

Loss Function Based Ranking in Two-Stage, Hierarchical Models

Rongheng Lin

National Institute of Environmental Health Science

Thomas A. Louis

Susan M. Paddock

Johns Hopkins Bloomberg School of Public Health

RAND Corporation

Greg Ridgeway

RAND Corporation

Abstract. Performance evaluations of health services providers burgeons. Similarly, analyzing spatially related health information, ranking teachers and schools, and identification of differentially expressed genes are increasing in prevalence and importance. Goals include valid and efficient ranking of units for profiling and league tables, identification of excellent and poor performers, the most differentially expressed genes, and determining “exceedances” (how many and which unit-specific true parameters exceed a threshold). These data and inferential goals require a hierarchical, Bayesian model that accounts for nesting relations and identifies both population values and random effects for unit-specific parameters. Furthermore, the Bayesian approach coupled with optimizing a loss function provides a framework for computing non-standard inferences such as ranks and histograms.

Estimated ranks that minimize Squared Error Loss (SEL) between the true and estimated ranks have been investigated. The posterior mean ranks minimize SEL and are “general purpose,” relevant to a broad spectrum of ranking goals. However, other loss functions and optimizing ranks that are tuned to application-specific goals require identification and evaluation. For example, when the goal is to identify the relatively good (e.g., in the upper 10%) or relatively poor performers, a loss function that penalizes classification errors produces estimates that minimize the error rate. We construct loss functions that address this and other goals, developing a unified framework that facilitates generating candidate estimates, comparing approaches and producing data analytic performance summaries. We compare performance for a fully parametric, hierarchical model with Gaussian sampling distribution under Gaussian and a mixture of Gaussians prior distributions. We illustrate approaches via analysis of standardized mortality ratio data from the United States Renal Data System.

Results show that SEL-optimal ranks perform well over a broad class of loss functions but can be improved upon when classifying units above or below a percentile cut-point. Importantly, even optimal rank estimates can perform poorly in many real-world settings; therefore, data-analytic performance summaries should always be reported.

Keywords: percentiling, Bayesian models, decision theory, operating characteristic

1 Introduction

Performance evaluation burgeons in many areas including health services (Goldstein and Spiegelhalter 1996; Christiansen and Morris 1997; Normand et al. 1997; McClellan and Staiger 1999; Landrum et al. 2000, 2003; Daniels and Normand 2006; Austin and Tu 2006), drug evaluation (DuMouchel 1999), disease mapping (Devine and Louis 1994; Devine et al. 1994; Conlon and Louis 1999; Wright et al. 2003; Diggle et al. 2006), and education (Lockwood et al. 2002; Draper and Gittoes 2004; McCaffrey et al. 2004; Rubin et al. 2004; Tekwe et al. 2004; Noell and Burns 2006). Goals of such investigations include valid and efficient estimation of population parameters such as average performance (over clinics, physicians, health service regions or other “units of analysis”), estimation of between-unit variation (variance components) and unit-specific evaluations. The latter includes estimating unit specific performance, computing the probability that a unit’s true, underlying performance is in a specific region, ranking units for use in profiling and league tables (Goldstein and Spiegelhalter 1996), identification of excellent and poor performers.

Bayesian models coupled with optimizing a loss function provide a framework for computing non-standard inferences such as ranks and histograms and producing data-analytic performance assessments. Inferences depend on the posterior distribution, and how the posterior is used should depend on inferential goals. Gelman and Price (1999) showed that no single set of estimates can simultaneously optimize loss functions targeting the unit-specific parameters (e.g, unit-specific means, optimized by the posterior mean) and those targeting the ranks of these parameters. For example, as Shen and Louis (1998) and Liu et al. (2004) showed, ranking the unit-specific maximum likelihood estimates (MLEs) performs poorly as does ranking Z-scores for testing whether a unit’s mean equals the population mean. In some situations, ranking the posterior means of unit-specific parameters can perform well, but in general an optimal approach to estimate ranks is needed.

In the Shen and Louis (1998) approach, SEL operates on the difference between the estimated and true ranks. But, in many applications interest focuses on identifying the relatively good (e.g., in the upper 10%) or relatively poor performers, a down/up classification. For example, quality improvement initiatives should be targeted at health care providers that have the highest likelihood of being the poorest performers; geography-specific, environmental assessments should be targeted at the most likely high incidence locations (Wright et al. 2003); genes with differential expression in the top 1% (say) should be selected for further study.

We construct new loss functions that focus on down/up classification and derive the optimizers for a subset of them. We develop connections between the new optimizers and others in the literature; report performance evaluations among the new ranking methods and other candidates; identify appropriate uncertainty assessments including a new performance measure. We evaluate performance for a fully parametric hierarchical model with unit-specific Gaussian sampling distributions and assuming either a Gaussian or a mixture of Gaussians prior. We evaluate performance and robustness under the prior and loss function that was used to generate the ranks as well as under

other priors and loss functions. Shen and Louis (1998) showed that when the posterior distributions are stochastically ordered, maximum likelihood estimate based ranks, posterior mean based ranks, SEL-optimal ranks and those based on most other rank-specific loss functions are identical. We report performance assessments for the stochastically ordered case and compare approaches for situations when the posterior distributions are not stochastically ordered. We illustrate approaches using Standardized Mortality Ratio (SMR) data from the United States Renal Data System (USRDS).

2 The two-stage, Bayesian hierarchical model

We consider a two-stage model with independent identically distributed (*iid*) sampling from a known prior G with density g and possibly different unit-specific sampling distributions f_k :

$$\begin{aligned} \theta_1, \dots, \theta_K & \text{ iid } G(\theta_k), k = 1, \dots, K \\ Y_k | \theta_k & \sim f_k(Y_k | \theta_k). \end{aligned} \tag{1}$$

From model (1) we can derive the independent (*ind*) posterior distributions for Bayesian inferences:

$$[\theta_k | Y_k] \text{ ind } g_k(\theta_k | Y_k) = \frac{f_k(Y_k | \theta_k)g(\theta_k)}{\int f_k(Y_k | u)g(u)du}.$$

For computing efficiency, we assume that the θ s are *iid*, though model (1) can be generalized to allow a regression structure in the prior and extended to three stages. Our theoretical results hold for these more general situations.

2.1 Loss functions and decisions

Let $\boldsymbol{\theta} = (\theta_1, \dots, \theta_K)$ and $\mathbf{Y} = (Y_1, \dots, Y_K)$. For a loss function $L(\boldsymbol{\theta}, \mathbf{a})$, the optimal Bayesian $\mathbf{a}(\mathbf{Y})$ minimizes the posterior Bayes risk,

$$\text{Risk}_G(\mathbf{a}(\mathbf{Y}), \mathbf{Y}) = E_{\boldsymbol{\theta}|\mathbf{Y}}[L(\boldsymbol{\theta}, \mathbf{a}(\mathbf{Y})) | \mathbf{Y}],$$

and thereby the pre-posterior risk

$$\text{Risk}_G(\mathbf{a}) = E_{\mathbf{Y}}[\text{Risk}_G(\mathbf{a}(\mathbf{Y}), \mathbf{Y})].$$

Also, for any $\mathbf{a}(\mathbf{Y})$ we can compute the frequentist risk:

$$\text{Risk}(\boldsymbol{\theta}, \mathbf{a}(\cdot)) = E_{\mathbf{Y}|\boldsymbol{\theta}}[L(\boldsymbol{\theta}, \mathbf{a}(\mathbf{Y})) | \boldsymbol{\theta}].$$

3 Ranking

Laird and Louis (1989) represented the ranks by,

$$R_k(\boldsymbol{\theta}) = \text{rank}(\theta_k) = \sum_{j=1}^K I_{\{\theta_k \geq \theta_j\}}, \quad (2)$$

with the smallest θ having rank 1 and the largest having rank K . The non-linear form of (2) implies that, in general, the optimal ranks are neither the ranks of the observed data nor the ranks of the posterior means of the θ s. A loss function is necessary to formalize developing estimates and related uncertainties.

3.1 Squared-error loss (SEL)

Square error loss (SEL) is the most common loss function used in estimation. It is optimized by the posterior mean of the target parameter. For example, under the model (1), with the unit-specific θ s as the target, the loss function is $L(\theta, a) = (\theta - a)^2$ and the optimal estimator is posterior mean (PM) $\theta_k^{pm} = E(\theta_k | \mathbf{Y})$.

When ranks are the target, producing SEL-optimal ranks by minimizing

$$\hat{L} = \hat{L}(\mathbf{R}^{est}, R(\boldsymbol{\theta})) = \frac{1}{K} \sum_k (R_k^{est} - R_k(\boldsymbol{\theta}))^2 \quad (3)$$

and setting R_k^{est} equal to,

$$\bar{R}_k(\mathbf{Y}) = E_{\boldsymbol{\theta}|\mathbf{Y}}[R_k(\boldsymbol{\theta}) | \mathbf{Y}] = \sum_{j=1}^K \text{pr}(\theta_k \geq \theta_j | \mathbf{Y}). \quad (4)$$

The \bar{R}_k are shrunk towards the mid-rank $(K + 1)/2$, and generally are not integers (Shen and Louis 1998). Optimal integer ranks are reached by

$$\hat{R}_k(\mathbf{Y}) = \text{rank}(\bar{R}_k(\mathbf{Y})). \quad (5)$$

See Section Appendix A for additional details on producing optimal ranks under weighted SEL.

3.2 Notation

Henceforth, we drop dependency on $\boldsymbol{\theta}$ and omit conditioning on \mathbf{Y} whenever this does not cause confusion. For example, R_k stands for $R_k(\boldsymbol{\theta})$ and \hat{R}_k stands for $\hat{R}_k(\mathbf{Y})$. Furthermore, use of the ranks facilitates notation in mathematical proofs, but percentiles

$$P_k = R_k/(K + 1); \hat{P}_k = \hat{R}_k/(K + 1), \text{ etc.} \quad (6)$$

normalize large sample performance and aid in communication. For example, Lockwood et al. (2002) showed that mean square error (MSE) for percentiles rapidly converges to a function that does not depend on K ; the same normalization strategy applies in the loss functions below.

4 Upper $100(1 - \gamma)\%$ loss functions

\hat{L} (Equation 3) evaluates general performance without specific attention to identifying the relatively well or poorly performing units. To attend to this goal, for $0 < \gamma < 1$ we investigate loss functions that focus on identifying the upper $100(1 - \gamma)\%$ of the units, with loss depending on the correctness of classification and, possibly, a distance penalty; identification of the lower $100\gamma\%$ group is similar. For notational convenience, we assume that γK is an integer, so $\gamma(K + 1)$ is not an integer and in the following it is not necessary to distinguish between $(>, \geq)$ or $(<, \leq)$.

4.1 Summed, unit-specific loss functions

For $0 < \gamma < 1$, let

$$\begin{aligned} AB_k(\gamma, P_k, P_k^{est}) &= I_{\{P_k > \gamma, P_k^{est} < \gamma\}} = I_{\{R_k > \gamma(K+1), R_k^{est} < \gamma(K+1)\}}, \\ BA_k(\gamma, P_k, P_k^{est}) &= I_{\{P_k < \gamma, P_k^{est} > \gamma\}} = I_{\{R_k < \gamma(K+1), R_k^{est} > \gamma(K+1)\}}. \end{aligned} \quad (7)$$

AB_k and BA_k indicate the two possible modes of misclassification. AB_k indicates that the true percentile is above the cutoff, but the estimated percentile is below the cutoff. Similarly, BA_k indicates that the true percentile is below the cutoff while the estimated percentile is above it.

For $p, q, c \geq 0$ define,

$$\begin{aligned} \tilde{L}(\gamma, p, q, c) &= \frac{1}{K} \sum_k \{|\gamma - P_k^{est}|^p AB_k(\gamma, P_k, P_k^{est}) + c|P_k^{est} - \gamma|^q BA_k(\gamma, P_k, P_k^{est})\}, \\ L^\dagger(\gamma, p, q, c) &= \frac{1}{K} \sum_k \{|P_k - \gamma|^p AB_k(\gamma, P_k, P_k^{est}) + c|\gamma - P_k|^q BA_k(\gamma, P_k, P_k^{est})\}, \\ L^\ddagger(\gamma, p, q, c) &= \frac{1}{K} \sum_k \{|P_k - P_k^{est}|^p AB_k(\gamma, P_k, P_k^{est}) + c|P_k^{est} - P_k|^q BA_k(\gamma, P_k, P_k^{est})\}, \\ L_{0/1}(\gamma) &= \frac{\sum_k \{AB_k(\gamma, P_k, P_k^{est}) + BA_k(\gamma, P_k, P_k^{est})\}}{K} = 2 \frac{\sum_k AB_k(\gamma, P_k, P_k^{est})}{K} \\ &= \frac{\#(\text{misclassifications})}{K} = \tilde{L}(\gamma, 0, 0, 1) = L^\dagger(\gamma, 0, 0, 1) = L^\ddagger(\gamma, 0, 0, 1). \end{aligned} \quad (8)$$

The loss functions confer no penalty if the pair of estimated and true unit-specific percentiles, (P_k^{est}, P_k) , are either both above or both below the γ cut point. If they are on different sides of γ , \tilde{L} penalizes by an amount that depends on the distance of the estimated percentile from γ , L^\dagger by the distance of the true percentile from γ and L^\ddagger by the distance between the true and estimated percentiles. Parameters p and q adjust the intensity of the penalties; $p \neq q$ and $c \neq 1$ allow for different penalties for the two kinds of misclassification. $L_{0/1}(\gamma)$ counts the number of discrepancies and is equivalent to setting $p = q = 0, c = 1$; we use the relation $\sum_k AB_k(\gamma, P_k, P_k^{est}) = \sum_k BA_k(\gamma, P_k, P_k^{est})$ in its definition. In practice, L^\dagger and L^\ddagger would be harder to use than \tilde{L} because their penalizing

quantities depend on unknown P_k . However, inclusion of them in our investigation does help to calibrate the robustness of other estimators.

Our mathematical analyses apply to the general loss functions (8), but our simulations are conducted for $p = q = 2, c = 1$. Within this setting, we denote the first three loss functions in (8) as $\tilde{L}(\gamma)$, $L^\dagger(\gamma)$ and $L^\ddagger(\gamma)$.

We do not investigate the ‘‘all or nothing, experiment-wise’’ loss function with zero loss, when all units are correctly classified as above γ or below γ , and penalty 1 if any unit is misclassified. While this loss function is one of the most fundamental and provides framework in many multiple comparison methods, it does not provide a good guideline in our ranking problem. Finding the optimal classification is challenging in computation and there will be many nearly optimal solutions. Furthermore, in the spirit of computing the false detection rate, loss functions that compute average performance over units or a subset of units are more appropriate in most applications.

5 Optimizers and other candidate ranks

We find ranks/percentiles that optimize $L_{0/1}$ and \tilde{L} and study an estimator that performs well for L^\ddagger , but is not optimal. First, note that optimizers for the loss functions in (8) and \tilde{L} are equal when the posterior distributions are stochastically ordered (the $G_k(t | Y_k)$ never cross). So, in this case \tilde{P}_k , which minimizes \tilde{L} (see equations (3), (5) and (6)), is optimal for a broad class of loss functions (see Theorem 4). Also, it is straightforward to show that all rank/percentile estimators operating through the posterior distribution of the ranks are monotone transform invariant; that is, they are unchanged by monotone transforms of the target parameter.

5.1 Optimizing $L_{0/1}$

Theorem 1. $L_{0/1}$ loss is minimized by

$$\begin{aligned}\tilde{R}_k(\gamma) &= \text{rank}\{\text{pr}(P_k \geq \gamma | \mathbf{Y})\}, \\ \tilde{P}_k(\gamma) &= \tilde{R}_k(\gamma)/(K + 1).\end{aligned}\tag{9}$$

These are not unique optimizers.

Proof. See Section Appendix B. □

5.2 Optimizing \tilde{L}

Theorem 2. The $\tilde{P}_k(\gamma)$ optimize \tilde{L} .

Proof. In Section Appendix C, we show in detail that the $\tilde{P}_k(\gamma)$ are also optimal for more general loss functions with the distance penalties $|\gamma - P_k^{est}|^p$ and $c|P_k^{est} - \gamma|^q$

replaced by any nondecreasing functions of $|\gamma - P_k^{est}|$. The proof has three steps: first, classify the units into (above γ)/(below γ) groups; second, inside each group, rewrite the posterior risk as the inner product of the discrepancy vector and the misclassification probability vector; third, repeatedly use the rearrangement inequality (Hardy et al. 1967) to minimize the inner product. \square

5.2.1 The Normand et al. (1997) estimate

Normand et al. (1997) proposed using the posterior probability $\text{pr}(\theta_k > t | \mathbf{Y})$ and ranks based on it to compare the performance of medical care providers. The cut point t is an application-relevant threshold. Using this approach, we define $P_k^*(\gamma)$ with properly chosen cut point t and show that $P_k^*(\gamma)$ is essentially identical to $\tilde{P}_k(\gamma)$.

Definition of $P_k^*(\gamma)$: Let

$$\bar{G}_{\mathbf{Y}}(t) = \frac{1}{K} \sum_{j=1}^K \text{pr}(\theta_j \leq t | \mathbf{Y}), \quad (10)$$

and define $P_k^*(\gamma)$ as the percentiles produced by ranking the $\text{pr}(\theta_k \geq \bar{G}_{\mathbf{Y}}^{-1}(\gamma) | \mathbf{Y})$. Section 6.2 gives a relation among \bar{P}_k , $\tilde{P}_k(\gamma)$ and P_k^* . Theorems 5 and 6 show that $\tilde{P}_k(\gamma)$ is asymptotically equivalent to $P_k^*(\gamma)$.

By making a direct link to the original θ scale, $P_k^*(\gamma)$ is straightforward to explain and interpret. Furthermore, for a desired accuracy, computing $P_k^*(\gamma)$ is substantially faster than computing $\tilde{P}_k(\gamma)$, since the former requires only accurate computation of individual posterior distributions and of $\bar{G}_{\mathbf{Y}}$.

5.3 Optimizing L^\dagger

Section Appendix D presents an optimization procedure for the case $p = q = 2$, $-\infty < c < \infty$. However, other than use of brute force (complete enumeration), we have not found an algorithm for general (p, q, c) . As for $L_{0/1}$, performance depends only on optimal allocation into the (above γ)/(below γ) groups. Additional criteria are needed to specify the within-group order.

5.4 Optimizing L^\ddagger

We have not found a simple means of optimizing this loss function, but Section Appendix E develops a helpful relation.

5.5 Other ranking estimators

Traditional rank estimators include ranks based on maximum likelihood estimates, those based on posterior means of the θ s and those based on hypothesis testing statistics (Z-scores, P-values). MLE-based ranks are monotone transform invariant, but the others

are not. As shown in Liu et al. (2004), MLE-based ranks will tend to give units with relatively large variances extreme ranks, while hypothesis testing statistics based ranks will tend to place units with relatively small variances at the extremes. Though not an optimal solution to this problem, modified hypothesis testing statistics moderate this shortcoming by reducing the ratio of the variances (Tusher et al. 2001; Efron et al. 2001).

5.5.1 A two stage ranking estimator

Ranking estimator $\tilde{P}_k(\gamma)$ optimize the (above γ)/(below γ) misclassification loss $L_{0/1}$ and \hat{P}_k optimize the \hat{L} , which penalizes on the distance $|P_k^{est} - P_k|$. A convex combination loss function, $\hat{L}_{0/1}^w(\gamma) = (1 - w)L_{0/1}(\gamma) + w\hat{L}$, ($0 \leq w \leq 1$), thus addresses both inferential goals, as L^\ddagger similarly does, and motivates $\hat{\hat{P}}_k(\gamma)$, a two stage hybrid ranking estimator.

Definition of $\hat{\hat{P}}_k(\gamma)$: Use $\tilde{P}_k(\gamma)$ to classify into (above γ)/(below γ) percentile groups. Then, within each percentile group order the estimates by \hat{P}_k .

Theorem 3. $\hat{\hat{P}}_k(\gamma)$ minimizes $L_{0/1}$ and conditional on this (above γ)/(below γ) classification, produces optimal SEL estimates within the two groups.

Proof. See Section Appendix F. □

Furthermore, it is straightforward to show that for $\hat{L}_{0/1}^w(\gamma)$ there exists a $w_* > 0$ such that for all $w \leq w_*$, $\hat{\hat{P}}_k(\gamma)$ is optimal and there exists a $w^* < 1$ such that for all $w \geq w^*$, \hat{P}_k is optimal.

6 Relations among estimators

In this section we develop relations among estimators using ranks or percentiles depending on the context for convenience.

6.1 A general relation

Let $\nu = \lceil \gamma K \rceil$ and define

$$R_k^+(\nu) = \frac{K(K+1)}{2(K-\nu+1)} \text{pr}(R_k \geq \nu).$$

Then the ranked $R_k^+(\nu)$ equal the ranked $\text{pr}(R_k \geq \nu)$ and so each generates the $\tilde{R}_k(\gamma)$. Note that $\frac{K(K+1)}{2(K-\nu+1)}$ is a constant used to standardize $R_k^+(\nu)$ such that:

$$\sum_k \frac{K(K+1)}{2(K-\nu+1)} \text{pr}(R_k \geq \nu) = \sum_k R_k^+(\nu) = \frac{K(K+1)}{2} = \sum_k R_k.$$

Theorem 4. \bar{R}_k is a linear combination of the $R_k^+(\nu)$ with respect to ν and so for any convex loss function the \bar{R}_k outperform the $R_k^+(\nu)$ for at least one value of $\nu = 1, \dots, K$. For SEL, \bar{R}_k dominates for all ν . As shown in Section Appendix A, the $\hat{R}_k = \text{rank}(\bar{R}_k)$ also dominate $\text{rank}(R_k^+(\nu)) = \tilde{R}_k(\gamma)$ for all ν .

Proof. See Section Appendix G. □

6.2 Relating \hat{P}_k , $\tilde{P}_k(\gamma)$ and P_k^*

From (2), (4) and (10), we have that,

$$\begin{aligned}\bar{G}_{\mathbf{Y}}(\theta_k) &= \text{E}[R_k|\theta_k]/K, \\ \bar{R}_k &= \text{E}[R_k] = K\text{E}[\bar{G}_{\mathbf{Y}}(\theta_k)] = \text{E}\{\text{E}[R_k|\theta_k]\}.\end{aligned}$$

The $R_k^*(\gamma)$ are generated by ranking the $\text{pr}(\theta_k \geq \bar{G}_{\mathbf{Y}}^{-1}(\gamma))$, which is equivalent to ranking $\text{pr}(\bar{G}_{\mathbf{Y}}(\theta_k) \geq \gamma)$. By the foregoing, it is equivalent to ranking the $\text{pr}(\text{E}[R_k|\theta_k] \geq \gamma K)$, which is similar to $\text{pr}(R_k \geq \gamma K)$, the generator of $\tilde{P}_k(\gamma)$. The \hat{R}_k are produced by ranking the \bar{R}_k which is the same as ranking the expectation of the random variables used to produce the R_k^* or $\tilde{R}_k(\gamma)$.

6.3 Approximate equivalence of $\tilde{P}_k(\gamma)$ and P_k^*

Theorem 5. Assume that $\theta_k \stackrel{iid}{\sim} G$, $Y_k|\theta_k \stackrel{ind}{\sim} f(Y_k | \theta_k)$ and that the posterior cumulative distribution function (cdf) of each θ_k is continuous and differentiable at $G^{-1}(\gamma)$. If $G_k(\cdot|\mathbf{Y})$ has a universally bounded second moment, then for $K \rightarrow \infty$, $P_k^*(\gamma)$ is equivalent to $\tilde{P}_k(\gamma)$.

Proof. See Section Appendix H. □

Theorem 6. Assume that $\theta_k \stackrel{iid}{\sim} G$, $Y_k|\theta_k \stackrel{ind}{\sim} f(Y_k | \theta_k, \zeta_k)$ and that the posterior cumulative distribution function (cdf) of each θ_k is continuous and differentiable at $G^{-1}(\gamma)$. Furthermore, assume that the empirical distribution function (edf) of the ζ_k converges to a probability distribution. If $G_k(\cdot|\mathbf{Y}, \zeta)$ has a universally bounded second moment, then for $K \rightarrow \infty$, $P_k^*(\gamma)$ is equivalent to $\tilde{P}_k(\gamma)$.

Proof: Regard ζ_k as part of the observed data and use $Y'_k = (Y_k, \zeta_k)$ in Theorem 5.

Theorems 5 and 6 imply that $P_k^*(\gamma)$ is asymptotically optimal for \tilde{L} and provides a loss function basis for the Normand et al. (1997) estimates.

6.4 A unifying score function

We provide a unified approach to loss function based ranking. To this end, we define a non-negative, nondecreasing **scoring function** $S(P) : (0, 1) \rightarrow [0, 1]$. The function

can be regarded as the scores (reward) a unit would get if its percentile was P . It relates percentiles, P , to “consequences” $S(P)$. These relations can help in eliciting an application-relevant loss function and in interpreting loss-function based percentiles. We use SEL for $S(P)$ to produce percentiles and ranks, specifically:

$$L_s = L(\mathbf{S}^{\text{est}}, S(P(\boldsymbol{\theta}))) = \frac{1}{K} \sum_k (S_k^{\text{est}} - S(P_k(\boldsymbol{\theta})))^2. \quad (11)$$

The optimal \mathbf{S}^{est} satisfy $S_k^{\text{est}} = \mathbb{E}[S(P_k)|\mathbf{Y}]$ and we use the ranks and percentiles based on them.

When $S(P) = aP + b$, $a > 0$, i.e. the reward is linear in the estimated percentile, we have $\hat{L}_s = a^2 \hat{L}$ and so \hat{P}_k is associated with linear rewards. When $S(P) = \mathbb{I}_{\{P > \gamma\}}$, i.e., the reward only depends on whether the percentile of a unit is beyond the threshold γ , there is no constraint on the rankings of units inside each of the (above γ)/(below γ) groups. With this setting, there exist many optimizers and $\tilde{P}_k(\gamma)$ is one of them. For the two stage ranking estimator $\hat{\tilde{P}}_k(\gamma)$, let $S(P) = aP + \mathbb{I}_{\{P > \gamma\}}$, and so $S_k^{\text{est}} = a\bar{P}_k + \text{pr}(P_k > \gamma)$. When a is sufficiently close to zero, $\hat{\tilde{P}}_k(\gamma)$ is optimal. More $S(P)$ are given in Section Appendix I.

7 Performance evaluations and comparisons

7.1 Posterior and pre-posterior performance evaluations

In a Bayesian model, an estimator’s performance can be evaluated by using the posterior distribution (data analytic evaluations) and by using the marginal distribution of the data (pre-posterior evaluations). For example, preposterior SEL performance is the sum of expected posterior variance plus expected squared posterior bias.

We provide evaluations relative to the loss function used to produce the estimate and other potentially relevant loss functions. For example, performance with respect to SEL should be computed for the SEL optimizer and for other estimators. These comparisons help to determine the efficiency of an estimator that optimizes one loss function when evaluated for other loss functions. Procedures that are robust to the choice of loss functions will be attractive in applications.

7.2 The (above γ)/(below γ) operating characteristic

For (above γ)/(below γ) classification, plots of the posterior probability of exceeding γ versus estimated percentiles are informative (see Figure 4). Such plots can be summarized by the *a posteriori* operating characteristic (OC). For any percentiling method, define,

$$\begin{aligned} OC(\gamma) &= \text{pr}(P_k < \gamma | P_k^{\text{est}} > \gamma, \mathbf{Y}) + \text{pr}(P_k > \gamma | P_k^{\text{est}} < \gamma, \mathbf{Y}) \\ &= \text{pr}(P_k > \gamma | P_k^{\text{est}} < \gamma, \mathbf{Y}) / \gamma = \frac{\mathbb{E}_{\theta|\mathbf{Y}} L_{0/1}(\gamma)}{2\gamma(1-\gamma)}, \end{aligned} \quad (12)$$

with the last equality following from identity $\sum_k BA_k(\gamma, P_k, P_k^{est}) = \sum_k AB_k(\gamma, P_k, P_k^{est})$. $OC(\gamma)$ is the sum of two misclassification probabilities and so is optimized by $\tilde{P}_k(\gamma)$. It is normalized so that if the data provide no information on the θ_k , then for all γ , $OC(\gamma) \equiv 1$. Evaluating performance using only one of the probabilities, e.g., $\text{pr}(P_k > \gamma | P_k^{est} < \gamma, \mathbf{Y})$ is analogous to computing the false discovery rate (Benjamini and Hochberg 1995; Storey 2002, 2003).

7.3 Unit-specific performance

For loss functions that sum over unit-specific components, performance can also be evaluated for individual units and, in a frequentist evaluation, for individual $\boldsymbol{\theta}$ vectors. These evaluations are in Section 9.3.

8 Simulation scenarios

We evaluate pre-posterior performance for the Gaussian sampling distribution with $K = 200$ using 2000 simulation replications. We compute $\text{pr}(R_k = \ell | \mathbf{Y})$ using an independent sample Monte Carlo with 2000 draws. All simulations are for loss functions with $p = q = 2$ and $c = 1$.

8.1 The Gaussian-Gaussian model

We evaluate estimators for the Gaussian/Gaussian, two-stage model with a Gaussian prior and Gaussian sampling distributions and allow for varying unit-specific variances. Without loss of generality we assume that the prior mean is $\mu = 0$ and the prior variance is $\tau^2 = 1$. Specifically,

$$\begin{aligned} \theta_k & \text{ iid } N(0, 1), \\ Y_k | \theta_k & \sim N(\theta_k, \sigma_k^2). \end{aligned}$$

This derives:

$$\theta_k | Y_k \text{ ind } N(\theta_k^{pm}, (1 - B_k)\sigma_k^2),$$

where $\theta_k^{pm} = (1 - B_k)Y_k$ and $B_k = \sigma_k^2 / (\sigma_k^2 + 1)$. When unit-specific variances ($\sigma_k^2 \equiv \sigma^2$) are all equal, the posterior distributions are stochastically ordered and all ranking methods we investigate are identical. Evaluation for this case provides a baseline performance with respect to the set of loss functions. In practice, the $\{\sigma_k^2\}$'s can vary substantially and we evaluate this situation using two departures from the $\sigma_k^2 \equiv \sigma^2$ case. In each case, the equal variance scenario is produced by $rls = 1$:

log uniform: Ordered, geometric sequences of the $\{\sigma_k^2\}$ with ratio of the largest σ^2 to the smallest $rls = \sigma_K^2 / \sigma_1^2$ and geometric mean $gmv = GM(\sigma_1^2, \dots, \sigma_K^2)$.

two clusters: The first half of the $\sigma_k^2 \equiv (rls)^{-\frac{1}{2}}$; for the second half, $\sigma_k^2 \equiv (rls)^{\frac{1}{2}}$. Here, $rls = \sigma_K^2 / \sigma_1^2$ and $gmv = 1$.

In both cases the variance sequence is monotone in k , but simulation results would be the same if the indices were permuted. These variance sequences — constant, uniform in the log scale, clustered at the extremes of the range — triangulate possible patterns, though specific applications can, of course, have their unique features.

8.2 A Mixture prior

This prior is a mixture of two Gaussian distributions with the mixture constrained to have mean 0 and variance 1:

$$\theta_k \stackrel{iid}{\sim} (1 - \epsilon)N\left(-\frac{\epsilon\Delta}{A}, \frac{1}{A^2}\right) + \epsilon N\left(\frac{(1 - \epsilon)\Delta}{A}, \frac{\xi^2}{A^2}\right)$$

where

$$A^2 = A^2(\epsilon, \Delta, \xi) = (1 - \epsilon) + \epsilon\xi^2 + \epsilon(1 - \epsilon)\Delta^2.$$

We present results for $\epsilon = 0.1$, $\Delta = 3.40$, $\xi^2 = .25$, $\gamma = 0.9$ and compute the preposterior risk for estimators that are computed from the posterior produced by this mixture and for estimators that are based on a standard, Gaussian prior.

9 Simulation results

9.1 SEL for \hat{P}_k and estimated θ -based percentiles

Table 1 documents $SEL(\hat{L})$ performance for \hat{P}_k , the optimal estimator, for percentiled Y_k (the MLE), percentiled θ_k^{pm} and percentiled $\exp\left\{\theta_k^{pm} + \frac{(1 - B_k)\sigma_k^2}{2}\right\}$ (the posterior mean of e^{θ_k}).

The posterior mean of e^{θ_k} is presented to assess performance for a monotone, non-linear transform of the target parameters. For $rls = 1$, the posterior distributions are stochastically ordered and the four sets of percentiles are identical, as is their performance. As rls increases, performance of Y_k -derived percentiles degrades, those based on the θ_k^{pm} are quite competitive with \hat{P}_k but performance for percentiles based on the posterior mean of e^{θ_k} rapidly degrades. Results show that though the posterior mean can perform well for some models and target parameters, in general it is not competitive with rank-based approaches.

9.2 Comparisons among loss function-based estimates

Table 2 reports results for \hat{P}_k , $\tilde{P}_k(\gamma)$ and $\hat{\hat{P}}_k(\gamma)$ under four loss functions and for the “log-uniform” variance pattern. For the “two-clusters” pattern, differences between estimators are modified relative to those for the log-uniform pattern, but preference relations are unchanged. For example, the \hat{L} risks are generally smaller for the “two-clusters” variance pattern than for the “log-uniform” pattern, but the reverse is true for \tilde{L} .

rls	percentiles based on			
	\hat{P}_k	θ_k^{pm}	$\exp\left\{\theta_k^{pm} + \frac{(1-B_k)\sigma_k^2}{2}\right\}$	Y_k
1	516	516	516	516
25	517	517	534	582
100	522	525	547	644

Table 1: Simulated preposterior SEL ($10000\hat{L}$) for $gmv = 1$.

When $rls = 1$, $\tilde{P}_k(\gamma) \equiv \hat{P}_k \equiv \hat{P}_k(\gamma)$ and so differences in the *SEL* results in the first and seventh rows quantify residual simulation variation and Monte Carlo uncertainty in computing the probabilities used in equation (1) to produce the $\tilde{P}_k(\gamma)$. Results for other values of rls show that under \hat{L} , \hat{P}_k outperforms $\tilde{P}_k(\gamma)$ and $\hat{P}_k(\gamma)$ as must be the case, since \hat{P}_k is optimal under SEL. Similarly, $\tilde{P}_k(\gamma)$ optimizes $L_{0/1}$ and \tilde{L} , and for $rls \neq 1$ outperforms competitors. Though $\hat{P}_k(\gamma)$ optimizes $\hat{L}_{0/1}^w(\gamma)$ (see Section 5.4) for sufficiently small w , it performs relatively poorly for the seemingly related L^\ddagger ; $\tilde{P}_k(\gamma)$ appears to dominate and \hat{P}_k performs well. The poor performance of $\hat{P}_k(\gamma)$ shows that unit-specific combining of a misclassification penalty with squared-error loss is fundamentally different from using them in an overall convex combination.

Similar relations among the estimators hold for the two component Gaussian mixture prior and for a ‘‘frequentist scenario’’ with a fixed set of parameters and repeated sampling only from the Gaussian sampling distribution conditional on these parameters.

Results in Table 2 are based on $gmv = 1$. Relations among the estimators for other values of gmv are similar, but a look at extreme gmv is instructive. Results (not shown) indicate that for $rls = 1$, the risk associated with $L_{0/1}$ is of the form $a(gmv)\gamma(1 - \gamma)$, where $a(gmv)$ is a constant depending only on the value of gmv . By identity (12), this implies that the expectation of $OC(\gamma)$ is approximately constant. When $gmv = 0$, the data are fully informative, $Y_k \equiv \theta_k$ and all risks are 0. When $\sigma_k^2 \rightarrow \infty$, $gmv = \infty$ and the Y_k provide no information on the θ s nor on P_k . Table 3 displays the preposterior risk for this no information case, with values providing an upper bound for results in Table 2.

Under $L_{0/1}$ $\tilde{P}_k(\gamma)$ is the optimal and the difference between \hat{P}_k and $\tilde{P}_k(\gamma)$ depends on the magnitude of rls . That $\tilde{P}_k(\gamma)$ is only moderately better than \hat{P}_k under \tilde{L} is due in part to our having considered only the case $p = q = 2, c = 1$, which makes \tilde{L} very similar to \hat{L} . For larger p and q there would be a more substantial difference.

Figures 1-3 are based on the Gaussian-Gaussian model. Figure 1 displays the dependence of risk on gmv for the exchangeable model ($rls = 1$). As expected, risk increases

γ	rls	$L_{0/1}$		\hat{L}			\tilde{L}			L^\ddagger		
		\hat{P}_k	$\tilde{P}_k(\gamma)$	\hat{P}_k	$\tilde{P}_k(\gamma)$	$\hat{P}_k(\gamma)$	\hat{P}_k	$\tilde{P}_k(\gamma)$	$\hat{P}_k(\gamma)$	\hat{P}_k	$\tilde{P}_k(\gamma)$	$\hat{P}_k(\gamma)$
0.5	1	2508	2511	517	518	517	104	105	104	336	337	336
0.5	25	2506	2508	519	524	519	98	96	98	340	335	340
0.5	100	2503	2503	521	530	521	93	90	93	342	334	342
0.6	100	2432	2422	522	537	523	91	87	91	324	316	327
0.8	25	1740	1717	517	558	517	67	59	67	175	170	181
0.8	100	1742	1689	523	595	523	71	57	71	178	170	189
0.9	1	1059	1058	515	520	515	30	30	30	73	73	73
0.9	25	1060	1032	518	609	519	37	29	37	75	72	81
0.9	100	1048	1005	523	673	523	43	29	42	77	70	86
0.8	1	1469	1471	565	567	565	54	54	54	150	150	150
0.8	25	1524	1494	566	606	567	59	51	59	161	158	168
0.9	1	782	782	565	575	565	14	14	14	42	42	42
0.9	25	823	783	564	699	564	23	14	23	51	48	58
0.8	1	1473	1470	565	566	565	54	54	54	150	150	150
0.8	25	1496	1493	567	615	567	58	52	59	159	158	168
0.9	1	782	782	565	570	565	14	14	14	42	42	42
0.9	25	788	783	565	664	564	22	14	23	50	45	58

Table 2: Simulated preposterior risk for $gmv = 1$. All values are $10000 \times (\text{Loss})$. The first block is for the Gaussian-Gaussian model; the second for the Gaussian mixture prior assuming the mixture; the third for the Gaussian mixture prior, but with analysis based on a single Gaussian prior.

γ	$L_{0/1}$	\hat{L}	L^\ddagger	\tilde{L} and L^\ddagger
	$200\gamma(1-\gamma)$	1667	$3333\gamma(1-\gamma)$	$3333\gamma(1-\gamma)[\gamma^3 + (1-\gamma)^3]$
0.5	5000	1667	833	208
0.6	4800	1667	800	224
0.8	3200	1667	533	277
0.9	1800	1667	300	219

Table 3: Preposterior risk for $rls = 1$ when $gmv = \infty$. All values are $10000 \times \text{Risk}$.

with gmv . For $rls = 1$, expected unit-specific loss equals the overall average risk and so the box plots summarize the sampling distribution of unit-specific risk.

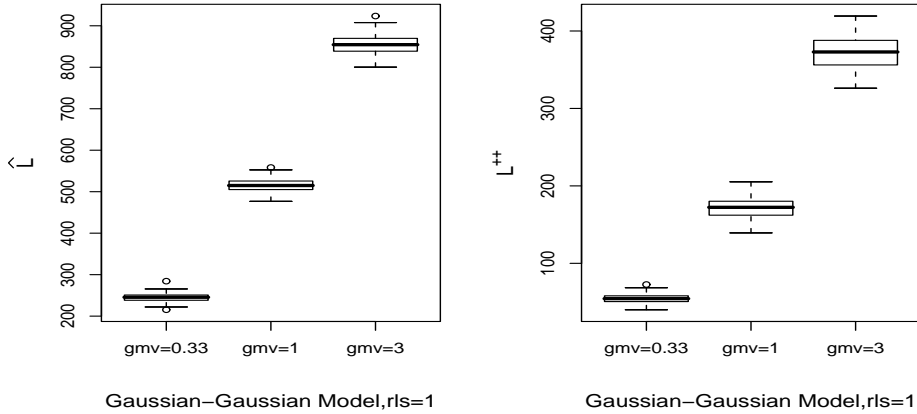


Figure 1: Unit-specific, \hat{L} and L^{\ddagger} risk classified by gmv for $K = 200$, $\gamma = 0.8$.

9.3 Unit-specific performance

When $rls = 1$, pre-posterior risk is the same for all units. However, when $rls > 1$, the σ_k^2 form a geometric sequence and preposterior risk depends on the unit. We study this non-exchangeable situation by simulation. Figure 2 displays loess smoothed performance of \hat{P}_k , $\tilde{P}_k(\gamma)$ and $\hat{P}_k(\gamma)$ for $L_{0/1}$, \hat{L} and L^{\ddagger} as a function of unit-specific variance for $gmv = 1$ and 3 , $rls = 100$ and $\gamma = 0.8$. Results for \hat{L} ($gmv = 3$) and $L_{0/1}$ ($gmv = 1$) are intuitive in that risk increases with increasing unit-specific variance. However, in the displays for $L_{0/1}$ ($gmv = 3$) and for L^{\ddagger} , for all estimators the risk increases and then decreases as a function of σ_k^2 . For gmv and rls sufficiently large, similar patterns hold for other γ -values with the presence and location of a downturn depending on $|\gamma - 0.5|$.

These apparent anomalies are explained as follows. If γ is near 1 (or equivalently, near 0) and if the σ_k^2 differ sufficiently ($rls \gg 1$), estimates for the high variance units perform better than for those with mid-level variance. This relation is due to the improved classification of high-variance units into (above γ)/(below γ) groups, with substantial shrinkage of the percentile estimates towards 0.5. For example, with $\gamma = 0.8$, *a priori* 80% of the percentiles should be below 0.8. Estimated percentiles for the high variance units are essentially guaranteed to be below 0.8 and so the classification error for the large-variance units converges to 0.20 as $rls \rightarrow \infty$. Generally, low variance units have small misclassification probabilities, but percentiles for units with intermediate variances are not shrunken sufficiently toward 0.5 to produce a low $L_{0/1}$.

9.4 Classification performance

As shown in the foregoing tables and by Liu et al. (2004) and Lockwood et al. (2002), even the optimal ranks and percentiles can perform poorly unless the data are very informative. Figure 3 displays average posterior classification probabilities as a function of the optimally estimated percentile for $gmv = 0.33, 1, 10, 100$ and $\gamma = 0.6, 0.8, 0.9$, when $rls = 1$. The pattern shown by the three panels should hold for other γ choices and we use the $\gamma = 0.8$ panel as the typical example for further comments. Discrimination improves with decreasing gmv , but even when $gmv = 0.33$ (the σ_k are $1/3$ of the prior variance), the model-based, posterior probability of $P_k > 0.8$ is only 0.42 for a unit with $\tilde{P}_k(0.8) = 0.8$. For this probability to exceed 0.95 (i.e., to be reasonably certain that $P_k > 0.80$) requires that $\tilde{P}_k(0.8) > 0.97$. It can be shown that as $gmv \rightarrow \infty$ the plots converge to a horizontal line at $(1 - \gamma)$ and that as $gmv \rightarrow 0$ the plots converge to a step function that jumps from 0 to 1 at γ .

9.5 The Poisson-Gamma model

We conducted investigations analogous to the all of the foregoing for the Poisson sampling distribution with a Gamma prior, with constant or unequal variances (e.g., expected values) for the unit-specific MLEs. Results are qualitatively and quantitatively very similar to those we report for the Gaussian sampling distribution.

10 Analysis of USRDS standardized mortality ratios

The United States Renal Data System (USRDS) uses provider specific, standardized mortality ratios (SMRs) as a quality indicator for its nearly 4000 dialysis centers (Lacson et al. 2001; End-Stage Renal Disease (ESRD) Network 2000; United States Renal Data System (USRDS) 2005). Under the Poisson likelihood (last line of model (13)), with Y_k the observed and m_k the expected deaths computed from a case-mix adjustment (Wolfe et al. 1992), the MLE is $\hat{\rho}_k = Y_k/m_k$, with variance ρ_k/m_k . For the “typical” dialysis center $\rho_k \approx 1$ and the m_k control the variance of the MLEs. The observed m_k s range from around 0 to greater than 100. The ratio of the largest m_k to the smallest m_k , which is analogous to the rls in the foregoing simulations, is around 258,000.

In this “profiling” application, the loss function should reflect the end use of the ranks or percentiles. For example, suppose that the following monetary reward (increased reimbursement) and penalty (increased scrutiny) system is in place:

- A provider either does or does not receive the reward depending on whether its percentile is or is not beyond (for SMRs, below) a γ threshold.
- Providers that do get rewards receive varying amounts depending on their position among those receiving rewards.
- Providers not receiving rewards undergo increased scrutiny from an oversight committee.

- The distance from the threshold γ is used to monetize rewards or intensify scrutiny.

Loss function \tilde{L} embodies this system with the values of p , q and c controlling the award/penalty differences within the (above γ)/(below γ) groups. If rewarded providers all get the same amount, then $L_{0/1}(\gamma)$ can be used. Alternatively, rewards and scrutiny can depend on the posterior probability of exceeding or falling below γ , with both $\tilde{P}_k(\gamma)$ and $P_k^*(\gamma)$ optimizing the evaluation.

Liu et al. (2004) analyzed 1998 data and Lin et al. (2004) extended these analyses to 1998 – 2001 for 3173 centers with complete data using an autoregressive model. We illustrate the new loss functions and performance measures using 1998 data and the model,

$$\begin{aligned} \xi &\stackrel{iid}{\sim} N(0, 10), & \lambda = \tau^{-2} &\stackrel{iid}{\sim} \text{Gamma}(0.05, 0.2) & (13) \\ [\theta_1, \dots, \theta_K \mid \xi, \tau] &\stackrel{iid}{\sim} N(\xi, \tau^2), & \theta_k &= \log(\rho_k) \\ [Y_k \mid m_k, \rho_k] &\sim \text{Poisson}(m_k \rho_k). \end{aligned}$$

For these data, $\bar{G}_K^{-1}(0.8) = 0.18$ ($\rho = 1.20$). Table 4 gives the posterior risks. For all loss functions investigated, the MLE based rank has the poorest performance; methods based on the posterior distributions generally perform well. As Theorem 6 indicates, $\tilde{P}_k(\gamma)$ and $P_k^*(\gamma)$ have almost identical risk.

Figure 4 displays $\text{pr}(\theta_k > 0.18 \mid \mathbf{Y})$ with X-axis percentiles determined separately by the three percentiling methods. As shown by Theorem 6, the $P_k^*(0.8)$ and $\tilde{P}_k(0.8)$ curves are monotone and approximately equal; the \hat{P}_k curve is not monotone, but is close to the other curves. The $OC(0.8)$ value for $P_k^*(\gamma)$ and $\tilde{P}_k(0.8)$ is 0.64 (the optimal classification produces an error rate that is 64% of that for the no information case) and for \hat{P}_k is 0.65, showing that for $\gamma = 0.8$, using \hat{P}_k to classify is nearly fully efficient. Figure 4 also shows that for centers classified in the top 10%, the probability that they are truly in the top 20% ($\gamma = 0.8$) can be as low as 0.45. Lin et al. (2004) showed that, by using data from 1998-2001, this probability increases to 0.57. Evaluators should take this relatively poor classification performance into account by tempering rewards and scrutiny.

Figure 5 displays the relation between $\tilde{P}_k(0.8)$ and \hat{P}_k for 50 dialysis centers spaced uniformly according to $\tilde{P}_k(0.8)$. Since $\tilde{P}_k(0.8)$ is based on the $\text{pr}(P_k > .8 \mid \mathbf{Y})$ calculated from MCMC samples, ties appear when this probability is close to zero. Among 3173 dialysis centers, 249 centers have the exceeding probability 0 and all are estimated with percentile $125/3174=0.039$. Though \hat{P}_k is highly efficient, some percentiles are substantially different from the optimal. As further evidence of this discrepancy, of the 635 dialysis centers classified by $\tilde{P}_k(0.8)$ in the top 20%, 39 are not so classified by \hat{P}_k with most of these near the $\gamma = 0.8$ threshold. Estimated percentiles are very similar for centers classified in the top 10%.

	\hat{P}_k^{PM}	\hat{P}_k^{MLE}	\hat{P}_k	$\tilde{P}_k(\gamma)$	$\hat{\tilde{P}}_k(\gamma)$	$P_k^*(\gamma)$
\hat{L}	741	872	740	769	741	769
\tilde{L}	108	164	107	100	107	100
L^\dagger	97	130	96	102	102	102
L^\ddagger	281	401	279	275	285	276
$L_{0/1}$	2001	2062	2006	1992	1991	1995

Table 4: Posterior, loss function risk for different ranking methods using the USRDS 1998 data. All values are $10000 \times (\text{Loss})$. \hat{P}_k^{PM} and \hat{P}_k^{MLE} are percentiles based on the θ_k^{pm} and the Y_k respectively.

11 Discussion

Effective ranking or percentiling should be based on a loss function computed from the estimated and true ranks, or be asymptotically equivalent to such loss function based estimates. Doing so produces optimal or near optimal performance and ensures desirable properties such as monotone transform invariance. In general, percentiles based on MLEs or on posterior means of the target parameter can perform poorly. Similarly, hypothesis test-based percentiles perform poorly.

Our performance evaluations are primarily for the fully parametric model with a Gaussian sampling distribution, though we do investigate departures from the Gaussian prior. Simulations for the Poisson/Gamma model produce relative performance very similar to those for the Gaussian. The \hat{P}_k that optimize \hat{L} (SEL) are “general purpose” with no explicit attention to optimize the (above γ)/(below γ) classification. The optimal (above γ)/(below γ) ranks are asymptotically equivalent to the “exceedance probability” procedure proposed in Normand et al. (1997). This near-equivalence provides insight into goals and a route to efficient computation.

We report loss function comparisons and plots based on unit-specific performance. These can be augmented by bivariate and multivariate summaries of properties, for example pair-wise posterior distributions or pair-wise operating characteristics.

When posterior distributions are not stochastically ordered and the choice of ranking methods does matter, our simulations show that though $\tilde{P}_k(\gamma)$ and $\hat{\tilde{P}}_k(\gamma)$ are optimal for their respective loss functions and outperform \hat{P}_k , \hat{P}_k performs well for a broad range of γ values. And, $\tilde{P}_k(\gamma)$ can have poor SEL performance. However, for some scenarios the relative benefit of using an optimal procedure is considerable and so a choice of estimator should be guided by goals.

Performance evaluations for three-level models with a hyper-prior and robust analyses based on the non-parametric maximum likelihood prior or a fully Bayesian nonparametric prior (Paddock et al. 2006) showed that SEL-optimal ranks perform well over a wide range of prior specifications.

Other loss functions and estimates can be considered. Weighted combinations of several loss functions can be used to broaden the class of loss functions. If an application-

relevant loss function cannot be optimized, our evaluations provide a framework to compare candidate estimators. Our scoring function approach can help practitioners elicit a meaningful loss function with an intuitive interpretation.

Though there are a wide variety of candidate loss functions and, thereby, candidate estimated percentiles, our investigations show that in most applications one can choose between \hat{P}_k and $P_k^*(\gamma)$ (equivalently, $\tilde{P}_k(\gamma)$). The \hat{P}_k are for general purpose and are recommended in situations where the full spectrum of percentiles is important, for identifying units as low, medium or high performers. This is the case in educational assessments. Schools and school districts want to track their performance over time irrespective of whether they are low, high or in the middle. The $P_k^*(\gamma)$ focus on a specific (above γ)/(below γ) cut point and are recommended in situations where identifying one extreme is the dominant goal. Selection of the most differentially expressed genes, with γ selected to deliver a manageable number for further analysis, is a prime example.

Whatever percentiling method is used, plots such as Figure 4 can be constructed with those percentiles on the X-axis. In general, the plot will not be monotone unless the $P_k^*(\gamma)$ are used, but use of the \hat{P}_k produces a nearly monotone plot and very good $OC(\gamma)$ performance. Therefore, unless there is a compelling reason to optimize relative to a specific (above γ)/(below γ) cut point, the \hat{P}_k are preferred.

Importantly, as do Liu et al. (2004) and Lockwood et al. (2002), we show that unless data are highly informative, even the optimal estimates can perform poorly. It is thus very important to select proper estimates for as good as possible inference, especially when performance differences between estimators are large. Data analytic performance summaries such as SEL , $OC(\gamma)$ and plots like Figures 3 and 4 should accompany any analysis.

Appendix

Appendix A Optimizing weighted squared error loss (WSEL)

Theorem 7. *Under weighted squared error loss:*

$$\sum_k \omega_k (R_k^{est} - R_k)^2, \quad (14)$$

the optimal rank estimates are

$$\bar{R}_k = E(R_k | \mathbf{Y}) = \sum_j pr(\theta_k \geq \theta_j | \mathbf{Y}).$$

Proof. (We drop conditioning on \mathbf{Y})

$$\begin{aligned} \mathbb{E} \sum_k \omega_k (R_k^{est} - R_k)^2 &= \sum_k \omega_k \mathbb{E} (R_k^{est} - \bar{R}_k + \bar{R}_k - R_k)^2 \\ &= \sum_k \omega_k \mathbb{E} [(R_k^{est} - \bar{R}_k)^2 + (\bar{R}_k - R_k)^2] \\ &\geq \sum_k \omega_k \mathbb{E} (\bar{R}_k - R_k)^2 \end{aligned}$$

Thus, the \bar{R}_k are optimal.

When all $w_k \equiv w$,

$$\hat{R}_k = \text{rank of } (\bar{R}_k)$$

optimizes (14) subject to the R_k^{est} exhausting the integers $(1, \dots, K)$. To see this, if $0 \leq \mathbb{E}(R_i) = m_i \leq \mathbb{E}(R_j) = m_j$, $r_i < r_j$, then

$$\begin{aligned} \mathbb{E}(R_i - r_i)^2 + \mathbb{E}(R_j - r_j)^2 &= \text{Var}(R_i) + \text{Var}(R_j) + (m_i - r_i)^2 + (m_j - r_j)^2 \\ &< \text{Var}(R_i) + \text{Var}(R_j) + (m_i - r_j)^2 + (m_j - r_i)^2 \\ &= \mathbb{E}(R_i - r_j)^2 + \mathbb{E}(R_j - r_i)^2 \end{aligned}$$

and the \hat{R}_k are optimal. \square

For general w_k there is no closed form solution, but a sorting-based algorithm based on,

$$\begin{aligned} \omega_i(m_i - r_i)^2 + \omega_j(m_j - r_j)^2 &< \omega_i(m_i - r_j)^2 + \omega_j(m_j - r_i)^2, \text{ if } r_j > r_i \\ \iff (r_j - r_i) \left(\left(1 - \frac{\omega_j}{\omega_i}\right) (r_i + r_j - 2m_j) + 2(m_j - m_i) \right) &> 0, \text{ if } r_j > r_i \\ \iff \left(1 - \frac{\omega_j}{\omega_i}\right) (r_i + r_j - 2m_j) + 2(m_j - m_i) &> 0, \text{ if } r_j > r_i. \end{aligned} \quad (15)$$

guides the optimization. By the above inequality, reversing any two estimated ranks that do not align with \bar{R}_k results in a smaller squared error.

Theorem 8. *Start from any initial ranks and implement the recursion: If inequality (15) is satisfied, switch the position of unit i and unit j , $i, j = 1, \dots, K$. The unique fixed point will minimize weighted squared error loss (14).*

Proof. Since each switch will decrease the expected loss and there are at most $n!$ possible values of the expected loss, a fixed point will be reached. At the fixed point, no (i, j) pair produces inequality (15) and so gives the SEL minimum. \square

In the standard sorting problem, the quantities to sort do not depend on the current positions of the units, while the quantity in (15) does. For this reason, the convergence of the algorithm can be very slow. After units i and j are compared and ordered, if unit i is compared to some other unit k and a switch happens, then unit i should be compared to unit j again and so this pairwise-switch optimization algorithm is computationally impractical.

Appendix B Optimizing $L_{0/1}$

Proof of Theorem 1

Proof. Rewrite the loss function as a function of the number of units that not classified in the top $(1 - \gamma)K$, but that should have been. Then, $L_{0/1} = \frac{1}{K}(K - |A \cap T|)$, where A is the set of indices of the observations classified in the top and T is the true set of indices for which $\text{rank}(\theta_k) > (1 - \gamma)K$. We need to maximize the expected number of correctly classified coordinates:

$$\begin{aligned} \mathbb{E}|A \cap T| &= \mathbb{E} \sum I(k \in A \cap T) \\ &= \mathbb{E} \sum_{k \in A} I(k \in T) = \sum_{k \in A} \text{pr}(P_k > \gamma | \mathbf{Y}). \end{aligned}$$

To optimize $L_{0/1}$, for each θ_k calculate $\text{pr}(P_k > \gamma | \mathbf{Y})$, rank these probabilities and select the largest $(1 - \gamma)K$ of them to minimize $L_{0/1}$, creating the optimal (above γ)/(below γ) classification. This computation can be time-consuming, but is Monte Carlo implementable.

The $\tilde{P}_k(\gamma)$ optimize $L_{0/1}$. There are other optimizers because $L_{0/1}$ requires only the optimal (above γ)/(below γ) categorization but not the optimal ordering. For example, permutations of the ranks of units classified in A or permutations of the ranks in A^C yield the same posterior risk for $L_{0/1}$. \square

Appendix C Optimizing \tilde{L}

Lemma 1. *If $a_1 + a_2 \geq 0$ and $b_1 \leq b_2$, then*

$$a_1 b_1 + a_2(1 - b_2) \leq a_1 b_2 + a_2(1 - b_1).$$

Proof.

$$\begin{aligned} (a_1 + a_2)b_1 \leq (a_1 + a_2)b_2 &\Rightarrow a_1 b_1 - a_2 b_2 \leq a_1 b_2 - a_2 b_1 \\ &\Rightarrow a_1 b_1 + a_2(1 - b_2) \leq a_1 b_2 + a_2(1 - b_1). \end{aligned}$$

\square

Lemma 2. *Rearrangement Inequality (Hardy et al. 1967): If $a_1 \leq a_2 \leq \dots \leq a_n$ and $b_1 \leq b_2 \leq \dots \leq b_n$, $b_{(1)}, b_{(2)}, \dots, b_{(n)}$ is a permutation of b_1, b_2, \dots, b_n , then*

$$\sum_{i=1}^n a_i b_{n+1-i} \leq \sum_{i=1}^n a_i b_{(i)} \leq \sum_{i=1}^n a_i b_i.$$

Proof. For $n = 2$ we use the ranking inequality:

$$a_1 b_2 + a_2 b_1 \leq a_1 b_1 + a_2 b_2 \Leftrightarrow (a_2 - a_1)(b_2 - b_1) \geq 0.$$

For $n > 2$, there exists a minimum and a maximum in all $n!$ combinations of sums of products. By the result for $n = 2$, the necessary condition for the sum to reach the minimum is that any pair of indices (i_1, i_2) , (a_{i_1}, a_{i_2}) and (b_{i_1}, b_{i_2}) must have the inverse order; to reach the maximum, they must have same order. Therefore, except in the trivial cases where there are ties inside $\{a_i\}$ or $\{b_i\}$, $\sum_{i=1}^n a_i b_{n+1-i}$ is the only candidate to reach the minimum and $\sum_{i=1}^n a_i b_i$ is the only candidate to reach the maximum. \square

Proof of Theorem 2 Denote by $R_{(i)}$ the rank random variables for units whose ranks are estimated as i . Then,

$$\begin{aligned} E(L_{R_K^{est}}(\gamma, p, q, c)) &= \sum_{i=1}^{\lceil \gamma(K+1) \rceil} |\gamma(K+1) - i|^p \text{pr}(R_{(i)} \geq \gamma(K+1)) \\ &+ \sum_{i=\lceil \gamma(K+1) \rceil + 1}^K c|i - \gamma(K+1)|^q (1 - \text{pr}(R_{(i)} \geq \gamma(K+1))). \end{aligned}$$

For optimum ranking, the following conditions are necessary:

1. By Lemma 1, for any (i_1, i_2) satisfying $(1 \leq i_1 \leq \lceil \gamma(K+1) \rceil, \lceil \gamma(K+1) \rceil + 1 \leq i_2 \leq K)$, it is required that $\text{pr}(R_{(i_1)} \geq \gamma(K+1)) \leq \text{pr}(R_{(i_2)} \geq \gamma(K+1))$. To satisfy this condition, divide the units into two groups by picking the units with largest $(1 - \gamma)K$ values of $\text{pr}(R_k \geq \gamma(K+1))$ into the (above γ) group.
2. By Lemma 2
 - (a) For the set $\{k : R_k = R_{(i)}, i = 1, \dots, \lceil \gamma(K+1) \rceil\}$, since $|\gamma(K+1) - i|^p$ is a decreasing function of i , we require that $\text{pr}(R_{(i_1)} \geq \gamma(K+1)) \geq \text{pr}(R_{(i_2)} \geq \gamma(K+1))$ if $i_1 > i_2$. Therefore, for the units with ranks $(1, \dots, \gamma K)$, the ranks should be determined by ranking the $\text{pr}(R_k \geq \gamma(K+1))$.
 - (b) For the set $\{k : R_k = R_{(i)}, i = \lceil \gamma(K+1) \rceil + 1, \dots, K\}$, since $|i - \gamma(K+1)|^q$ is an increasing function of i , we require that $\text{pr}(R_{(i_1)} \geq \gamma(K+1)) \geq \text{pr}(R_{(i_2)} \geq \gamma(K+1))$ if $i_1 > i_2$. Therefore, for the units with ranks $(\gamma K + 1, \dots, K)$, the ranks should be determined by ranking the $\text{pr}(R_k \geq \gamma(K+1))$.

These conditions imply that the $\tilde{R}_k(\gamma)$ ($\tilde{P}_k(\gamma)$) are optimal. By the proof of Lemma 2, we know that the optimization is not unique, when there are ties in $\text{pr}(R_k \geq \gamma(K+1))$.

Appendix D Optimization procedure for L^\dagger

As in the proof of Theorem 2, we begin with a necessary condition for optimization. Denote by $R_{(i_1)}, R_{(i_2)}$ the rank random variables for units whose ranks are estimated as i_1, i_2 , where $i_1 < \gamma(K+1), i_2 > \gamma(K+1)$. Let,

$$\text{pr}(R_{(i_1)} \geq \gamma(K+1)) = p_1, \text{pr}(R_{(i_2)} \geq \gamma(K+1)) = p_2.$$

For the index selection to be optimal,

$$\begin{aligned} & \mathbb{E}[(R_{(i_1)} - \gamma(K+1))^2 | R_{(i_1)} \geq \gamma(K+1)]p_1 + c\mathbb{E}[(R_{(i_2)} - \gamma(K+1))^2 | R_{(i_2)} < \gamma(K+1)](1-p_2) \\ & \leq c\mathbb{E}[(R_{(i_1)} - \gamma(K+1))^2 | R_{(i_1)} < \gamma(K+1)](1-p_1) + \mathbb{E}[(R_{(i_2)} - \gamma(K+1))^2 | R_{(i_2)} \geq \gamma(K+1)]p_2. \end{aligned}$$

The following is equivalent to the foregoing:

$$\begin{aligned} & \mathbb{E}[(R_{(i_1)} - \gamma(K+1))^2 | R_{(i_1)} \geq \gamma(K+1)]p_1 - c\mathbb{E}[(R_{(i_1)} - \gamma(K+1))^2 | R_{(i_1)} < \gamma(K+1)](1-p_1) \\ & \leq \mathbb{E}[(R_{(i_2)} - \gamma(K+1))^2 | R_{(i_2)} \geq \gamma(K+1)]p_2 - c\mathbb{E}[(R_{(i_2)} - \gamma(K+1))^2 | R_{(i_2)} < \gamma(K+1)](1-p_2). \end{aligned}$$

Therefore, with $p_k = \text{pr}(R_k \geq \gamma(K+1))$ the optimal ranks split the θ s into a lower fraction and an upper fraction by ranking the quantity,

$$\mathbb{E}[(R_k - \gamma(K+1))^2 | R_k \geq \gamma(K+1)]p_k - c\mathbb{E}[(R_k - \gamma(K+1))^2 | R_k < \gamma(K+1)](1-p_k).$$

This result is useful and different from that of WSEL in Section Appendix A in the sense that we can now successfully get a quantity depend on unit index k only. However, as for $L_{0/1}$ optimization of L^\dagger does not induce an optimal ordering in the two groups. A second stage loss, for example SEL, can be imposed within the two groups.

Appendix E Optimizing L^\ddagger

As for optimizing WSEL in Section Appendix A, a pairwise switch algorithm is computationally challenging, since the decision on switching a pair of units depends on their relative position and on their estimated ranks. Thus, in each iteration all pairwise relations have to be checked. We have not identified a general representation or efficient algorithm for the optimal ranks. However, we have developed the following relation between L^\dagger , \tilde{L} and L^\ddagger . Note that when either $AB_k(\gamma, P_k, P_k^{est}) \neq 0$ or $BA_k(\gamma, P_k, P_k^{est}) \neq 0$ it must be the case that either $P_k^{est} \geq \gamma \geq P_k$ or $P_k \geq \gamma \geq P_k^{est}$. Equivalently,

$$|P_k - \gamma| + |P_k^{est} - \gamma| = |P_k - P_k^{est}| \text{ or } \frac{|P_k - \gamma|}{|P_k - P_k^{est}|} + \frac{|P_k^{est} - \gamma|}{|P_k - P_k^{est}|} = 1.$$

Now, suppose $c > 0, p \geq 1, q \geq 1$ and let $m = \max(p, q)$. Then, using the inequality $2^{1-m} \leq a^m + (1-a)^m \leq 1$ for $0 \leq a \leq 1$, we have that $(\tilde{L} + L^\dagger) \leq L^\ddagger \leq 2^{m-1}(\tilde{L} + L^\dagger)$. Specifically, if $p = q = 1$, $L^\ddagger = \tilde{L} + L^\dagger$; if $p = q = 2$, then $(\tilde{L} + L^\dagger) \leq L^\ddagger \leq 2(\tilde{L} + L^\dagger)$. Similarly, when $c > 0, p \leq 1, q \leq 1$, $(\tilde{L} + L^\dagger) \geq L^\ddagger \geq 2^{m-1}(\tilde{L} + L^\dagger)$. Therefore, \tilde{L} and L^\dagger can be used to control L^\ddagger .

Appendix F Proof of Theorem 3

Proof. Since the (above γ)/(below γ) groups are formed by $\tilde{P}_k(\gamma)$, $\hat{P}_k(\gamma)$ minimizes $L_{0/1}$. For constrained SEL minimization we prove the more general result that for any (above γ)/(below γ) categorization, ordering within the groups by \hat{P}_k produces the

constrained solution. To see this, without loss of generality, assume that coordinates $(1, \dots, \gamma K)$ are in the (below γ) group and $(\gamma K + 1, \dots, K)$ are in the (above γ) group. Similar to Section Appendix A,

$$\mathbb{E} \sum_k (R_k^{est} - R_k)^2 = \sum_k V(R_k) + \sum_k (R_k^{est} - \bar{R}_k)^2.$$

Nothing can be done to reduce the variance terms. The summed squared bias partitions into,

$$\sum_k (R_k^{est} - \bar{R}_k)^2 = \sum_{k=1}^{\gamma K} (R_k^{est} - \bar{R}_k)^2 + \sum_{k=\gamma K+1}^K (R_k^{est} - \bar{R}_k)^2$$

which must be minimized subject to the constraints that $(R_1^{est}, \dots, R_{\gamma K}^{est}) \in \{1, \dots, \gamma K\}$ and $(R_{\gamma K+1}^{est}, \dots, R_K^{est}) \in \{\gamma K + 1, \dots, K\}$. We deal only with the (below γ) group; the (above γ) group is handled in the same manner. Without loss of generality assume that $\bar{R}_1 < \bar{R}_2 < \dots < \bar{R}_{\gamma K}$ and compare SEL for $R_k^{est} = \text{rank}(\bar{R}_k) = k, k = 1, \dots, \gamma K$ to any other assignment. It is straightforward to show that switching any pair that does not follow the \bar{R}_k order reduces SEL. Iterating this and noting that the $\hat{R}_k = \text{rank}(\bar{R}_k)$ produces the result. \square

Appendix G Proof of Theorem 4

Proof. Recall that for a positive, discrete random variable the expected value can be computed as the sum of $(1 - \text{cdf})$ at mass points, where cdf is the cumulative distribution function, so

$$\begin{aligned} \bar{R}_k &= \sum_{\nu=1}^K \nu \text{pr}[R_k = \nu] = \sum_{\nu=1}^K \text{pr}[R_k \geq \nu] \\ &= \sum_{\nu=1}^K \frac{2(K - \nu + 1)}{K(K + 1)} \frac{K(K + 1)}{2(K - \nu + 1)} \text{pr}[R_k \geq \nu] \\ &= \sum_{\nu=1}^K \frac{2(K - \nu + 1)}{K(K + 1)} R_k^+(\nu). \end{aligned} \tag{16}$$

Relation (16) can be used to show that when the posterior distributions are stochastically ordered, $\hat{R}_k \equiv \tilde{R}_k(\gamma)$ because the order of $\text{pr}[R_k \geq \nu]$ does not depend on γ and the \bar{R}_k inherit their order. \square

Appendix H Proof of Theorem 5

Proof. In this proof we use \mathbf{Y}_K rather than \mathbf{Y} to stress that as K goes to infinity, the length of \mathbf{Y} changes. For $G_{\mathbf{Y}_K}(t) = \frac{1}{K} \sum_{k=1}^K \text{pr}(\theta_k \leq t | \mathbf{Y}_K)$, we prove: as $K \rightarrow \infty$,

$|\text{pr}(P_k \geq \gamma | \mathbf{Y}_{\mathbf{K}}) - \text{pr}(\theta_k \geq \bar{G}_{\mathbf{Y}_{\mathbf{K}}}^{-1}(\gamma) | \mathbf{Y}_{\mathbf{K}})| \rightarrow 0$, where P_k is the true percentile of θ_k , $\mathbf{Y}_{\mathbf{K}}$ is the vector (Y_1, Y_2, \dots, Y_K) .

The posterior independence of θ_k is straightforward. Denote $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k, \dots, \theta_K)$ and $\boldsymbol{\theta}^{(-k)} = (\theta_1, \dots, \theta_{k-1}, \theta_{k+1}, \dots, \theta_K)$, where $\theta_k \stackrel{\text{ind}}{\sim} g(\cdot | Y_k) = g_k(\cdot)$. Let $\theta_{(\gamma)} = \theta_{(i,K)}$ be the γ th quantile of $\boldsymbol{\theta}$, if $\frac{i}{K} \leq \gamma < \frac{i+1}{K}$, where $\theta_{(i,K)}$ is the i th largest number of $\boldsymbol{\theta}$. Respectively, $\theta_{(\gamma)}^{(-k)}$ is the γ th quantile of $\boldsymbol{\theta}^{(-k)}$. We also denoted $\theta_{(i-1, K-1)}$ as the $(i-1)$ th largest number of $\boldsymbol{\theta}^{(-k)}$.

For the $\tilde{P}_k(\gamma)$'s generator:

$$\begin{aligned} \text{pr}(P_k \geq \gamma | \mathbf{Y}_{\mathbf{K}}) &= \text{E}[I(\theta_k \geq \theta_{(\gamma)}) | \mathbf{Y}_{\mathbf{K}}] \\ &= \text{E}[I(\theta_k \geq \theta_{(\gamma)}^{(-k)}) | \mathbf{Y}_{\mathbf{K}}] + \text{E}[I(\theta_k \geq \theta_{(\gamma)}) - I(\theta_k \geq \theta_{(\gamma)}^{(-k)}) | \mathbf{Y}_{\mathbf{K}}] \end{aligned} \quad (17)$$

For the second term in (17)

$$\text{E}[I(\theta_k \geq \theta_{(\gamma)}) - I(\theta_k \geq \theta_{(\gamma)}^{(-k)}) | \mathbf{Y}_{\mathbf{K}}] = -\text{pr}(\theta_{(\gamma)}^{(-k)} \leq \theta_k < \theta_{(\gamma)} | \mathbf{Y}_{\mathbf{K}}) + \text{pr}(\theta_{(\gamma)} \leq \theta_k < \theta_{(\gamma)}^{(-k)} | \mathbf{Y}_{\mathbf{K}})$$

We have the inequality $\frac{i-1}{K-1} < \frac{i}{K} \leq \gamma < \frac{i+1}{K}$, $\theta_{(\gamma)} = \theta_{(i,K)}$ by definition. Consider the relation between $\frac{i}{K-1}$ and γ :

- If $\gamma < \frac{i}{K-1}$, then $\frac{i-1}{K-1} < \frac{i}{K} \leq \gamma < \frac{i}{K-1} < \frac{i+1}{K}$, $\theta_{(\gamma)}^{(-k)} = \theta_{(i-1, K-1)}$
 $\text{pr}(\theta_{(i-1, K-1)} \leq \theta_k < \theta_{(i, K)} | \mathbf{Y}_{\mathbf{K}}) = 0$ and
 $\text{pr}(\theta_{(i, K)} < \theta_k \leq \theta_{(i-1, K-1)} | \mathbf{Y}_{\mathbf{K}}) = 0$;
- If $\frac{i}{K-1} \leq \gamma$, then $\frac{i-1}{K-1} < \frac{i}{K} < \frac{i}{K-1} \leq \gamma < \frac{i+1}{K}$, $\theta_{(\gamma)}^{(-k)} = \theta_{(i, K-1)}$
 $\text{pr}(\theta_{(i, K-1)} < \theta_k \leq \theta_{(i, K)} | \mathbf{Y}_{\mathbf{K}}) = 0$ and
 $\text{pr}(\theta_{(i, K)} < \theta_k \leq \theta_{(i-1, K-1)} | \mathbf{Y}_{\mathbf{K}}) = 0$.

Thus the second term in (17) is zero,

$$\begin{aligned} \text{pr}(P_k \geq \gamma | \mathbf{Y}_{\mathbf{K}}) &= \text{E}[I(\theta_k \geq \theta_{(\gamma)}^{(-k)}) | \mathbf{Y}_{\mathbf{K}}] = \text{E}[\text{E}[I(\theta_k \geq \theta_{(\gamma)}^{(-k)}) | \boldsymbol{\theta}^{(-k)}] | \mathbf{Y}_{\mathbf{K}}] \\ &= \text{E}[\text{pr}(\theta_k \geq G^{-1}(\gamma) | \mathbf{Y}_{\mathbf{K}}) + g_k(G^{-1}(\gamma))(\theta_{(\gamma)}^{(-k)} - G^{-1}(\gamma)) + o_p(\theta_{(\gamma)}^{(-k)} - G^{-1}(\gamma)) | \mathbf{Y}_{\mathbf{K}}] \\ &= \text{pr}(\theta_k \geq G^{-1}(\gamma) | Y_k) + g_k(G^{-1}(\gamma))\text{E}[(\theta_{(\gamma)}^{(-k)} - G^{-1}(\gamma)) + o_p(\theta_{(\gamma)}^{(-k)} - G^{-1}(\gamma)) | \mathbf{Y}_{\mathbf{K}}] \end{aligned} \quad (18)$$

In (18), $\theta_{(\gamma)}^{(-k)}$ is the γ th quantile of non-iid $K-1$ samples from $K-1$ posterior distributions. By theorem 5.2.1 of David and Nagaraja (2003) and large sample theorem of order statistics from iid sampling, we have $\theta_{(\gamma)}^{(-k)} \rightarrow G^{-1}(\gamma)$ in probability as K goes to ∞ . Since we assume that $\theta_k | Y_k$ has a uniformly bounded finite second moment, so does $\theta_{(\gamma)}^{(-k)} | \mathbf{Y}_{\mathbf{K}}$. Thus $\text{E}[\theta_{(\gamma)}^{(-k)} | \mathbf{Y}_{\mathbf{K}}] \rightarrow G^{-1}(\gamma)$.

The generator of $P_k^*(\gamma)$ is:

$$\begin{aligned} \text{pr}(\theta_k \geq \bar{G}_{\mathbf{Y}_K}^{-1}(\gamma) | \mathbf{Y}_K) &= \text{pr}(\theta_k \geq G^{-1}(\gamma) | \mathbf{Y}_K) + g_k(\bar{G}^{-1}(\gamma))(\bar{G}_{\mathbf{Y}_K}^{-1}(\gamma) - G^{-1}(\gamma)) \\ &\quad + o(\bar{G}_{\mathbf{Y}_K}^{-1}(\gamma) - G^{-1}(\gamma)) \\ &= \text{pr}(\theta_k \geq G^{-1}(\gamma) | Y_k) + g_k(\bar{G}^{-1}(\gamma))(\bar{G}_{\mathbf{Y}_K}^{-1}(\gamma) - G^{-1}(\gamma)) \\ &\quad + o(\bar{G}_{\mathbf{Y}_K}^{-1}(\gamma) - G^{-1}(\gamma)) \end{aligned} \quad (19)$$

Since $\bar{G}_{\mathbf{Y}_K}^{-1}(\gamma) \rightarrow G^{-1}(\gamma)$, by (18) and (19), $|\text{pr}(P_k \geq \gamma | \mathbf{Y}_K) - \text{pr}(\theta_k \geq \bar{G}_{\mathbf{Y}_K}^{-1}(\gamma) | \mathbf{Y}_K)| \rightarrow 0$. \square

Appendix I Scoring function

For each function $S(P)$, there will be an optimal SEL ranking estimator. For instance,

$$S(P) = (aP + b) * \mathbf{I}_{\{P > \gamma\}}, a > 0$$

indicates that the reward or penalty is the same for all units below the threshold γ ; for units above γ the reward/penalty is linearly related to the rank.

We study the (above γ)/(below γ) classification, but more than two ordinal categories can be of interest. For example, educational institutions might be classified into three categories, the poor, the average and the excellent. The following two $S(P)$ capture this goal. Let $J \geq 3$ be the number of ordered categories, then

$$S(P) = \sum_{j=1}^J a_j I(P \leq \gamma_j), \quad 0 < \gamma_1 < \dots < \gamma_J, \quad a_j \geq 0$$

or

$$S(P) = \sum_{j=1}^J a_j I(P \leq \gamma_j) + a_0 P, \quad 0 < \gamma_1 < \dots < \gamma_J, \quad a_j \geq 0.$$

References

- Austin, P. C. and Tu, J. V. (2006). "Comparing Clinical Data with Administrative Data for Producing Acute Myocardial Infarction Report Cards." *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 169(1): 115–126.
- Benjamini, Y. and Hochberg, Y. (1995). "Controlling the false discovery rate: A practical and powerful approach to multiple testing." *Journal of the Royal Statistical Society, Series B, Methodological*, 57: 289–300.
- Christiansen, C. L. and Morris, C. N. (1997). "Improving the statistical approach to health care provider profiling." *Annals of Internal Medicine*, 127: 764–768.
- Conlon, E. M. and Louis, T. A. (1999). "Addressing Multiple Goals in Evaluating Region-specific Risk using Bayesian methods." In Lawson, A., Biggeri, A., Böhning,

- D., Lesaffre, E., Viel, J.-F., and Bertollini, R. (eds.), *Disease Mapping and Risk Assessment for Public Health*, chapter 3, 31–47. Wiley.
- Daniels, M. and Normand, S.-L. T. (2006). “Longitudinal profiling of health care units based on continuous and discrete patient outcomes.” *Biostatistics*, 7: 1–15.
- David, H. A. and Nagaraja, H. N. (2003). *Order Statistics*. Wiley, third edition.
- Devine, O. J. and Louis, T. A. (1994). “A constrained empirical Bayes estimator for incidence rates in areas with small populations.” *Statistics in Medicine*, 13: 1119–1133.
- Devine, O. J., Louis, T. A., and Halloran, M. E. (1994). “Empirical Bayes estimators for spatially correlated incidence rates.” *Environmetrics*, 5: 381–398.
- Diggle, P. J., Thomson, M. C., Christensen, O. F., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, J. H., Boussinesq, M., and Molyneux, D. H. (2006). “Spatial modelling and prediction of Loa loa risk: decision making under uncertainty.” Technical report, Department of Mathematics and Statistics, Lancaster University.
- Draper, D. and Gittoes, M. (2004). “Statistical Analysis of Performance Indicators in UK Higher Education.” *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 167(3): 449–474.
- DuMouchel, W. (1999). “Bayesian Data Mining in Large Frequency Tables, With An Application to the FDA Spontaneous Reporting System (with discussion).” *The American Statistician*, 53: 177–190.
- Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. (2001). “Empirical Bayes analysis of a microarray experiment.” *Journal of the American Statistical Association*, 96(456): 1151–1160.
- End-Stage Renal Disease (ESRD) Network (2000). “1999 Annual Report: ESRD Clinical Performance Measures Project.” Technical report, Health Care Financing Administration.
- Gelman, A. and Price, P. N. (1999). “All maps of parameter estimates are misleading.” *Statistics in Medicine*, 18: 3221–3234.
- Goldstein, H. and Spiegelhalter, D. J. (1996). “League tables and their limitations: statistical issues in comparisons of institutional performance (with discussion).” *Journal of the Royal Statistical Society Series A*, 159: 385–443.
- Hardy, G. H., Littlewood, J. E., and Polya, G. (1967). *Inequalities*. Cambridge University Press, 2nd edition.
- Lacson, E., Teng, M., Lazarus, J. M., Lew, N., Lowrie, E. G., and Owen, W. F. (2001). “Limitations of the facility-specific standardized mortality ratio for profiling health care quality in Dialysis.” *American Journal of Kidney Diseases*, 37: 267–275.

- Laird, N. M. and Louis, T. A. (1989). “Empirical Bayes ranking methods.” *Journal of Educational Statistics*, 14: 29–46.
- Landrum, M. B., Bronskill, S. E., and Normand, S.-L. T. (2000). “Analytic methods for constructing cross-sectional profiles of health care providers.” *Health Services and Outcomes Research Methodology*, 1: 23–48.
- Landrum, M. B., Normand, S.-L. T., and Rosenheck, R. A. (2003). “Selection of Related Multivariate Means: Monitoring Psychiatric Care in the Department of Veterans Affairs.” *Journal of the American Statistical Association*, 98(461): 7–16.
- Lin, R., Louis, T. A., Paddock, S. M., and Ridgeway, G. (2004). “Ranking of USRDS, provider-specific SMRs from 1998-2001.” Technical Report 67, Johns Hopkins University, Dept. of Biostatistics Working Papers, <http://www.bepress.com/jhubiostat/paper67>.
- Liu, J., Louis, T. A., Pan, W., Ma, J., and Collins, A. (2004). “Methods for estimating and interpreting provider-specific, standardized mortality ratios.” *Health Services and Outcomes Research Methodology*, 4: 135–149.
- Lockwood, J. R., Louis, T. A., and McCaffrey, D. F. (2002). “Uncertainty in rank estimation: Implications for value-added modeling accountability systems.” *Journal of Educational and Behavioral Statistics*, 27(3): 255–270.
- McCaffrey, D. F., Lockwood, J. R., Koretz, D., Louis, T. A., and Hamilton, L. (2004). “Models for value-added modeling of teacher effects.” *Journal of Educational and Behavioral Statistics*, 29(1): 67–101.
- McClellan, M. and Staiger, D. (1999). “The Quality of Health Care Providers.” Technical Report 7327, National Bureau of Economic Research, Working Paper.
- Noell, G. H. and Burns, J. L. (2006). “Value-added assessment of teacher preparation - An illustration of emerging technology.” *Journal of Teacher Education*, 57(1): 37–50.
- Normand, S.-L. T., Glickman, M. E., and Gatsonis, C. A. (1997). “Statistical methods for profiling providers of medical care: Issues and applications.” *Journal of the American Statistical Association*, 92: 803–814.
- Paddock, S. M., Ridgeway, G., Lin, R., and Louis, T. A. (2006). “Flexible distributions for triple-goal estimates in two-stage hierarchical models.” *Computational Statistics & Data Analysis*, 50/11: 3243–3262.
- Rubin, D. B., Stuart, E. A., and Zanutto, E. L. (2004). “A potential outcomes view of value-added assessment in education.” *Journal of Educational and Behavioral Statistics*, 29(1): 103–116.
- Shen, W. and Louis, T. A. (1998). “Triple-goal estimates in two-stage, hierarchical models.” *Journal of the Royal Statistical Society, Series B*, 60: 455–471.

- Storey, J. D. (2002). “A direct approach to false discovery rates.” *Journal of the Royal Statistical Society, Series B, Methodological*, 64(3): 479–498.
- (2003). “The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value.” *The Annals of Statistics*, 31(6): 2013–2035.
- Tekwe, C. D., Carter, R. L., Ma, C. X., Algina, J., Lucas, M. E., Roth, J., Ariet, M., Fisher, T., and Resnick, M. B. (2004). “An empirical comparison of statistical models for value-added assessment of school performance.” *Journal of Educational and Behavioral Statistics*, 29(1): 11–35.
- Tusher, V. G., Tibshirani, R., and Chu, G. (2001). “Significance analysis of microarrays applied to the ionizing radiation response.” *Proceedings of National Academy of Sciences*, 98(9): 5116–5121.
- United States Renal Data System (USRDS) (2005). “2005 Annual Data Report: Atlas of end-stage renal disease in the United States.” Technical report, Health Care Financing Administration.
- Wolfe, R., Gaylin, D., Port, F., Held, P., and Wood, C. (1992). “Using USRDS generated mortality tables to compare local ESRD mortality rates to national rates.” *Kidney Int*, 42(4): 991–6.
- Wright, D. L., Stern, H. S., and Cressie, N. (2003). “Loss functions for estimation of extrema with an application to disease mapping.” *The Canadian Journal of Statistics*, 31(3): 251–266.

About the Authors

Rongheng Lin is Research Fellow, Biostatistics Branch, NIH, National Institute of Environmental Health Science, 111 T.W. Alexander Drive, Research Triangle Park, NC 27709, U.S.A. (E-mail:linr2@niehs.nih.gov), to whom correspondences should be addressed.

Thomas A. Louis is Professor, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, U.S.A.

Susan M. Paddock and Greg Ridgeway are both Full Statisticians, RAND Corporation, Santa Monica, CA 90401, U.S.A.

Acknowledgments

This work was supported by grant 1-R01-DK61662 from the U.S. NIH, National Institute of Diabetes, Digestive and Kidney Diseases. The authors are grateful to the editors and referees for helpful comments.

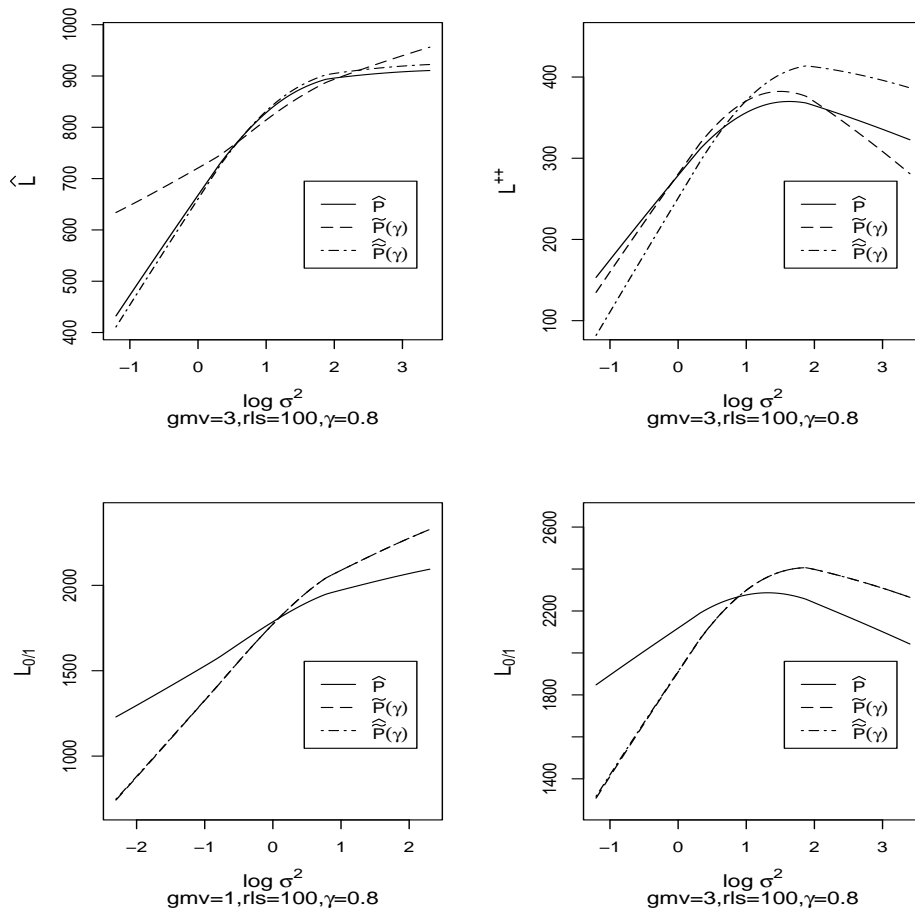


Figure 2: Loess smoothed, unit-specific performance of \hat{P}_k , $\tilde{P}_k(\gamma)$ and $\hat{P}_k(\gamma)$ under \hat{L} , L^{++} , and $L_{0/1}$ as a function of unit-specific variance (σ_k^2).

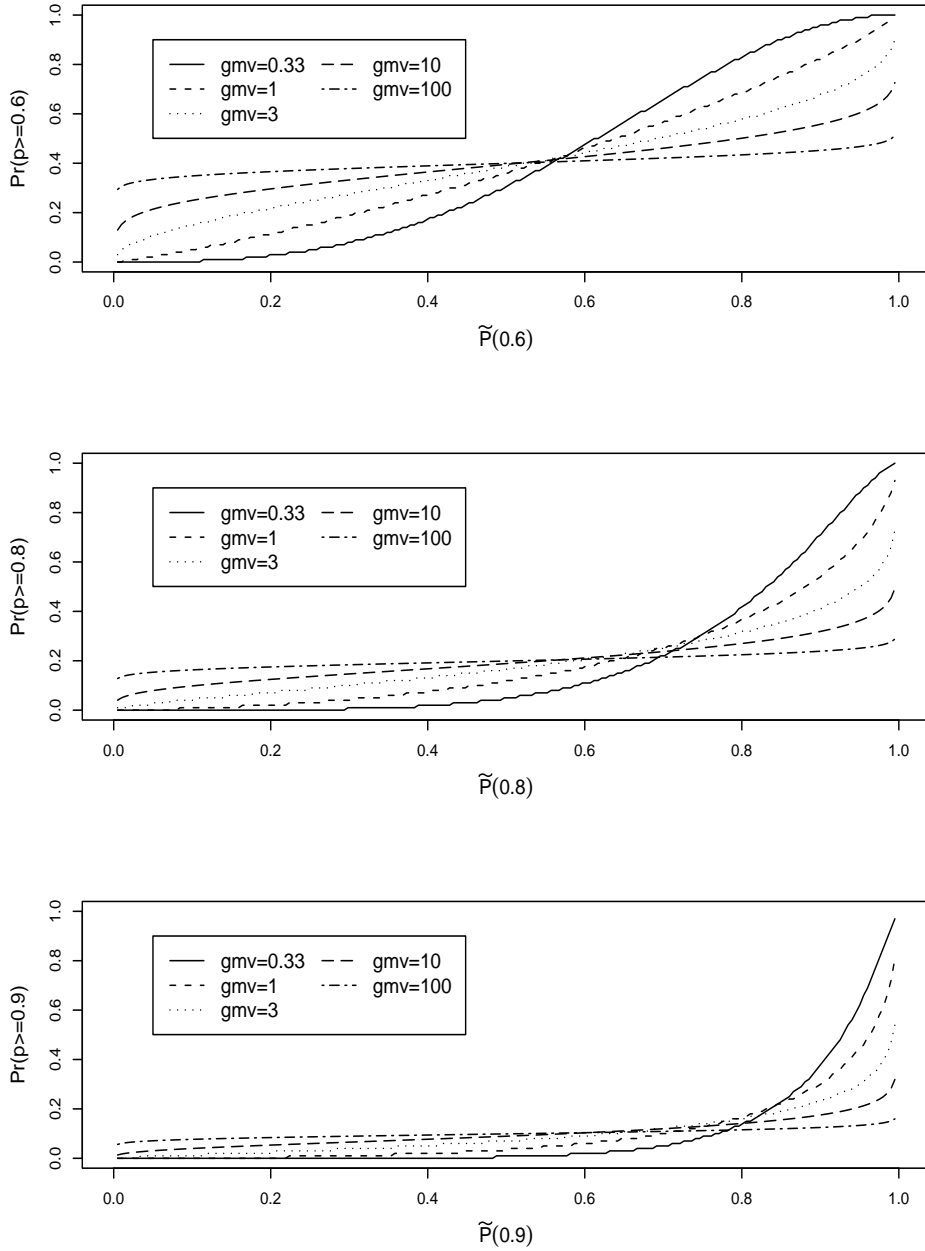


Figure 3: Average posterior classification probabilities as a function of the optimally estimated percentiles for $rls = 1$, $\gamma = (0.6, 0.8, 0.9)$.

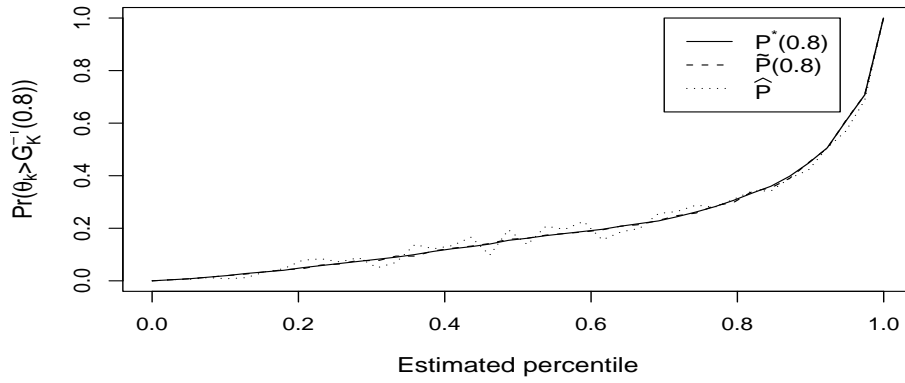


Figure 4: $\Pr(\theta_k > 0.18 \mid \mathbf{Y})$ with X-axis percentiles determined separately by the three percentiling methods.

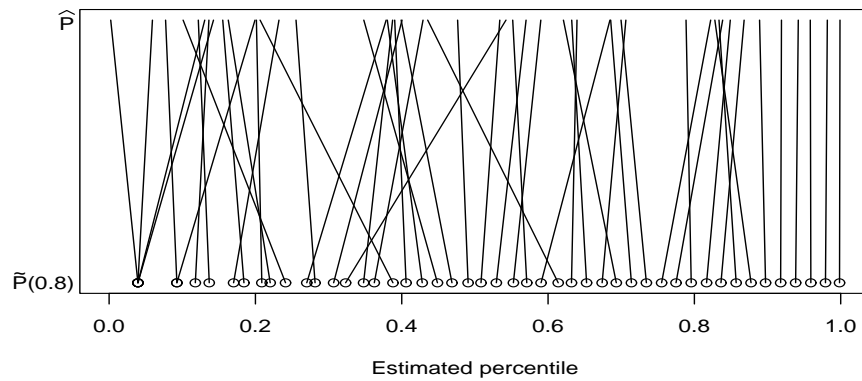


Figure 5: Circles represent 50 USRDS dialysis centers evenly spread across percentiles determined by $\tilde{P}_k(0.8)$ using 1998 SMR data. Lines connect $\tilde{P}_k(0.8)$ and \hat{P}_k . Ties appear in the lower percentile area.