Loss Function Based Ranking in Two-Stage, Hierarchical Models

Rongheng Lin, Thomas A. Louis, Susan M. Paddock, Greg Ridgeway[1]

April 17, 2005

[1]Rongheng Lin is Ph.D candidate, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, 615 N. Wolfe St., Baltimore MD 21205, U.S.A. (E-mail:rlin@jhsph.edu), 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 Statistician, RAND, Santa Monica, CA 90401, U.S.A. Research was supported by grant 1-R01-DK61662 from U.S. NIH National Institute of Diabetes, Digestive and Kidney Diseases.

Abstract

Performance evaluations of health services providers burgeons. Similarly, analyzing spatially structured 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 "exceedences" (how many and which unit-specific true parameters exceed a threshold). The data structure and inferential goals require a hierarchical, Bayesian model that accounts for nesting relations and specifies both population values and random effects for unit-specific parameters. Guided by a loss function, the approach structures non-standard inferences such as ranking.

Several authors have studied the performance of optimal ranks (based on posterior mean of the rank) based on Squared Error Loss (SEL) . These posterior means of the ranks are effective in many applications, but other loss functions and optimal ranks might be more appropriate, for example when the goal is to identify the relatively good (e.g., in the upper 10%) or relatively poor performers. We construct loss functions that address this and other goals, using mathematical analysis and simulation, compare these to SEL-based ranks and other candidate approaches. We develop relations among candidates and study performance for a fully parametric hierarchical model with a Gaussian prior and Gaussian sampling distributions. We illustrate estimates and performance assessments by analyzing 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 may not be preferred when classifying units above or below a percentile cutpoint. Even optimal rank estimates can perform quite poorly in most real-world settings and data-analytic performance indicators should always be reported.

KEY WORDS: Ranking/Percentiling, Bayesian Models, Decision Theory, Operating Characteristic

1 Introduction

The prevalence of performance evaluations of health services providers using ranks or percentiles (Goldstein and Spiegelhalter 1996; Christiansen and Morris 1997; McClellan and Staiger 1999; Landrum et al. 2000; Normand et al. 1997), using post-marketing information to evaluate drug side-effects (DuMouchel 1999), ranking geographic regions using disease incidence, (Devine and Louis 1994; Devine et al. 1994; Conlon and Louis 1999) and ranking teachers and schools (value added modeling (Lockwood et al. 2002) burgeons. 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. These latter include 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 are very effective in structuring these assessments. Inferences depend on the posterior distribution, and how the posterior is used should depend on inferential goals. Gelman and Price (1999) show that no single set of estimates can effectively address the foregoing multiple goals. For example, the histogram of maximum likelihood estimates (MLEs) of unit-specific parameters is always over-dispersed relative to the histogram of the true random effects and the histogram of Bayesian posterior means is always under-dispersed. However, as Shen and Louis (1998) show, a loss function directly addressing the histogram produces effective estimates. They structure estimating ranks by squared error loss (SEL) operating on the ranks and show that SEL-optimal ranks outperform competitors such as MLE-based ranks. Lockwood et al. (2002) present simulation-based evaluations of these ranks; Liu et al. (2004) provide additional evaluations.

SEL applies the same distance penalty to all pairs of estimated and true ranks, but in many applications interest focuses on identifying the relatively good (e.g., in the upper 10% of the target parameter distribution) or relatively poor performers. For example, quality improvement initiatives should be targeted at health care providers that truly perform poorly; environmental assessments should be targeted at the truly high incidence locations (see Wright et al. (2003));

genes with the truly highest differential expression should be selected for further study. To this end, we construct (above γ)/(below γ) loss functions that penalize for misclassification and, in some instances an additional estimation loss. We derive optimal ranks, evaluate and compare pre-posterior performance and develop posterior distribution computed, data analytic, performance indicators including a new operating characteristic for (above γ)/(below γ) classification.

We evaluate performance for a fully parametric hierarchical model with a Gaussian sampling distribution (constant or unit-specific sampling variance) and either a Gaussian prior or a mixture of two Gaussians. To study optimal performance and robustness to assumptions, we evaluate under the prior and loss function used to generate the ranks and other priors and loss functions. We illustrate approaches using Standardized Mortality Ratio (SMR) data from the from United States Renal Data System (USRDS).

Results show that though SEL-optimal ranks and percentiles do not specifically focus on classifying with respect to a percentile cut point, they perform very well over a broad range of loss functions. However, optimal (above γ)/(below γ) ranks are superior for classification loss and 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.

Even optimal rank estimates can perform quite poorly in most real-world settings and data-analytic performance indicators should always be reported.

2 The two-stage, Bayesian hierarchical model

We consider a two-stage model with iid sampling from a known prior G with density g and possibly different sampling distributions f_k:

$$\begin{array}{rcl} \theta_1, \ldots, \theta_K & iid & G(\theta_k), k = 1, \ldots, K \\ Y_k \mid \theta_k & \sim & f_k(Y_k \mid \theta_k) \\ \theta_k \mid Y_k & ind & g_k(\theta_k \mid Y_k) = \dfrac{f_k(Y_k \mid \theta_k) g(\theta_k)}{\int f_k(Y_k \mid u) g(u) du}. \end{array} \tag{1}$$

The model can be generalized to allow a regression structure in the prior and can be extended to three stages with a hyper-prior to structure "Bayes empirical Bayes" (Conlon and Louis 1999;

Louis and Shen 1999; Carlin and Louis 2000).

2.1 Loss functions and decisions

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

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

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:

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

3 Ranking

In general, the optimal ranks for the θ_k are neither the ranks of the observed data nor the ranks of their posterior means. Laird and Louis (1989) structure estimation by representing the ranks as,

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

with the smallest θ having rank 1.

3.1 Squared-error loss (SEL)

Square error loss is the most common estimation loss function, optimized by the posterior mean of the target parameter. For example, with the unit-specific θs as the target, the loss function is $L(\theta, a) = (\theta - a)^2$ and $\theta_k^{pm} = E(\theta_k \mid \mathbf{Y})$. When ranks are the target, to produce SEL-optimal ranks, minimize

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

by setting R_k^{est} equal to,

$$\bar{R}_k(\mathbf{Y}) = E_G[R_k(\boldsymbol{\theta}) \mid \mathbf{Y}] = \sum_{j=1}^{K} P_G[\theta_k \geq \theta_j \mid \mathbf{Y}]. \tag{4}$$

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

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

See Section A.1 for additional details on estimating optimal ranks under weighted SEL (WSEL).

In the subsequent sections, generally we drop dependency on $\boldsymbol{\theta}$ and omit conditioning on $\mathbf{Y}$. 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

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

normalize large sample performance and aid in applied communication. For example, Lockwood et al. (2002) show that mean square error (MSE) for percentiles rapidly converges to a function that does not depend on K; similar normalization applies to 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 correct classification and, possibly, a distance penalty (identification of the lower 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

To structure loss functions, for $0 < \gamma < 1$, let

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

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,

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

The loss functions confer no penalty if an estimated and true unit-specific percentile pair (P_k^{est}, P_k) are either both above or both below the γ cut point. If they are on different sides of γ, the loss functions penalize by an amount that depends on the distance of either the estimated percentile (in $\tilde{L}$) or the true percentile (in $L^\dagger$) from γ or on the distance between the true and estimated percentiles (in $L^\ddagger$). The parameters p and q adjust the intensity of the two penalties; $p \neq q$ and $c \neq 1$ allow for differential penalties for the two kinds of misclassification. $L_{0/1}(\gamma)$ counts the number of discrepancies and is equivalent to setting $p = q = 0$, $c = 1$.

Our mathematical analyses apply to the loss functions (8), but our simulations are conducted for $p = q = 2, c = 1$. In this case, the foregoing simplify to:

$$\begin{aligned}
\tilde{L}(\gamma) &= \frac{1}{K}\sum_k (\gamma - P_k^{est})^2 \{AB_k(\gamma, P_k, P_k^{est}) + BA_k(\gamma, P_k, P_k^{est})\} \\
L^\dagger(\gamma) &= \frac{1}{K}\sum_k (P_k - \gamma)^2 \{AB_k(\gamma, P_k, P_k^{est}) + BA_k(\gamma, P_k, P_k^{est})\} \qquad (9) \\
L^\ddagger(\gamma) &= \frac{1}{K}\sum_k (P_k - P_k^{est})^2 \{AB_k(\gamma, P_k, P_k^{est}) + BA_k(\gamma, P_k, P_k^{est})\}.
\end{aligned}$$

We do not investigate the "all or nothing" loss function that confers no penalty if all units are correctly classified (above γ)/(below γ) and penalty 1 if any unit is misclassified. Finding the optimal classification is challenging and there will be many nearly optimal solutions. Furthermore, loss functions that compute average performance over units are more appropriate in most applications.

5 Optimizers and other candidate estimators

5.1 Loss function optimizers

The $\hat{P}_k$ minimize $\hat{L}$ (see Equation 5) and when the posterior distributions are stochastically ordered (the $G_k(t \mid Y_k)$ never cross), they are optimal for broad class of loss functions (see Theorem 4). Below, 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. We note that it is straightforward to show that 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.

5.1.1 Optimizing $L_{0/1}$

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

$$\begin{aligned} \tilde{R}_k(\gamma) &= \text{rank}\{\text{pr}(P_k \geq \gamma|\mathbf{Y})\} \\ \tilde{P}_k(\gamma) &= \tilde{R}_k(\gamma)/(K+1) \end{aligned} \tag{10}$$

These are not unique optimizers.

Proof. Rewrite the loss function as a function of the number of observations 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:

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

To optimize $L_{0/1}$, for each θ_k calculate $\mathsf{pr}(P_k > \gamma|\mathbf{Y})$, 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 $L_{0/1}$, but there are other optimizers because $L_{0/1}$ does not require optimal ordering, only optimal (above γ)/(below γ) categorization. For example, permutations of the ranks of observations in A or permutations of the ranks in A^C yield the same posterior risk for $L_{0/1}$. Other loss functions provide additional structure and resolution. □

The $\tilde{P}_k(\gamma)$ are similar to percentiles proposed by Normand et al. (1997), who suggest using the posterior probability $\text{pr}(\theta_k > t|\mathbf{Y})$ to compare the performance of medical care providers, where t is a relevant threshold. We relate their approach to the $\tilde{P}_k(\gamma)$ via $P_k^*(\gamma)$:

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

$$\bar{G}_{\mathbf{Y}}(t) = \frac{1}{K}\sum_{k=1}^{K} \text{pr}(\theta_k \leq t|\mathbf{Y}), \tag{11}$$

and define $P_k^*(\gamma)$ as the percentiles induced 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 $P_k^*(\gamma)$, therefore approximately loss function based and monotone transform invariant. By making a direct link to the original θ scale, $P_k^*(\gamma)$ is easier to explain and interpret. Furthermore, for a desired accuracy, computing $P_k^*(\gamma)$ is substantially faster $\tilde{P}_k(\gamma)$, since the former requires only accurate computation of individual posterior distributions and of $\bar{G}_{\mathbf{Y}}$.

5.1.2 Optimizing $\tilde{L}$

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

Proof. In Section A.2, we show 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 $|\gamma - P_k^{est}|$. □

5.1.3 Optimizing $L^\dagger$

Section A.3 presents an optimization procedure for the case $p = q = 2, -\infty < c < \infty$, but other than 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.

5.1.4 Optimizing $L^\ddagger$

We have not found a simple means of optimizing this loss function, but Section A.4 develops a helpful relation. An alternative to $L^\ddagger$ is the convex combination loss, $\hat{L}_{0/1}^w(\gamma) = (1-w)L_{0/1}(\gamma) + w\hat{L}, (0 \leq w \leq 1)$. It motivates $\hat{\tilde{P}}_k(\gamma)$, which we will show is very effective though not optimal under $L^\ddagger$.

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

Theorem 3. The $\hat{\tilde{P}}_k(\gamma)$ minimize $L_{0/1}$ and conditional on this (above γ)/(below γ) grouping, produce optimal constrained SEL estimates.

Proof. Since the (above γ)/(below γ) groups are formed by $\tilde{P}_k(\gamma)$, $\hat{\tilde{P}}_k(\gamma)$ 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 $(1, \ldots, \gamma K)$ are in the (below γ) group and $(\gamma K + 1, \ldots, K)$ are in the (above γ) group. Similar to Section A.1,

$$\mathsf{E} \sum_k \left(R_k^{est} - R_k\right)^2 \;=\; \sum_k V(R_k) + \sum_k \left(R_k^{est} - \overline{R}_k\right)^2 .$$

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

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

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

Consequence: For the convex combination loss function $\hat{L}_{0/1}^w(\gamma)$, it is straightforward to show that there exists a $w_* > 0$ such that for all $w \leq w_*$, $\hat{\tilde{P}}_k(\gamma)$ is optimal (motivating it as a candidate estimator). Similarly, there exists a $w^* < 1$ such that for all $w \geq w^*$, $\hat{P}_k$ is optimal.

5.2 Other ranking estimators

Traditional rank estimators include ranks based on maximum likelihood estimates (MLE), posterior means, and 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 reward or punish the units with large variance while hypothesis test-based ranks will reward or punish unfairly the units with small variance. Modified hypothesis test statistics moderate this shortcoming by reducing the ratio of variances (Tusher et al. (2001); Efron et al. (2001)).

6 Relations among estimators

In this section, we relate $\bar{P}_k, P_k^*$ and $\tilde{P}_k(\gamma)$. Proofs are generally in terms fo the ranks (R).

6.1 A general relation

Let $\nu = [\gamma K]$ and define

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

The rank$[R_k^+(\nu)]$ = rank[pr$(R_k \geq \nu)$] and so each generate 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:

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

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

Proof. Recall that for a positive, discrete random variable the expected value can be computed as the sum of (1 - cdf), so

$$\begin{aligned}
\overline{R}_k &= \sum_{\nu=1}^{K} \nu \mathsf{pr}[R_k = \nu] = \sum_{\nu=1}^{K} \mathsf{pr}[R_k \geq \nu] \qquad (12)\\
&= \sum_{\nu=1}^{K} \frac{2(K-\nu+1)}{K(K+1)} \frac{K(K+1)}{2(K-\nu+1)} \mathsf{pr}[R_k \geq \nu]\\
&= \sum_{\nu=1}^{K} \frac{2(K-\nu+1)}{K(K+1)} R_k^+(\nu)
\end{aligned}$$

□

Relation (12) 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 pr$[R_k \geq \nu]$ does not depend on γ and the $\overline{R}_k$ inherit their order.

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

From (2), (4) and (11), we have that,

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

The $R_k^*(\gamma)$ are generated by ranking the $\mathsf{pr}(\theta_k \geq \bar{G}_{\mathbf{Y}}^{-1}(\gamma))$ which is equivalent to ranking $\mathsf{pr}(\bar{G}_{\mathbf{Y}}(\theta_k) \geq \gamma)$. By the foregoing is equivalent to ranking the $\mathsf{pr}(\mathsf{E}[R_k|\theta_k] \geq \gamma K)$. 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^*.

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

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

Proof. See section A.5. □

Theorem 6. If $\theta_k \overset{iid}{\sim} G$, $Y_k|\theta_k \overset{ind}{\sim} f(Y_k \mid \theta_k, \zeta_k), \zeta_k$ observed with an edf that converges to a probability distribution and the posterior cdf of each θ_k is continuous and differentiable at $G^{-1}(\gamma)$. If $G_k(\cdot|\mathbf{Y})$ has a universally bounded finite second moment , then for $K \to \infty$, $P_k^*(\gamma)$ is equivalent to $\tilde{P}_k(\gamma)$

Proof: 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.

7 Posterior and pre-posterior performance evaluations

With Bayesian structuring, any estimator's performance can be evaluated using the posterior distribution (producing data analytic evaluations) and mixing over the marginal distribution of the data produces pre-posterior performance. These are valid, subject to model validity. For example, SEL performance can be computed by summing posterior variance and bias (this latter

is 0 for the posterior mean). Similarly, performance for other loss functions can be evaluated and estimators that are optimal for one loss function can be evaluated under other loss functions.

For (above γ)/(below γ) classification, the following posterior operating characteristic (OC) provides a informative evaluation. For any percentiling method, define,

$$\begin{aligned} OC(\gamma) &= \mathsf{pr}(P_k < \gamma | P_k^{est} > \gamma, \mathbf{Y}) + \mathsf{pr}(P_k > \gamma | P_k^{est} < \gamma, \mathbf{Y}) \\ &= \mathsf{pr}(P_k > \gamma | P_k^{est} < \gamma, \mathbf{Y})/\gamma = \frac{\mathsf{E}_{\theta|\mathbf{Y}} L_{0/1}(\gamma)}{2\gamma(1-\gamma)}, \end{aligned} \tag{13}$$

with the last equality following from $\sum_k BA_k(\gamma, P_k, P_k^{est}) = \sum_k AB_k(\gamma, P_k, P_k^{est})$. $OC(\gamma)$ 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., $\mathsf{pr}(P_k > \gamma | P_k^{est} < \gamma, \mathbf{Y})$ is similar to computing the false discovery rate (Benjamini and Hochberg (1995); Storey (2002, 2003)).

8 Simulation Scenarios and Results

8.1 Scenarios

We evaluate pre-posterior performance for $K = 200$ using $nrep = 2000$ simulation replications. We compute the $\mathsf{pr}(R_k = \ell \mid \mathbf{Y})$ using an independent sample Monte Carlo with $n = 2000$ draws. All simulations are for loss functions with $p = q = 2;\ c = 1$. Reported simulations are for the Gaussian sampling distribution. We have conducted similar investigations for the Poisson sampling distribution and results are very similar.

8.1.1 The Gaussian-Gaussian model

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

$$
\begin{aligned}
\theta_k &\overset{iid}{} N(0,\,1) \\
Y_k|\theta_k &\sim N(\theta_k, \sigma_k^2) \\
g_k(\theta_k \mid Y_k) &\sim N\left(\theta_k^{pm},\, (1-B_k)\sigma_k^2\right) \\
\theta_k^{pm} &= (1-B_k)Y_k \\
B_k &= \sigma_k^2/(\sigma_k^2+1).
\end{aligned}
$$

The $\{\sigma_k^2\}$ form an ordered, geometric sequence with geometric mean $gmv = GM(\sigma_1^2, \ldots, \sigma_K^2)$ and ratio of the largest σ^2 to the smallest $rls = \sigma_K^2/\sigma_1^2$.

8.1.2 Mixture model

The θ_k come from a two-component, mean 0 and variance 1 mixture of Gaussian distributions:

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

where

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

We compute the preposterior risk both assuming this mixture model and assuming that the the θ_k are from N(0,1). We present results for $\epsilon = 0.1, \Delta = 3.40, \xi^2 = .25, \gamma = 0.9$.

8.2 Results

Table 1 documents SEL $(\hat{L})$ performance for $\hat{P}_k$, the optimal estimator, for percentiled Y_k, 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}), this latter 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 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} degrades. Results show that though the posterior mean can perform well for some models and target parameters, in general it is not competitive.

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 $gmv = 1$.

Table 2 reports results for $\hat{P}_k$, $\tilde{P}_k(\gamma)$ and $\hat{\tilde{P}}_k(\gamma)$ under four loss functions. When $rls = 1$, $\tilde{P}_k(\gamma) \equiv \hat{P}_k \equiv \hat{\tilde{P}}_k(\gamma)$ 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 $\hat{L}$, $\hat{P}$ outperforms $\tilde{P}_k(\gamma)$ and $\hat{\tilde{P}}_k(\gamma)$ 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 $\hat{\tilde{P}}_k(\gamma)$ optimizes $\hat{L}^w$ 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 $\hat{\tilde{P}}_k(\gamma)$ 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 fnc$(gmv)\gamma(1-\gamma)$. By identity (13), this translates to 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 $\sigma_k^2 \to \infty$, $gmv = \infty$ and the Y_k provide no information on the θs or on the 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 $L_{0/1}$ percentile estimates $\hat{P}$ and $\tilde{P}$ perform similarly with many of the tabulated values are identical up to rounding. That $\tilde{P}_k(\gamma)$ is only moderately better than $\hat{P}_k$ under $\tilde{L}$ is due, in part

		$L_{0/1}$		$\hat{L}$			$\tilde{L}$			$L^{\ddagger}$		
γ	rls	$\hat{P}$	$\tilde{P}$	$\hat{P}$	$\tilde{P}$	$\hat{\tilde{P}}$	$\hat{P}$	$\tilde{P}$	$\hat{\tilde{P}}$	$\hat{P}$	$\tilde{P}$	$\hat{\tilde{P}}$
0.5	1	25	25	517	518	517	104	105	104	336	337	336
0.5	25	25	25	519	524	519	98	96	98	340	335	340
0.5	100	25	25	521	530	521	93	90	93	342	334	342
0.6	100	24	24	522	537	523	91	87	91	324	316	327
0.8	25	17	17	517	558	517	67	59	67	175	170	181
0.8	100	17	17	523	595	523	71	57	71	178	170	189
0.9	1	11	11	515	520	515	30	30	30	73	73	73
0.9	25	11	10	518	609	519	37	29	37	75	72	81
0.9	100	10	10	523	673	523	43	29	42	77	70	86
0.8	1	5	5	565	567	565	54	54	54	150	150	150
0.8	25	10	10	566	606	567	59	51	59	161	158	168
0.9	1	12	12	565	575	565	14	14	14	42	42	42
0.9	25	21	21	564	699	564	23	14	23	51	48	58
0.8	1	5	5	565	566	565	54	54	54	150	150	150
0.8	25	10	10	567	615	567	58	52	59	159	158	168
0.9	1	12	12	565	570	565	14	14	14	42	42	42
0.9	25	21	21	565	664	564	22	14	23	50	45	58

Table 2: Simulated preposterior risk for $gmv = 1$. $L_{0/1}$ is the percent; other values are $10000\times$(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	$L_{0/1}$ $200\gamma(1-\gamma)$	$\hat{L}$ 1667	$L^{\ddagger}$ $3333\gamma(1-\gamma)$	$\tilde{L}$ and $L^{\dagger}$ $3333\gamma(1-\gamma)[\gamma^3+(1-\gamma)^3]$
0.5	50	1667	833	208
0.6	48	1667	800	224
0.8	32	1667	533	277
0.9	18	1667	300	219

Table 3: Preposterior risk for $rls = 1$ when $gmv = \infty$. $L_{0/1}$ is the percentage; other values are $10000\times$Risk.

to our having considered only the case $p = q = 2, c = 1$ which makes $\tilde{L}$ very similar to $\widehat{L}$. 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 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.

8.2.1 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 $\hat{\tilde{P}}_k(\gamma)$ 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 $L_{0/1}$ ($gmv = 1$) 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 ($rls >> 1$), estimates for the high variance units perform better than for those with mid-level variance. This relation is due to improved classification into the (above γ)/(below γ) groups, with the improvement due to the substantial shrinkage towards 0.5 of the percentile estimates for high-variance units. 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 $rls \to \infty$. 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 $L_{0/1}$.

8.2.2 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 $\gamma = 0.8$ panel is typical. Discrimination improves with decreasing gmv, but even when $gmv = 0.33$ (the σ_k are $1/3$ of the prior variance), for a unit with $\tilde{P}_k(0.8) = 0.8$, the model-based, posterior probability that $P_k > 0.8$ is only 0.42. 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 $gmv \to \infty$ the plots converge to a horizontal line at $(1 - \gamma)$ and that as $gmv \to 0$ the plots converge to a step function that jumps from 0 to 1 at γ.

9 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), ESRD (2000), USRDS (2003)). Under the Poisson likelihood (see the last equation of (14)), the MLE is $\hat{\rho}_k = Y_k/m_k$, with Y_k the observed and m_k the expected deaths computed from a case-mix adjustment (Wolfe et al. (1992)). For the "typical" dialysis center $\rho_k \approx 1$. The expecteds control variability and there is a considerable range of m_k from 0 to greater than 100.

Liu et al. (2004) analyzed 1998 data; Lin et al. (2004) extend these analyses to $1998 - 2001$ for 3173 centers with complete data using an autoregressive model; we illustrate using 1998 data and the model

$$\begin{aligned} \xi \overset{iid}{\sim} N(0, 10), \qquad & \lambda = \tau^{-2} \overset{iid}{\sim} \text{Gamma}(0.05, 0.2) \\ [\theta_1, \ldots, \theta_K \mid \xi, \tau] \quad \overset{iid}{\sim} \quad & N(\xi, \tau^2),\ \theta_k = \log(\rho_k) \\ [Y_k \mid m_k, \rho_k] \quad \sim \quad & \text{Poisson}(m_k \rho_k). \end{aligned} \tag{14}$$

For these data, $\bar{G}_K^{-1}(0.8) = 0.18$ ($\rho = 1.20$). Figure 4 displays 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 $P_k^*(\gamma)$ 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 and for $\hat{P}_k$ is 0.65, showing that use of $\hat{P}_k$ is nearly fully efficient and that the optimal classification produces a 64% error rate for the $\gamma = 0.8$ threshold. Also, Figure 4 shows that for centers classified in the top 10%, the probability of that they are truly in the top 20% can be as low as 0.5.

Figure 5 displays the relation between $\tilde{P}_k(0.8)$ and $\hat{P}_k$ for 50 dialysis centers spaced uniformly according to $\tilde{P}_k(0.8)$. 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%.

10 Discussion

Table 1 shows that percentiles based on the Y_k or on posterior means of the target parameter can perform poorly. Similarly, hypothesis test-based percentiles perform poorly. Effective approaches should be loss function based or approximately so and should be invariant to monotone transforms of the target parameter. Basing inferences on the posterior distribution of (2) ensures this invariance, but procedures such as $P_k^*(\gamma)$ are also invariant.

The $\hat{P}_k$ that optimize $\hat{L}$ (SEL) are "general purpose" with no explicit attention to optimizing the (above γ)/(below γ) classification. When posterior distributions are not stochastically ordered (i.e., when choice of loss function matters), our simulations show that though $L_{0/1}(\gamma)$, $\tilde{P}_k(\gamma)$ 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. Similarly, $\tilde{P}_k(\gamma)$ can perform poorly with respect to SEL. In some scenarios relative benefits are considerable and so a choice must be made guided by goals.

Our performance evaluations are for the fully parametric model with a Gaussian sampling distribution, though we do investigate departures from the Gaussian prior. Very similar comparisons to those we report apply to the Poisson sampling distribution. Evaluation of performance for three-level models with a hyper-prior and for robust priors based on the non-parametric maximum likelihood prior or a fully Bayesian approach Paddock et al. (2005) show that alternative prior choices perform well across a variety of scenarios for SEL-optimal ranks.

Though our estimates depend on the joint distribution of ranks, we provide only univariate summaries of their properties and performance. These can be generalized to joint properties (e.g., pair-wise, posterior distributions or pair-wise operating characteristics). And, other loss functions such as "all or nothing" loss and estimates such as posterior median ranks can be considered.

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, and data analytic performance summaries such as SEL, $OC(\gamma)$ and plots like Figures 3 and 4 should accompany any analysis.

A Appendix

A.1 Optimizing weighted squared error loss (WSEL)

Theorem 7. *Under weighted squared error loss:*

$$\sum_k \omega_k \left(R_k^{est} - R_k\right)^2, \tag{15}$$

the optimal rank estimates are

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

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

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

Thus, the $\overline{R}_k$ are optimal.

When all $w_k \equiv w$,

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

optimizes (15) subject to the R_k^{est} exhausting the integers $(1, \ldots, K)$. To see this, if $0 \leq \mathsf{E}(R_i) =$

$m_i \leq \mathsf{E}(R_j) = m_j$, $r_i < r_j$, then

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

and the $\hat{R}_k$ are optimal. □

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

$$\begin{aligned}
& \omega_i(m_i - r_i)^2 + \omega_j(m_j - r_j)^2 < \omega_i(m_i - r_j)^2 + \omega_j(m_j - r_i)^2, \text{if } r_j > r_i \\
\iff\ & (r_j - r_i)((1 - \frac{\omega_j}{\omega_i})(r_i + r_j - 2m_j) + 2(m_j - m_i)) > 0, \text{if } r_j > r_i \\
\iff\ & (1 - \frac{\omega_j}{\omega_i})(r_i + r_j - 2m_j) + 2(m_j - m_i) > 0, \text{if } r_j > r_i. \qquad (16)
\end{aligned}$$

will help to solve the problem. By above inequality, reversing any two estimated ranks that do not align with $\bar{R}_k$ results in a smaller squared error.

Theorem 8. *Starting from any initial ranks, iteratively switch the position of unit i and unit j, $i, j = 1, ...K$, if inequality (16) is satisfied will lead to optimal ranking under the weighted square loss function (15).*

Proof. Since each switch will decrease the expected loss and there are at most $n!$ possible values of the expected loss , the switches will stop at some step. When this iteration stops, for any (i, j), inequality (16) will not be satisfied. On the other hand, the optimal ranking result exists due to the fact that there are at most $n!$ possible ranking results, and a necessary condition of the optimality is that no pair of coordinates (i, j) will satisfy inequality (16). This means that the optimal ranking result has the exact same order as the ranking result at stop of iterations. □

Convergence 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 a pairwise-switch optimization algorithm is impractical.

A.2 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.

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

□

Lemma 2. *(Rearrangement Inequality,see Hardy et al. (1967)) If $a_1 \leq a_2 \leq ... \leq a_n$ and $b_1 \leq b_2 \leq ... \leq b_n$, $b_{(1)}, b_{(2)}, ...b_{(n)}$ is a permutation of $b_1, b_2, ...b_n$, 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_{i1}, a_{i2}) and (b_{i1}, b_{i2}) 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. □

Proof of Theorem 2

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

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

For optimum ranking, the following conditions are necessary.

1. By Lemma 1, for any (i_1, i_2) satisfying $(1 \leq i_1 \leq [\gamma(K+1)],\ [\gamma(K+1)]+1 \leq i_2 \leq K)$, 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 largest $K - [\gamma(K+1)]$ from $\{\text{pr}(R_k \geq \gamma(K+1)) : k = 1, ...K\}$ to be the largest $(1-\gamma)K$ ranks.

2. By Lemma 2

   (a) For the set $\{k : R_k = R_{(i)}, i = 1, \cdots, [\gamma(K+1)]\}$, 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 $(1, \ldots \gamma K)$, the ranks should be determined by ranking the $\text{pr}(R_k \geq \gamma(K+1))$.

   (b) For the set $\{k : R_k = R_{(i)}, i = [\gamma(K+1)]+1, \cdots, K\}$, 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 $(\gamma K + 1, \ldots, K)$, 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 $\text{pr}(R_k \geq \gamma(K+1))$.

A.3 Optimization procedure for $L^\dagger$

Similar to 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,

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

The following is equivalent to the foregoing.

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

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

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

This result is useful and different from that of WSEL in section A.1 in the sense that we can now successfully get a quantity depend on unit index i only. However, as for $L_{0/1}$ optimizing 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.

A.4 Optimizing $L^\ddagger$

Similar to that of WSEL in section A.1, pairwise switch algorithm is computationally challenging since whether to switch a pair of units depend not only their relative position, but also their exact estimated ranks. Thus in each iteration, all pairwise relations have to be checked again. 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 $L^\ddagger$.

A.5 Proof of Theorem 5

In this proof, we use use $\mathbf{Y_K}$ rather than $\mathbf{Y}$ to stress that as K goes to infinity, the length of $\mathbf{Y}$ changes. For $\bar{G}_{\mathbf{Y_K}}(t) = \frac{1}{K}\sum_{k=1}^{K} \text{pr}(\theta_k \leq t | \mathbf{Y_K})$, we prove a stronger statement: as $K \to \infty$, $|\text{pr}(P_k \geq \gamma | \mathbf{Y_K}) - \text{pr}(\theta_k \geq \bar{G}^{-1}_{\mathbf{Y_K}}(\gamma) | \mathbf{Y_K})| \to 0$, where P_k is the true percentile of θ_k, $\mathbf{Y_K}$ is the vector $(Y_1, Y_2, \cdots, Y_K)$.

The posterior independence of θ_k is straightforward. Denote $\boldsymbol{\theta} = (\theta_1, \cdots, \theta_k, \cdots, \theta_K)$ and $\boldsymbol{\theta}^{(-k)} = (\theta_1, \cdots, \theta_{k-1}, \theta_{k+1}, \cdots, \theta_K)$, where $\theta_k \overset{ind}{\sim} g(\cdot|Y_k) = g_k(\cdot)$.

Let $\theta_{(\gamma)} = \theta_{(i,K)}$ be the γ_{th} quantile of $\boldsymbol{\theta}$, if $\frac{i}{K} \leq r < \frac{i+1}{K}$, where $\theta_{(i,K)}$ is the i_{th} largest number of $\boldsymbol{\theta}$. $\theta_{(\gamma)}^{(-k)}$ is respectively the γ_{th} quantile of $\boldsymbol{\theta}^{(-k)}$. We also denoted $\theta_{(i-1,K-1)}$ as the $(i-1)_{\text{th}}$ largest number of $\boldsymbol{\theta}^{(-k)}$.

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

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

For the second term in (17)

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

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

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

Thus the second term in (17) is zero,

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

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 $\theta_{(\gamma)}^{(-k)} \to G^{-1}(\gamma)$ in probability as K goes to ∞. Since we assume that $\theta_k|Y_k$ has a uniformly bounded finite second moment, so does $\theta_{(\gamma)}^{(-k)}|\mathbf{Y_K}$. Thus $\mathsf{E}[\theta_{(\gamma)}^{(-k)}|\mathbf{Y_K}] \to G^{-1}(\gamma)$.

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

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

Since $\bar{G}^{-1}_{\mathbf{Y_K}}(\gamma) \to G^{-1}(\gamma)$, by (18) and (19), $|\mathsf{pr}(P_k \geq \gamma|\mathbf{Y_K}) - \mathsf{pr}(\theta_k \geq \bar{G}^{-1}_{\mathbf{Y_K}}(\gamma)|\mathbf{Y_K})| \to 0$

References

Benjamini, Y. and Hochberg, Y. "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 (1995).

Carlin, B. and Louis, T. *Bayes and Empirical Bayes Methods for Data Analysis*. Boca Raton, FL: Chapman and Hall/CRC Press, 2^{nd} edition (2000).

Christiansen, C. and Morris, C. "Improving the statistical approach to health care provider profiling." *Annals of Internal Medicine*, 127:764–768 (1997).

Conlon, E. and Louis, T. "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 (1999).

David, H. A. and Nagaraja, H. N. *Order Statistics*. Wiley, third edition (2003).

Devine, O. and Louis, T. "A constrained empirical Bayes estimator for incidence rates in areas with small populations." *Statistics in Medicine*, 13:1119–1133 (1994).

Devine, O., Louis, T., and Halloran, M. "Empirical Bayes estimators for spatially correlated incidence rates." *Environmetrics*, 5:381–398 (1994).

DuMouchel, W. "Bayesian Data Mining in Large Frequency Tables, With An Application to the FDA Spontaneous Reporting System (with discussion)." *The American Statistician*, 53:177–190 (1999).

Efron, B., Tibshirani, R., Storey, J. D., and Tusher, V. "Empirical Bayes analysis of a microarray experiment." *Journal of the American Statistical Association*, 96(456):1151–1160 (2001).

ESRD. "1999 Annual Report: ESRD Clinical Performance Measures Project." Technical report, Health Care Financing Administration (2000).

Gelman, A. and Price, P. "All maps of parameter estimates are misleading." *Statistics in Medicine*, 18:3221–3234 (1999).

Goldstein, H. and Spiegelhalter, D. "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 (1996).

Hardy, G. H., Littlewood, J. E., and Polya, G. *Inequalitites*. Cambridge University Press, 2nd edition (1967).

Lacson, E., Teng, M., Lazarus, J., Lew, N., Lowrie, E., and Owen, W. "Limitations of the facility-specific standardized mortality ratio for profiling health care quality in Dialysis." *American Journal of Kidney Diseases*, 37:267–275 (2001).

Laird, N. M. and Louis, T. A. "Empirical Bayes ranking methods." *Journal of Educational Statistics*, 14:29–46 (1989).

Landrum, M., Bronskill, S., and Normand, S.-L. "Analytic methods for constructing cross-sectional profiles of health care providers." *Health Services and Outcomes Research Methodology*, 1:23–48 (2000).

Lin, R., Louis, T. A., Paddock, S. M., and Ridgeway, G. "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 (2004).

Liu, J., Louis, T., Pan, W., Ma, J., and Collins, A. "Methods for estimating and interpreting provider-specific, standardized mortality ratios." *Health Services and Outcomes Research Methodology*, 4:135–149 (2004).

Lockwood, J., Louis, T., and McCaffrey, D. "Uncertainty in rank estimation: Implications for value-added modeling accountability systems." *Journal of Educational and Behavioral Statistics*, 27(3):255–270 (2002).

Louis, T. and Shen, W. "Innovations in Bayes and empirical Bayes methods: Estimating parameters, populations and ranks." *Statistics in Medicine*, 18:2493–2505 (1999).

McClellan, M. and Staiger, D. "The Quality of Health Care Providers." Technical Report 7327, National Bureau of Economic Research, Working Paper (1999).

Normand, S.-L. T., Glickman, M. E., and Gatsonis, C. A. "Statistical methods for profiling providers of medical care: Issues and applications." *Journal of the American Statistical Association*, 92:803–814 (1997).

Paddock, S., Ridgeway, G., Lin, R., and Louis, T. A. "Flexible prior distributions and triple goal estimates in two-stage,hierarchical linear models." Technical report, Rand Statistics group (2005).

Shen, W. and Louis, T. "Triple-goal estimates in two-stage, hierarchical models." *Journal of the Royal Statistical Society, Series B*, 60:455–471 (1998).

Storey, J. D. "A direct approach to false discovery rates." *Journal of the Royal Statistical Society, Series B, Methodological*, 64(3):479–498 (2002).

—. "The Positive False Discovery Rate: A Bayesian Interpretation and the q-Value." *The Annals of Statistics*, 31(6):2013–2035 (2003).

Tusher, V. G., Tibshirani, R., and Chu, G. "Significance analysis of microarrays applied to the ionizing radiation response." *Procedings of National Academy of Sciences*, 98(9):5116–5121 (2001).

USRDS. "2003 Annual Data Report: Atlas of end-stage renal disease in the United States." Technical report, Health Care Financing Administration (2003).

Wolfe, R., Gaylin, D., Port, F., Held, P., and Wood, C. "Using USRDS generated mortality tables to compare local ESRD mortality rates to national rates." *Kidney Int*, 42(4):991–6 (1992).

Wright, D. L., Stern, H. S., and Cressie, N. "Loss functions for estimation of extrema with an application to disease mapping." *The Canadian Journal of Statistics*, 31(3):251–266 (2003).

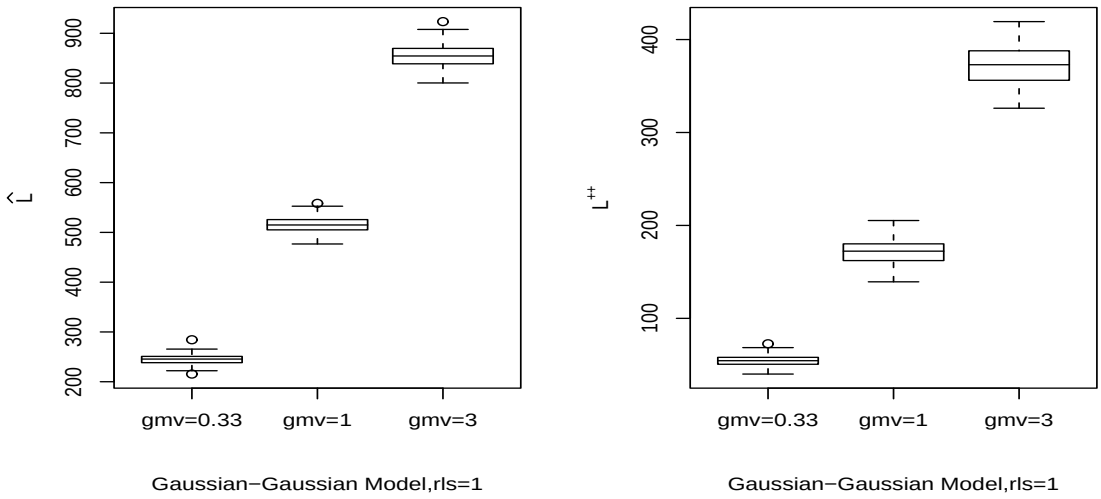

Figure 1: Unit-specific, $\hat{L}$ and $L^{\ddagger}$ risk as a function of gmv for $K = 200$, $\gamma = 0.8$, $rls = 1$.

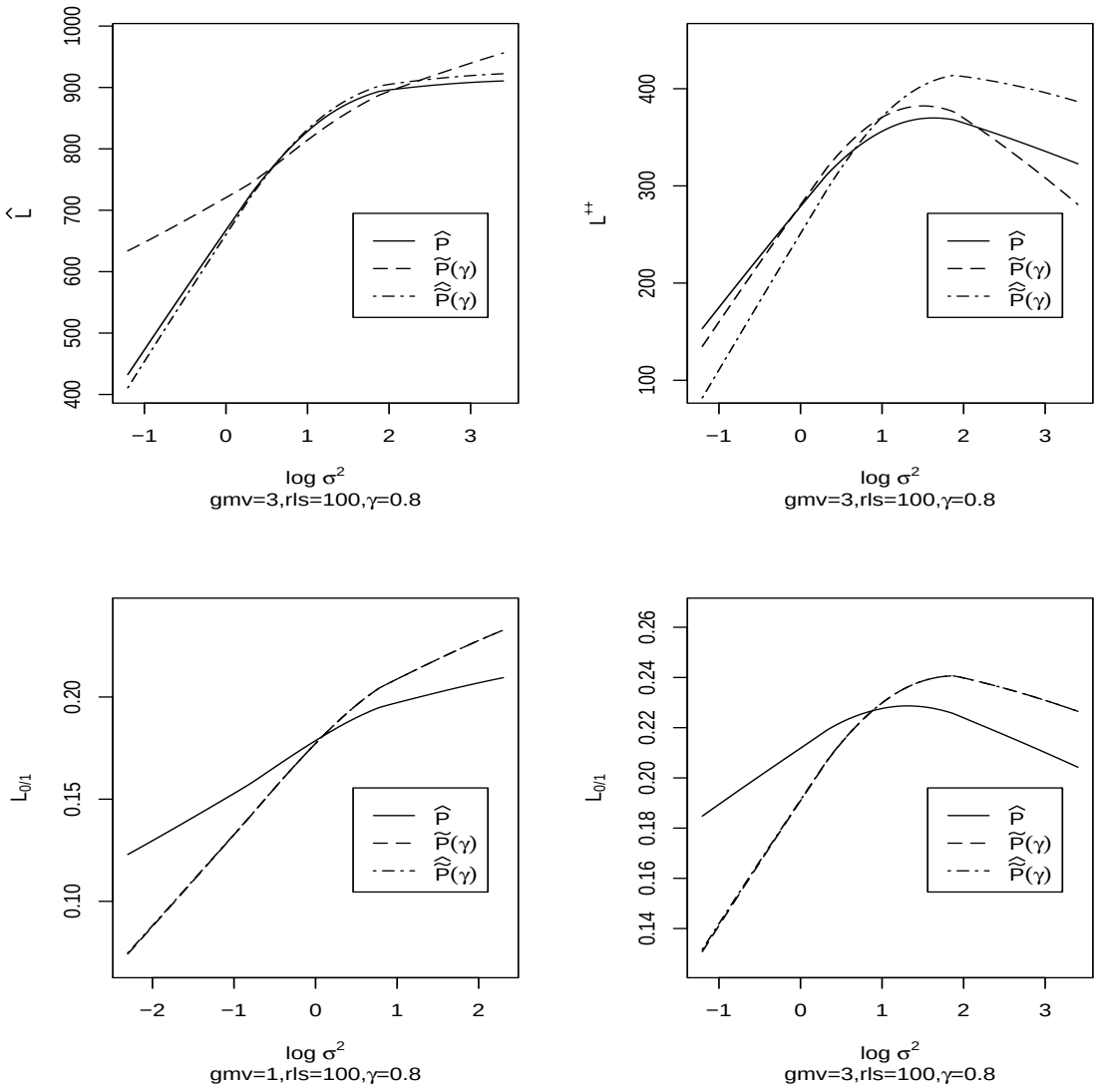

Figure 2: Loess smoothed, unit-specific performance of $\hat{P}_k$, $\tilde{P}_k(\gamma)$ and $\hat{\tilde{P}}_k(\gamma)$ under $\hat{L}$, $L^{\ddagger}$, and $L_{0/1}$ as a function of unit-specific variance (σ_k^2) for $\gamma = 0.8$, $rls = 100$ and $gmv = 1$ and 3.

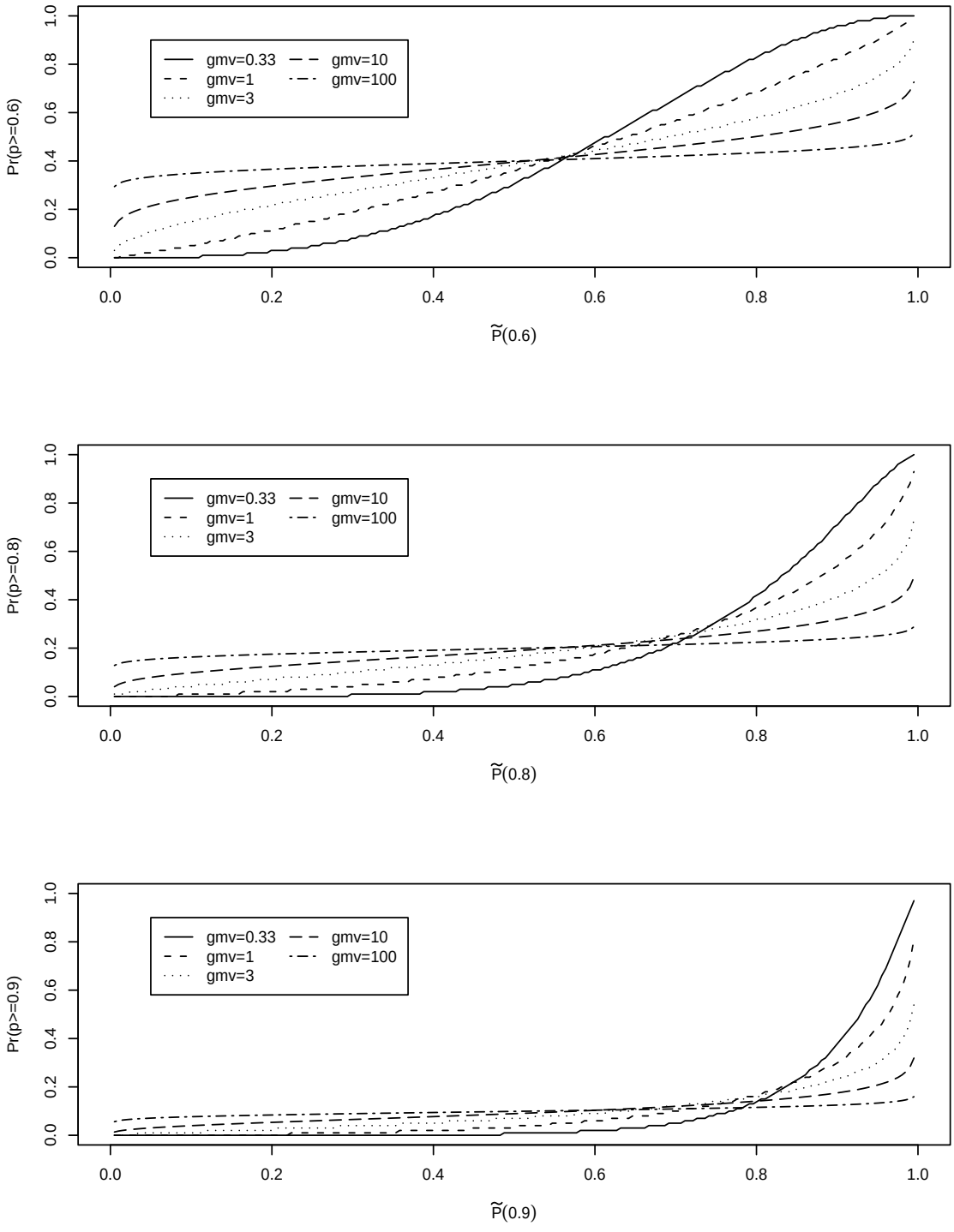

Figure 3: Average posterior classification probabilities as a function of the optimally estimated percentiles for $rls = 1$, $gmv = (0.33, 1, 3, 10, 100)$, $\gamma = (0.6, 0.8, 0.9)$.

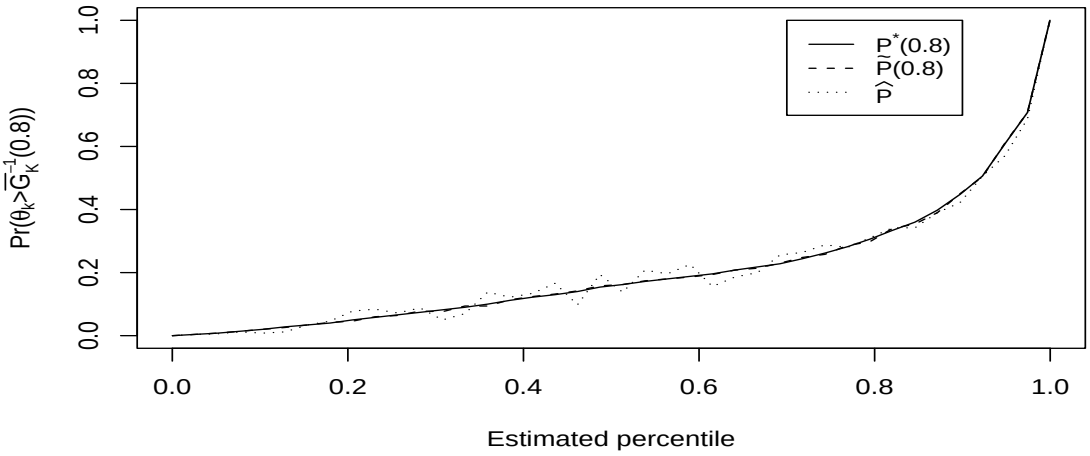

Figure 4: $\Pr(\theta_k > 0.18 \mid \mathbf{Y})$ with X-axis percentiles determined separately by the three percentiling methods.

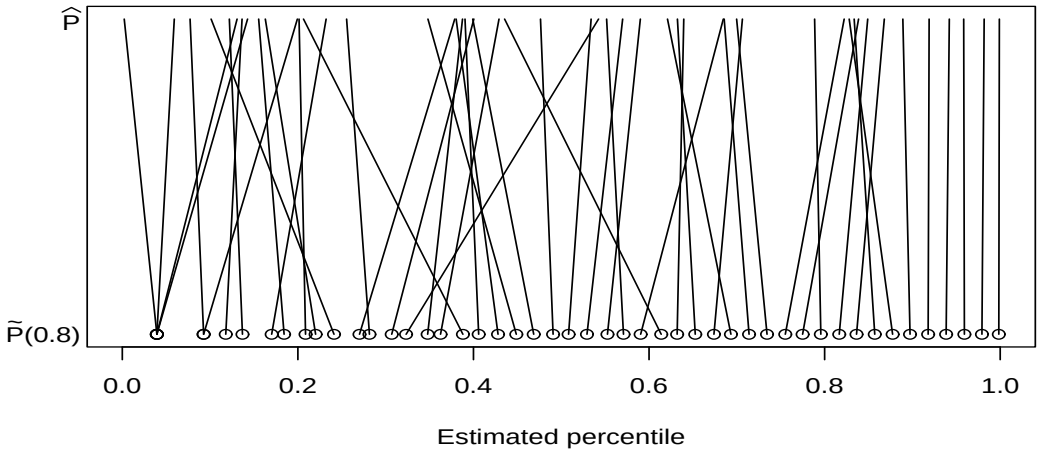

Figure 5: Comparison plot of SMR-based $\hat{P}_k$, $\tilde{P}_k(\gamma)$ and $\hat{\tilde{P}}_k(\gamma)$ for the 50 dialysis centers evenly distributed on $\tilde{P}_k(\gamma)$, $\gamma = 0.8$