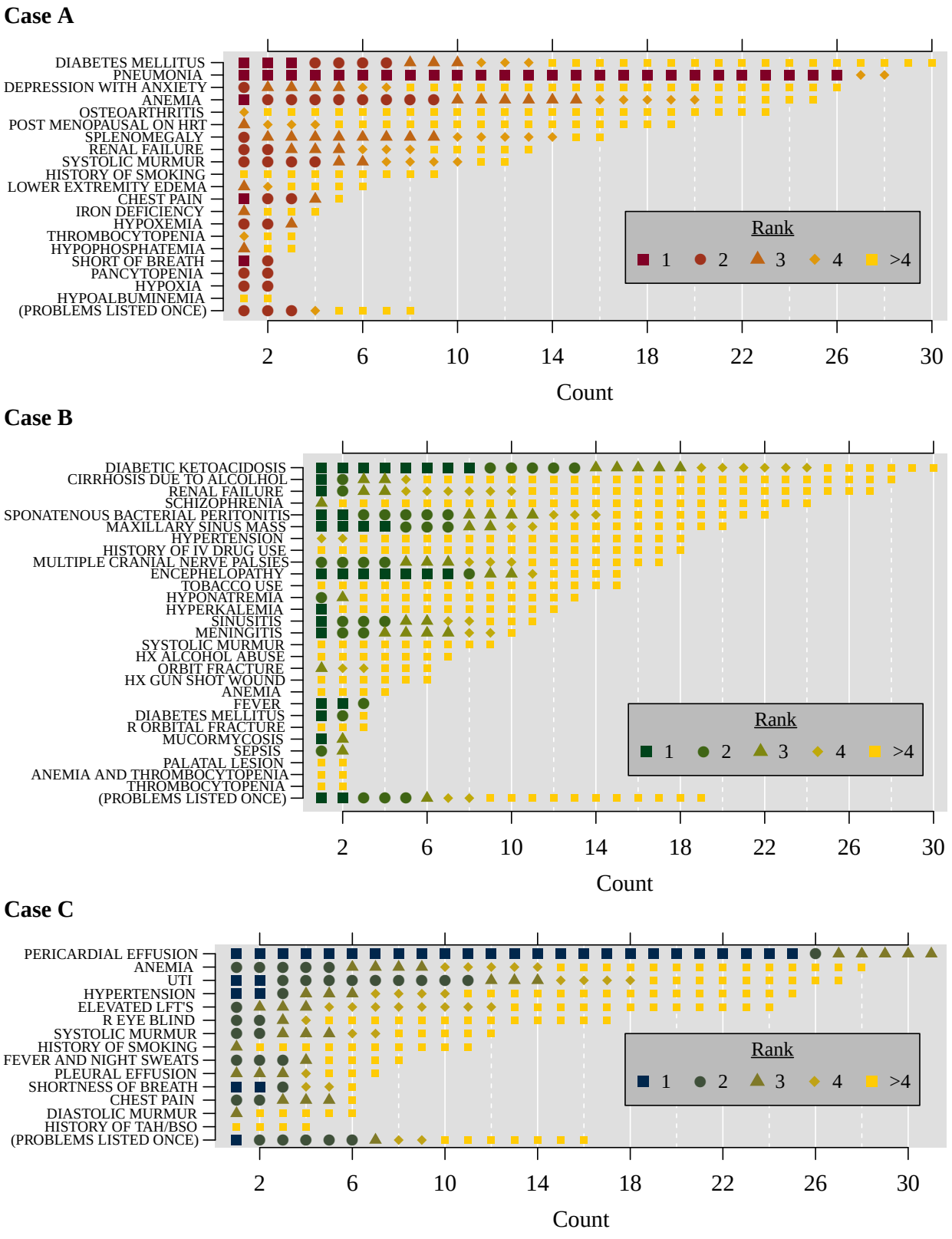

Figure 1: Counts of the frequency that each problem was listed on any of the $n = 32$ generated lists for each of three case abstracts, with shading and shape used to indicate the rank of that problem. For brevity, only those problems listed by at least two physicians are shown.

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 set of weights ψ^d. 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 in 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 an intuition that the pair $\{\boldsymbol{x}_{i_1}, \boldsymbol{x}_{i_2}\}$ have 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 the patient whose data are depicted in Figure 1, case A, 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 approach could 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.

2.2 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. The traditional multistage model for ordered lists, attributed to both Plackett (1975) and Luce (1959), describes the ith ranker's stage-wise process of generating 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 (WLOG), we may assume that $\min_j\{\theta_j\} \equiv 0$. See Section 5.6, Marden (1996) for an overview of classical multistage models.

An extension by Benter introduces a dampening effect 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, there is less distinction between items, and so, assuming that the strongest preferences are always in the first stages, it is reasonable to constrain $\delta(s)$ to be non-increasing with s. At the final stage, $\delta(v)$ may take any value WLOG, since there is no choice remaining. When $\delta(\cdot)$ is limited to the set of non-increasing functions, this dampening function serves an analogous purpose as the ψ parameter in the (LD)RBO measures of Webber et al. (2010) and Krauss et al. (2016), namely to reflect that agreement at higher ranks is relatively more important than at lower ranks.

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 $\Pr(x_{is} = k|\mathcal{O}_{is}) = 1_{[k\in\mathcal{O}_{is}]} \exp(\theta_k\delta(s)) / \sum_{j\in\mathcal{O}_{is}} \exp(\theta_j\delta(s))$, $s, k = 1, \ldots, v$. 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 of them having been yet ordered, are $\theta_{k_1} - \theta_{k_2}$ (PL) or $\delta(s)[\theta_{k_1} - \theta_{k_2}]$ (BPL). We now propose two novel extensions based upon the BPL model – one to the model itself and one to its estimation – to satisfy the objectives of the problem list analysis.

3 Modeling 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 have the model handle lists with lengths less than the number of items ($\ell_i < v$) and 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 have ranked 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 $\min_j\{\theta_j\} \equiv 0$ 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:

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

where $\mathcal{O}_{is}$ is as defined in Equation (1). This is the model we will use in the analysis of the problem list data. We discuss the choice of dampening function $\delta(\cdot)$ in Section 3.1 and consider strategies for inducing sparsity in the fitted model in Section 3.2

3.1 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 $\delta(\cdot)$ other than that $0 \leq \delta(s) \leq 1$ 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 ($\delta_1 < 1$).

3.2 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 the smallest θ_k equals zero, 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 exactly equal to zero, such that there is a set of $\hat{\theta}_k$'s that are simultaneously equal to zero (which is the smallest possible parameter value). 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 equal to one another. 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 $\{\hat{\theta}_1, \hat{\theta}_2, \hat{\theta}_3\} = \{0, \log(1.1), \log(2.9)\}$. The estimated probability of selecting item 3 at stage 1 is $2.9/(1 + 1.1 + 2.9) = 0.58$. If $\hat{\theta}_2$ is to be set to zero to reflect that items 1 and 2 seem to be equally least important, then the corresponding estimate of $\hat{\theta}_3$ must also be changed to approximately $\log(2.76)$ in order to maintain this probability: $2.76/(1 + 1 + 2.76) \approx 0.58$. That is, in order to change $\hat{\theta}_2$

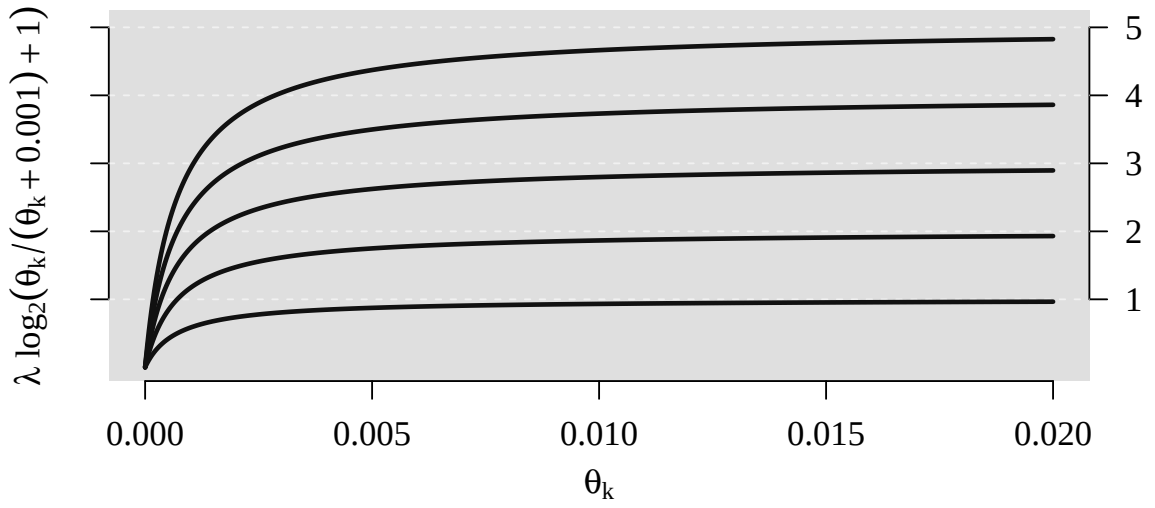

Figure 2: Seamless L_0 penalty under the default choice of $\tau = 0.001$ and different values of the penalty parameter (and asymptote) λ

from log(1.1) to 0 while maintaining the relative importance of $\hat{\theta}_3$, 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 capturing 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 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 or tuning 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.

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) 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 solutions that are only locally optimal. These are the main challenges our algorithm must overcome.

As is common 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 $\boldsymbol{\beta}$ across λ.

In more detail, at each iteration we propose and accept changes to the current parameter estimates that increase the penalized likelihood. 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 the proposal sequence $\Gamma = \{\gamma_{-t}, \gamma_{-t+1}, \ldots, \gamma_{-1}, \gamma_0 \equiv 0, \gamma_1, \ldots, \gamma_{t-1}, \gamma_t\}$, where $\gamma_{-t} = -\gamma_t$ for all $t > 0$. The inclusion of $\gamma_0 \equiv 0$ means that one proposal is to not change any values. Furthermore, 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 results in up to $2t - 1$ penalized log-likelihood calculations, and we identify $t_{\max} \in \{-t, -t+1, \ldots, 1, 0, 1, \ldots, t-1, t\}$, which is defined as the index of Γ yielding the largest penalized log-likelihood. 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 starts over 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 considering also multivariate proposals to speed convergence and discourage the algorithm from getting stuck in local optima. We incorporated several such proposals in our implementation. One proposal shifts all non-zero $\hat{\theta}_k$'s *towards* (but never less than) zero by an amount equal to a randomly selected positive element of Γ, and another shifts all non-zero $\hat{\theta}_k$'s and one randomly selected zero-valued $\hat{\theta}_k$ (if there is more than one such zero-valued $\hat{\theta}_k$) *away* from zero by a value equal to a randomly selected positive element of Γ. A third proposal considers the current estimated item weights $\hat{\Theta}^{(m)}$ in increasing order and, with a certain probability, exchanges the index of any neighboring parameter estimates. For example, if $\{\hat{\theta}_1^{(m)}, \hat{\theta}_2^{(m)}, \hat{\theta}_3^{(m)}\} = \{0, \log(1.1), \log(2.9)\}$, the proposal might swap $\hat{\theta}_1^{(m)}$ and $\hat{\theta}_2^{(m)}$, resulting in a proposal of $\{\log(1.1), 0, \log(2.9)\}$. 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 complicated formulation, since $\tilde{\delta}_2 = 0$. The penalized log-likelihood under each of these four proposals is calculated, and if any exceed the current penalized log-likelihood value, the corresponding proposal is accepted.

4.1 Default values

Our algorithm for approximating the maximized penalized log-likelihood requires choosing several 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 implementation in `R` provided on github, we give default choices that we used in our analyses, so that a user need only provide the data, comprising a set of ordered lists.

The default choice of convergence criterion is $\epsilon_{\text{def}} = 0.001$. In addition to indicating convergence, this also means that any $\hat{\theta}_k < 0.001$ is treated as equal to zero and that the number of significant digits retained by the algorithm is equal to $\lceil \log_{10}(1/\epsilon_{\text{def}}) \rceil = 3$ by default. 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 five sets of initial values in our analyses.

4.2 Model Selection

For each case study, we calculated an unpenalized BPL model as well as two penalized BPL model fits that select $\lambda > 0$ minimizing one of two information criteria. 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 the fatigue parameter 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 and $h(\tilde{p}_\lambda)=\log(n)\tilde{p}_\lambda/2$ for the BIC.

5 Consensus Problem List

Tables 1–3 give the parameter estimates for all models for cases A–C, respectively. The BIC-based results are given for comparison, and we focus primarily on the AIC-based fitted models. Figures 4–6 in the Supplement give the full solution paths from our algorithm, with the AIC and BIC solutions noted. For comparison, Tables 1–3 also include the consensus problem list from Krauss et al. (2016) and the sample mean and median ranking of each problem, which ignores the frequency that it appeared in a list. 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, only 10 were estimated to have non-trivial 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 decrease but 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 ANEMIA, say, at this stage (supposing it has not yet been ranked) decreases to $5.30\times 0.63=3.3$. The BPL models agree with the length-8 consensus problem list reported in Krauss et al. (2016) on what the three most important problems are. However, this agreement did not occur at the lower ranks, and, in fact, the BPL models did not even 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 easily translate into an expected list length, since that is a multidimensional function of all elements of $\boldsymbol{\beta}$. Thus, we directly simulated many lists from the fitted model. 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. Note also that the probabilities may slightly *increase* with stage due to the effect of the dampening function.

Case B, given in Table 2, was the most challenging, which was 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), 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

Table 1: Parameter estimates from fitted penalized BPL models applied to **case A** data, ordered by estimated values of θ_k from using λ as selected by AIC. For comparison, the consensus problem list ranks reported by Krauss et al. (2016) are given, as are the observed mean and median values of the ranks of each problem. The final row, $\tilde{p}_\lambda$, gives the number of non-zero parameters in the estimated model

	BPL				Mean	Median
Problem / Parameter	$\lambda = 0$	$\lambda = \lambda_{\text{AIC}}$	$\lambda = \lambda_{\text{BIC}}$	Krauss et al.	Rank	Rank
PNEUMONIA	9.44	6.86	7.33	1	1.2	1.0
DIABETES MELLITUS	7.84	5.31	5.76	2	4.8	5.0
ANEMIA	7.84	5.30	5.75	3	3.6	3.0
DEPRESSION WITH ANXIETY	7.05	4.55	4.99	5	6.0	6.0
OSTEOARTHRITIS	6.55	4.07	4.50	6	7.3	7.0
SPLENOMEGALY	6.32	3.79	4.22	4	3.8	3.0
POST MENOPAUSAL ON HRT	6.16	3.68	4.10		7.1	7.0
RENAL FAILURE	5.80	3.29	3.71	7	4.2	4.0
SYSTOLIC MURMUR	5.75	3.22	3.65		3.3	3.5
θ_0	3.90	2.57	2.79			
HISTORY OF SMOKING	4.67	2.26	2.66		7.4	8.0
CHEST PAIN	4.08	0.00	2.01		2.6	2.0
LOWER EXTREMITY EDEMA	4.01	0.00	2.00		6.2	5.0
IRON DEFICIENCY	3.28	0.00	0.00		6.5	7.0
HYPOXEMIA	2.95	0.00	0.00		2.3	2.0
HYPOPHOSPHATEMIA	2.77	0.00	0.00		7.7	10.0
THROMBOCYTOPENIA	2.74	0.00	0.00		6.7	7.0
SHORT OF BREATH	2.59	0.00	0.00		1.5	1.5
HYPOXIA	2.27	0.00	0.00	8	2.0	2.0
PANCYTOPENIA	2.21	0.00	0.00		2.0	2.0
HYPOALBUMINEMIA	2.05	0.00	0.00		6.0	6.0
FEVER	1.06	0.00	0.00		2.0	2.0
CONGESTIVE HEART FAILURE	1.01	0.00	0.00		2.0	2.0
PULMONARY EDEMA	0.98	0.00	0.00		2.0	2.0
PULMONARY EMBOLISM	0.86	0.00	0.00		4.0	4.0
TACHYCARDIA	0.83	0.00	0.00		6.0	6.0
DEPRESSION	0.77	0.00	0.00		5.0	5.0
HIGH HAPTOGLOBIN	0.77	0.00	0.00		8.0	8.0
ANXIETY	0.00	0.00	0.00		6.0	6.0
δ_1	0.99	1.00	1.00			
δ_2	0.61	0.62	0.61			
λ	0	3.17	2.27			
$\tilde{p}_\lambda$	30	12	14			

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 constructed by Krauss et al. (2016). 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 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 $8/32 = 0.25$.

Finally, the results for case C are given in Table 3. Thirty unique problems were listed

Table 2: Parameter estimates from fitted penalized BPL models applied to **case B** data, ordered by estimated values of θ_k from using λ as selected by AIC. For comparison, the consensus problem list ranks reported by Krauss et al. (2016) are given, as are the observed mean and median values of the ranks of each problem. The final row, $\tilde{p}_\lambda$, gives the number of non-zero parameters in the estimated model

	BPL				Mean	Median
Problem / Parameter	$\lambda = 0$	$\lambda = \lambda_{\text{AIC}}$	$\lambda = \lambda_{\text{BIC}}$	Krauss et al.	Rank	Rank
DIABETIC KETOACIDOSIS	5.27	4.56	4.84	2	3.2	3.0
RENAL FAILURE	4.39	3.68	3.96	5	5.4	5.0
SPONATENOUS BACTERIAL PERITONITIS	4.23	3.52	3.80	4	4.0	3.5
CIRRHOSIS DUE TO ALCOLHOL	4.10	3.39	3.67	8	7.5	7.0
MAXILLARY SINUS MASS	3.83	3.11	3.39	1	3.8	4.0
SCHIZOPHRENIA	3.60	2.89	3.17	11	9.0	8.5
ENCEPHELOPATHY	3.50	2.79	3.07	3	3.3	2.0
θ_0	3.50	2.79	3.07			
MULTIPLE CRANIAL NERVE PALSIES	3.46	2.75	3.03	6	4.4	4.0
HYPERTENSION	3.14	2.43	2.71	12	10.2	9.0
HISTORY OF IV DRUG USE	3.14	2.42	2.71	14	10.7	10.0
HYPONATREMIA	3.00	2.29	2.57	7	6.8	7.0
HYPERKALEMIA	2.89	2.17	2.46	9	7.3	7.0
TOBACCO USE	2.87	2.15	2.44		11.1	11.0
MENINGITIS	2.74	2.03	2.31		3.1	3.0
SINUSITIS	2.73	2.03	2.31	10	3.6	3.0
SYSTOLIC MURMUR	2.42	1.71	2.00		9.8	10.0
HX ALCOHOL ABUSE	2.05	0.00	1.62		11.3	12.0
ORBIT FRACTURE	1.93	0.00	1.50	13	5.8	5.5
HX GUN SHOT WOUND	1.84	0.00	1.42		12.7	13.5
ANEMIA	1.49	0.00	0.00		9.2	8.5
FEVER	1.14	0.00	0.00		1.3	1.0
DIABETES MELLITUS	1.14	0.00	0.00		2.7	2.0
R ORBITAL FRACTURE	1.12	0.00	0.00		9.7	9.0
THROMBOCYTOPENIA	0.77	0.00	0.00		12.0	12.0
MUCORMYCOSIS	0.75	0.00	0.00		2.0	2.0
SEPSIS	0.75	0.00	0.00		2.5	2.5
PALATAL LESION	0.75	0.00	0.00		6.0	6.0
ANEMIA AND THROMBOCYTOPENIA	0.73	0.00	0.00		9.0	9.0
TACHYCARDIA	0.05	0.00	0.00		12.0	12.0
HYPERSOSMOLAR COMA	0.04	0.00	0.00		2.0	2.0
PLASMA PROTEIN DISORDER	0.03	0.00	0.00		8.0	8.0
POTENTIAL CVA	0.02	0.00	0.00		1.0	1.0
DEHYDRATION	0.01	0.00	0.00	15	4.0	4.0
HEPTO-RENAL SYNDROME	0.01	0.00	0.00		3.0	3.0
LEUKOCYTOSIS	0.01	0.00	0.00		4.0	4.0
CEREBROVASCULAR ACCIDENT	0.01	0.00	0.00		7.0	7.0
ALCOHOLISM	0.01	0.00	0.00		8.0	8.0
ALCOHOLIC CIRRHOSIS WITH ASCITES	0.00	0.00	0.00		1.0	1.0
MALNUTRITION	0.00	0.00	0.00		11.0	11.0
HEPATOMEGALY	0.00	0.00	0.00		12.0	12.0
HX MEDICAL NONCOMPLIANCE	0.00	0.00	0.00		16.0	16.0
ALCOHOLIC CIRRHOSIS WITH SBP	0.00	0.00	0.00		2.0	2.0
ASCITES	0.00	0.00	0.00		2.0	2.0
RENAL FAILURE WITH HYPERKALEMIA	0.00	0.00	0.00		5.0	5.0
SMOKING	0.00	0.00	0.00		6.0	6.0
POLYSUBSTANCE ABUSE	0.00	0.00	0.00		8.0	8.0
PROTEIN CALORIE MALNUTRITION	0.00	0.00	0.00		8.0	8.0
δ_1	1.00	1.00	1.00			
δ_2	1.00	1.00	1.00			
λ	0	4.71	1.88			
$\tilde{p}_\lambda$	40	17	20			

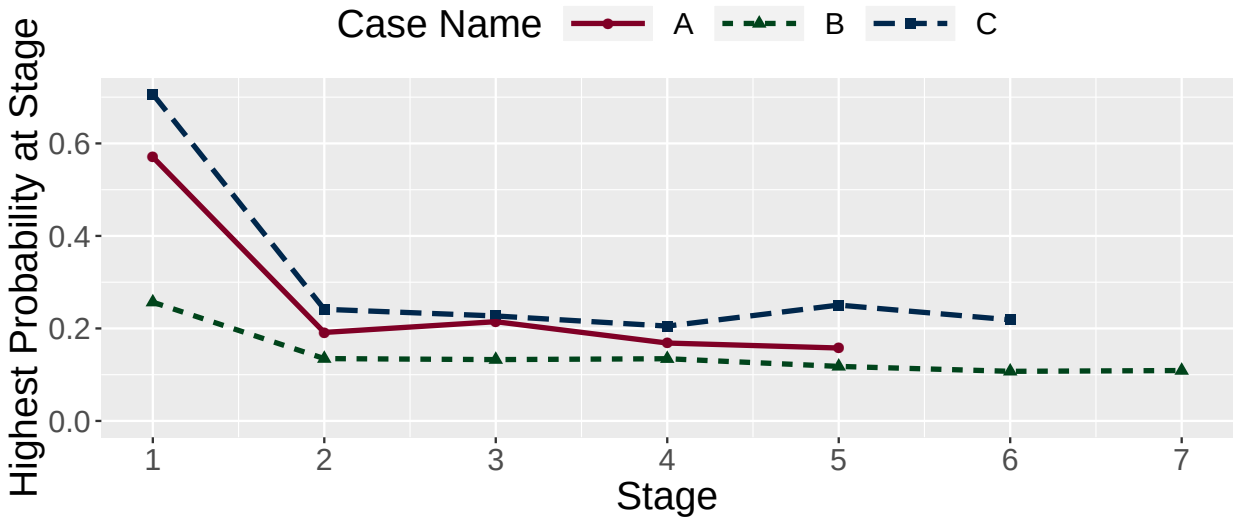

Figure 3: The probability of the most preferred item at each stage according to the AIC-estimated BPL model fit in Tables 1–3, conditional on all prior stages having also selected the most preferred item. Each such modal list continues until the item "0" is selected.

across all lists. The largest log-odds ratio was attributed to PERICARDIAL EFFUSION (7.25, 7.42 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 $7.25 - 5.17 \approx 2.08$, meaning that the model-estimated odds of ranking PERICARDIAL EFFUSION over UTI at stage 1 are approximately $\exp\{7.25 - 5.17\} \approx 8$. 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 on the ranking of 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. 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 Discussion

A challenging – but not unique – feature of the problem list data is that each list is of a potentially different length, rendering all but useless most correlation and similarity measures and making difficult the implementation of multistage models that assume a uniform list length. Thus, in order to best distill and aggregate this problem list data, we have extended classical, multistage models and amalgamated them with modern penalized likelihood ideas.

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 pair-

Table 3: Parameter estimates from fitted penalized BPL models applied to **case C** data, ordered by estimated values of θ_k from using λ as selected by AIC. For comparison, the consensus problem list ranks reported by Krauss et al. (2016) are given, as are the observed mean and median values of the ranks of each problem. The final row, $\tilde{p}_\lambda$, gives the number of non-zero parameters in the estimated model

	BPL				Mean	Median
Problem / Parameter	$\lambda = 0$	$\lambda = \lambda_{\text{AIC}}$	$\lambda = \lambda_{\text{BIC}}$	Krauss et al.	Rank	Rank
PERICARDIAL EFFUSION	7.42	7.25	7.42	1	1.4	1.0
UTI	5.34	5.17	5.34	2	3.6	3.0
ANEMIA	4.96	4.78	4.96	3	4.9	4.5
ELEVATED LFT'S	4.54	4.36	4.54	4	5.1	4.5
HYPERTENSION	4.51	4.33	4.50	6	5.3	6.0
R EYE BLIND	3.53	3.34	3.52	7	7.3	8.0
θ_0	3.47	3.28	3.45			
SYSTOLIC MURMUR	3.29	3.11	3.29		4.2	4.0
HISTORY OF SMOKING	2.98	2.79	2.97	5	7.6	7.0
FEVER AND NIGHT SWEATS	2.72	2.52	2.70		4.1	4.0
PLEURAL EFFUSION	2.54	2.34	2.52		4.3	4.0
SHORTNESS OF BREATH	2.50	2.30	2.48		2.8	3.0
CHEST PAIN	2.39	2.20	2.38		3.2	3.0
DIASTOLIC MURMUR	2.22	2.02	2.20		5.8	5.5
HISTORY OF TAH/BSO	1.68	0.00	1.66		7.8	7.5
AORTIC DISSECTION	0.18	0.00	0.00		1.0	1.0
MYOCARDIAL INFARCTION	0.07	0.00	0.00		5.0	5.0
CARDIOMYOPATHY	0.04	0.00	0.00		2.0	2.0
INCREASED JVP	0.04	0.00	0.00		4.0	4.0
EKG CHANGES, OLD MI	0.03	0.00	0.00		5.0	5.0
THROMBOCYTOSIS	0.03	0.00	0.00		8.0	8.0
CONGESTIVE HEART FAILURE	0.02	0.00	0.00		2.0	2.0
RENAL INSUFFICIENCY	0.02	0.00	0.00		8.0	8.0
HYPERTENSIVE HEART DISEASE	0.01	0.00	0.00		2.0	2.0
VALVULAR HEART DISEASE	0.01	0.00	0.00		2.0	2.0
CARDIOMEGALY	0.01	0.00	0.00		2.0	2.0
IRON DEFICIENCY	0.01	0.00	0.00		3.0	3.0
EKG CHANGES	0.00	0.00	0.00		4.0	4.0
PULMONARY EMBOLISM	0.00	0.00	0.00		6.0	6.0
ASCVD	0.00	0.00	0.00		6.0	6.0
R SIDED HEART FAILURE	0.00	0.00	0.00		8.0	8.0
δ_1	1.00	1.00	1.00			
δ_2	0.86	0.86	0.86			
λ	0	5.13	0.01			
$\tilde{p}_\lambda$	29	15	16			

wise similarity with the observed problem lists. One additional, yet-unmentioned advantage is that the penalized BPL models do not simply order the problems (items) but also give an explicit numerical assessment of their relative importance by way of an estimated relative log-odds ratio. Thus, in case B, we can conclude that there are a substantial number of problems for which the physicians were relatively conflicted about: the difference in log-odds ratios between the 6th and 15th ranked problems, SCHIZOPHRENIA and SINUSITIS, respectively, was just $2.89 - 2.03 = 0.86$, meaning that any differences in log-odds ratios between these ranks was even smaller. This is likely why the AIC-selected consensus list differed substantially from the LDRBO-based list.

The results in Tables 1–3 may seem to be in conflict with those in Figure 3, with the set of non-zero items in the tables being somewhat longer than the length of the modal lists plotted in Figure 3. Both 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 speaking, 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 (JCK) 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 likely one reason 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 likely 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.

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

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, 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* .

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

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.

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.

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

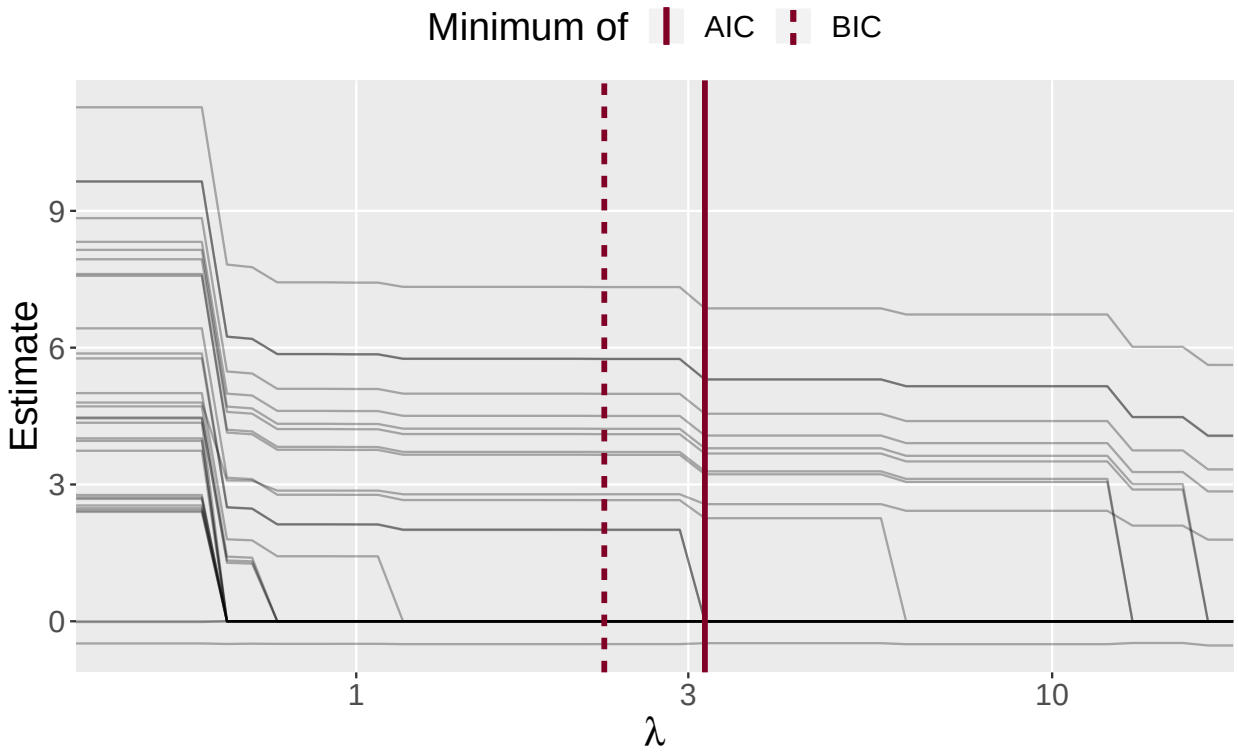

Figure 4: Solution path for **Case A**, with the choice of λ minimizing the AIC and BIC noted

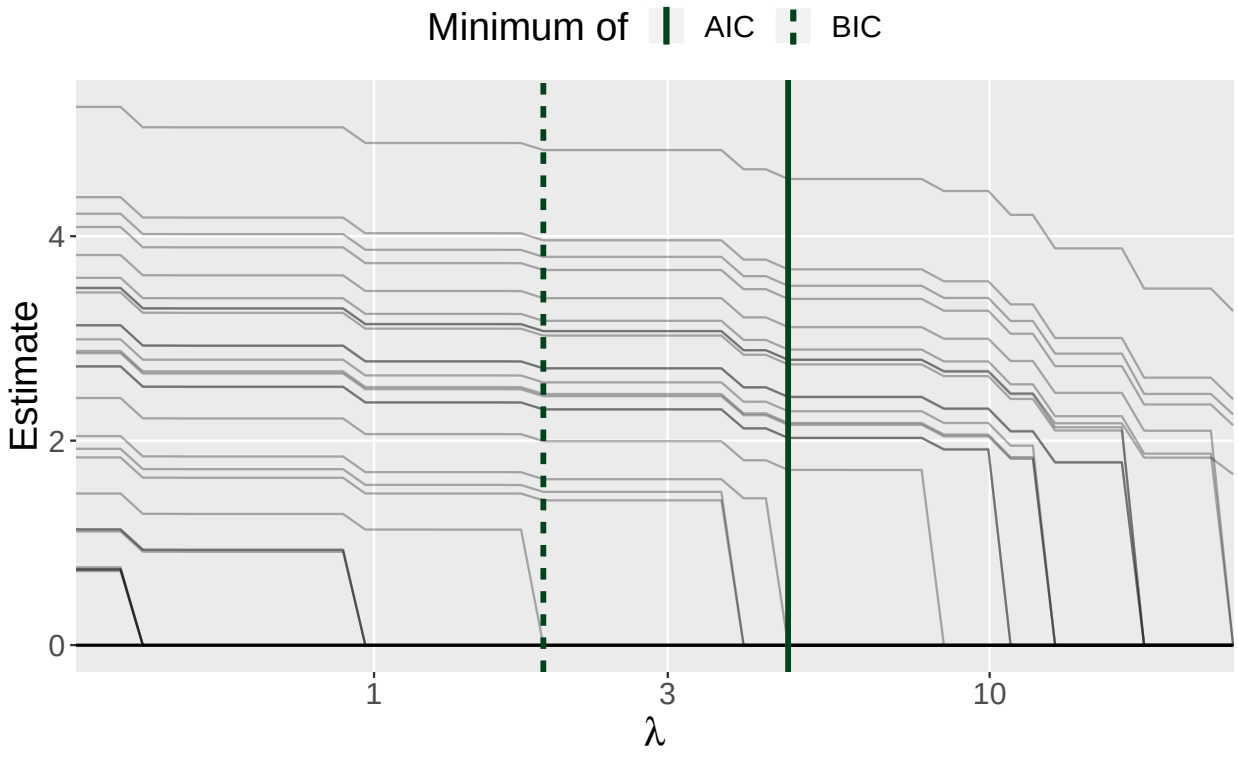

Figure 5: Solution path for **Case B**, with the choice of λ minimizing the AIC and BIC noted

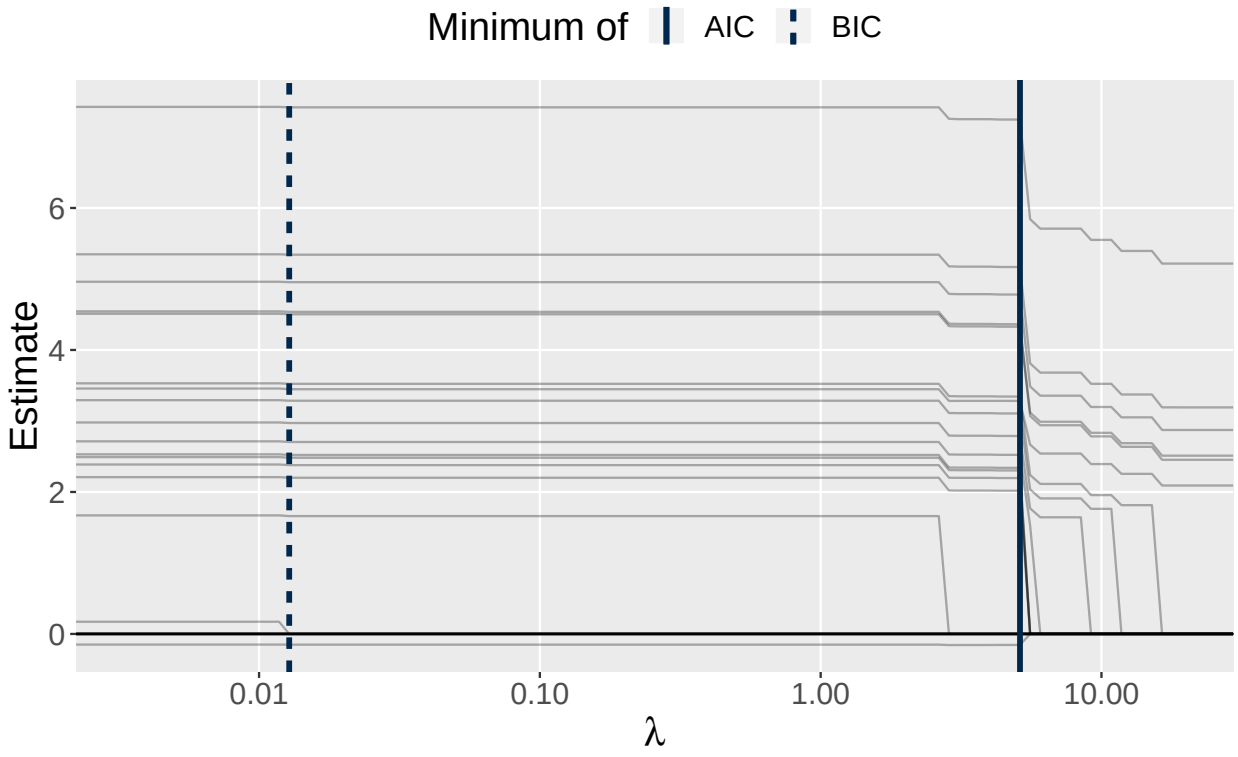

Figure 6: Solution path for **Case C**, with the choice of λ minimizing the AIC and BIC noted