Loss Function Based Ranking in Two-Stage, Hierarchical Models

Rongheng Lin[1], Thomas A. Louis, Susan M. Paddock, Greg Ridgeway[2]

Oct 26, 2004

[1]Address for correspondence: 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

[2]Rongheng Lin is Ph.D candidate and Thomas A. Louis is Professor, Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, Baltimore, MD 21205, U.S.A. Susan M. Paddock is Full Statistician and Greg Ridgeway is Full Statistician, RAND, Santa Monica, CA 90401, U.S.A. This work was supported by grant 1-R01-DK61662 from U.S. NIH National Institute of Diabetes, Digestive and Kidney Diseases.

Loss Function Based Ranking in Two-Stage, Hierarchical Models

Abstract

Several authors have studied the performance of optimal, squared error loss (SEL) estimated ranks. Though these are effective, in many applications interest focuses on identifying the relatively good (e.g., in the upper 10%) or relatively poor performers. We construct loss functions that address this goal and evaluate candidate rank estimates, some of which optimize specific loss functions. We study performance for a fully parametric hierarchical model with a Gaussian prior and Gaussian sampling distributions, evaluating performance for several loss functions. 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. We compare inferences produced by the candidate estimates using data from The Community Tracking Study.

KEY WORDS: Bayesian Methods, Percentiles, League Tables, Decision Theory

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 (i.e., value added modeling) burgeons. Goals of such investigations include valid and efficient estimation of population parameters such as the 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 (point estimates and confidence intervals), computing the probability that a unit's true, underlying performance is in an "acceptable" (or unacceptable) region, ranking/percentiling units as input to profiling and league tables (Goldstein and Spiegelhalter 1996), identification of excellent and poor performers.

Bayesian models are very effective in structuring these assessments. Inferences always depend on the posterior distribution, but how the posterior is used should depend on inferential goals. As Shen and Louis (1998) show and Gelman and Price (1999) present in detail, no single set of estimates can effectively address the multiple goals. For example, the histogram of maximum likelihood estimates (MLEs) is always over-dispersed relative to the histogram of the true random effects. In a Bayesian setting, the histogram of posterior means (PMs) is always under-dispersed. Therefore, when inferences address non-standard goals, they are best guided by a loss function. In this vein, Shen and Louis (1998) used squared-error loss (SEL) to estimate the cumulative distribution function of unit-specific parameters in a two stage hierarchical model and to estimate the ranks of these parameters. They show that ranks estimated in this way outperform competitors.

As the following authors show, even optimal percentile estimates can perform quite poorly in most real-world settings. Lockwood and co-authors (Lockwood, Louis, and McCaffrey 2002) present simulation-based evaluations of SEL-based optimal percentiles with application to comparing performance of teachers and schools (value added modeling in educational assessment), Liu et al. (2004) provide evaluations with specific reference to ranking dialysis centers with respect to standardized mortality ratios, Goldstein and Spiegelhalter (1996) compute "league tables" for schools and surgeons.

SEL applies the same distance function to all (estimated, true) percentile pairs, but in many applications interest focuses on identifying the relatively good (e.g., in the upper 10%) 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; genes with the truly highest differential expression should be the ones to generate hypotheses. In this report we construct loss functions that address such goals and evaluate candidate rank (percentile) estimates, some of which optimize specific loss functions. We study performance for a fully parametric hierarchical model with a Gaussian prior and Gaussian sampling distributions (both with constant and unit-specific sampling variance). We evaluate performance for the loss function used to generate (or motivate) the ranks and for other candidate loss functions, thereby evaluating both the

performance of the optimal or loss function tuned estimates and robustness to assumptions and goals.

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. We compare inferences produced by the candidate estimates using data from The Community Tracking Study (Kemper et al. 1996; Center for Studying Health System Change 2003).

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:

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

Note that the f_k can depend on k. This 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 decision rule $\mathbf{a}(\mathbf{Y})$ and loss function $L(\boldsymbol{\theta}, \mathbf{a})$, the optimal Bayesian procedure (the optimal $\mathbf{a}(\mathbf{Y})$) minimizes the posterior and preposterior Bayes risks,

$$\begin{aligned} R_G(\mathbf{a}(\mathbf{Y}), \mathbf{Y}) &= E_{\boldsymbol{\theta}|\mathbf{Y}}[L(\boldsymbol{\theta}, \mathbf{a}(\mathbf{Y}) \mid \mathbf{Y}] \\ R_G(\mathbf{a}) &= E_{\mathbf{Y}}[R_G(\mathbf{a}(\mathbf{Y}), \mathbf{Y})]. \end{aligned}$$

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

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

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

This is the most common estimation loss function. It is 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 $\theta_k^{pm} = E(\theta_k \mid \mathbf{Y})$. To produce SEL-optimal ranks, minimize the squared error loss (SEL) for them,

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

where R_k^{est} are generic estimates. $\hat{L}$ is optimized by the posterior means,

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

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

$$\hat{R}_k(\mathbf{Y}) = \text{rank}(\bar{R}_k(\mathbf{Y})), \tag{4}$$

which are the optimal, integer estimates. See Section A.1 for additional details on the SEL approach.

In the sequel, 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 in many applications percentiles,

$$P_k = R_k/(K+1) \tag{5}$$

are used, e.g. $\hat{P}_k = \hat{R}_k/(K+1)$. Also, percentiles have the advantage of normalizing with respect to the number of units K. 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 and their use similarly standardizes our new loss functions. We use both ranks and percentiles with the choice depending on presentation clarity.

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

When posterior distributions are not stochastically ordered, SEL ($\hat{L}$)-optimal percentiles (the $\hat{P}_k$) outperform ranking the θ_k^{pm} and substantially outperform ranking the MLEs. However, $\hat{L}$ evaluates general performance without specific attention to identifying the relatively good or relatively poor performers. 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 (identification of the lower group is similar) and may also include a distance penalty. For notational clarity we assume that γK is an integer, so $\gamma(K+1)$ is not an integer and in the following it is not necessary to make the distinction between $>$ and $\geq$ ($<$ and $\leq$).

4.1 Summed, unit-specific loss functions

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

$$\begin{aligned} AB_k(\gamma, P_k, P_k^{est}) &= I_{\{P_k > \gamma,\, P_k^{est} < \gamma\}} = I_{\{R_k > \gamma(K+1),\, R_k^{est} < \gamma(K+1)\}} \\ BA_k(\gamma, P_k, P_k^{est}) &= I_{\{P_k < \gamma,\, P_k^{est} > \gamma\}} = I_{\{R_k < \gamma(K+1),\, R_k^{est} > \gamma(K+1)\}}, \end{aligned} \tag{6}$$

AB_k and BA_k indicate the two possible modes of misclassification of the percentiles. AB_k indicates that the true percentile is above the cutoff, and therefore unit k should be flagged; however, 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 $p, q, c > 0$ define,

$$\begin{aligned}
\tilde{L}(\gamma, p, q, c) &= \frac{1}{K}\sum_k \{|\gamma - P_k^{est}|^p AB_k(\gamma, P_k, P_k^{est}) + c|P_k^{est} - \gamma|^q BA_k(\gamma, P_k, P_k^{est})\} \\
L^{\dagger}(\gamma, p, q, c) &= \frac{1}{K}\sum_k \{|P_k - \gamma|^p AB_k(\gamma, P_k, P_k^{est}) + c|\gamma - P_k|^q BA_k(\gamma, P_k, P_k^{est})\} \qquad (7) \\
L^{\ddagger}(\gamma, p, q, c) &= \frac{1}{K}\sum_k \{|P_k - P_k^{est}|^p AB_k(\gamma, P_k, P_k^{est}) + c|P_k^{est} - P_k|^q BA_k(\gamma, P_k, P_k^{est})\} \\
L_{0/1}(\gamma) &= \frac{\sum_k \{AB_k(\gamma, P_k, P_k^{est}) + BA_k(\gamma, P_k, P_k^{est})\}}{K} = 2\frac{\sum_k AB_k(\gamma, P_k, P_k^{est})}{K} \\
&= \frac{\#(\text{misclassifications})}{K} = \tilde{L}(\gamma, 0, 0, 1) = L^{\dagger}(\gamma, 0, 0, 1) = L^{\ddagger}(\gamma, 0, 0, 1) \\
L_{all}(\gamma) &= 0 \text{ or } 1 \text{ according as all classifications are correct or at least 1 is incorrect} \\
&= 1 - \prod_k \{1 - AB_k(\gamma, P_k, P_k^{est})\}\{1 - BA_k(\gamma, P_k, P_k^{est})\}
\end{aligned}$$

$L_{all}(\gamma)$ is "all or nothing," with no penalty if all (P_k^{est}, P_k) pairs are on the same side of the cut point and penalizing by 1 if there are any discrepancies. The other 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 the cut point, the loss functions penalize by an amount that depends on the distance of either the estimated or the true percentile from the cut point or on the distance between the true and estimated percentiles. 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 the preceding three functions with $p = q = 0$, $c = 1$.

Several of our mathematical results apply to the general form of the loss functions (7), but our simulations evaluate them for $p = q = 2, c = 1$. The foregoing simplify to:

$$\begin{aligned}
\tilde{L}(\gamma) &= \frac{1}{K}\sum_k (\gamma - P_k^{est})^2 \{AB_k(\gamma, P_k, P_k^{est}) + BA_k(\gamma, P_k, P_k^{est})\} \\
L^{\dagger}(\gamma) &= \frac{1}{K}\sum_k (P_k - \gamma)^2 \{AB_k(\gamma, P_k, P_k^{est}) + BA_k(\gamma, P_k, P_k^{est})\} \qquad (8) \\
L^{\ddagger}(\gamma) &= \frac{1}{K}\sum_k (P_k - P_k^{est})^2 \{AB_k(\gamma, P_k, P_k^{est}) + BA_k(\gamma, P_k, P_k^{est})\}.
\end{aligned}$$

4.2 A convex combination loss function

Though the loss functions (7) and (8) are most relevant in applications, the convex combination loss,

$$\hat{L}^{w}_{0/1}(\gamma) \;=\; (1-w)L_{0/1}(\gamma) + w\hat{L} \tag{9}$$

is a possible alternative to $L^{\ddagger}$. (Convex combination losses derived from $\tilde{L}$ and $L^{\dagger}$ are less interesting in that once the (above γ)/(below γ) groups are determined, no further control is possible). Once the (above γ)/(below γ) groups have been specified, risk is controlled by constrained optimization of $\hat{L}$. We do not report simulations for $\hat{L}^{w}_{0/1}(\gamma)$, but use it to motivate the estimator $\hat{\tilde{P}}_k(\gamma)$ defined in Section 5.5.

5 Optimizers and other candidate estimates

In addition to $\hat{P}_k$ which minimizes $\hat{L}$ (SEL, see equation (4)), we consider an estimator that optimizes both $L_{0/1}$ and $\tilde{L}$ and one that is motivated by $\hat{L}^{w}_{0/1}$. Note that when the posteriors are stochastically ordered the $\hat{P}_k$ are optimal for broad class of loss functions.

5.1 Optimizing L_{all}

Optimization requires finding the subscripts $(k_1, k_2, \ldots, k_{(1-\gamma)K})$ that maximize:

$$pr(\{R_{k_1}, R_{k_2}, \ldots, R_{k_{(1-\gamma)K}}\} = \{\gamma K, \gamma K + 1, \ldots, K\}).$$

Finding this optimal subset requires the full joint distribution of the ranks and evaluating all possible $(1-\gamma)K$-element subsets. Even if the joint distribution is available, the optimization is infeasible unless $\begin{pmatrix} K \\ [\gamma K] \end{pmatrix}$ is small.

5.2 Optimizing $L_{0/1}$

Let,

$$\begin{aligned} p_{k\ell} &= pr(R_k = \ell \mid \mathbf{Y}) \\ \pi_k(\gamma) &= pr(R_k \geq \gamma(K+1) \mid \mathbf{Y}) = pr(P_k \geq \gamma \mid \mathbf{Y}) = \sum_{\ell=[\gamma K]+1}^{K} p_{k\ell} \qquad (10) \\ \tilde{R}_k(\gamma) &= \text{rank}(\pi_k(\gamma));\ \tilde{P}_k(\gamma) = \tilde{R}_k(\gamma)/(K+1) \qquad (11) \end{aligned}$$

Theorem 1. The $\tilde{P}_k(\gamma)$ optimize $L_{0/1}$, but not uniquely.($L_{0/1}$ does not require optimal ordering, only optimal categorization above and below the cut point. Other loss functions provide additional structure and resolution.)

Proof. Rewrite the loss function as a function of the number of coordinates 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 classified set and T is the true set of $(1-\gamma)K$ numbers. We need to maximize the expected number of correctly classified coordinates:

$$\begin{aligned} \mathrm{E}|A \cap T| &= \mathrm{E}\sum I(\theta_i \in A \cap T) = \mathrm{E}\sum I(\theta_i \in A) I(\theta_i \in T) \\ &= \mathrm{E}\sum_{\theta_i \in A} I(\theta_i \in T) = \sum_{\theta_i \in A} pr(\theta_i \in T). \end{aligned}$$

So, to optimize $L_{0/1}$, for each θ_i calculate $pr(\theta_i \in T)$, rank these probabilities and select the largest $(1-\gamma)K$ of them. That is to calculate $pr(\theta_k \in T) = \sum_{\ell=(1-\gamma)K}^{K} p_{k\ell} = \pi_k(\gamma)$. Therefore, ordering by the $\pi_k(\gamma)$ creates the optimal "(above γ)/(below γ) " groups.

□

This computation can be difficult and time-consuming, but is Monte Carlo implementable. We note that here the optimizer $\tilde{P}_k(\gamma)$ produced by $\pi_k(\gamma)$ is in the similar spirt of the ranking proposed by Normand et al. (1997), which suggests to use $\Pr(\theta_k > c)$ to compare the performance.

5.3 Optimizing $\tilde{L}$

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

The proof, in Section A.2, shows that the $\tilde{P}_k(\gamma)$ are 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 $|\gamma - P_k^{est}|$.

5.4 Optimizing $L^\dagger$

We have not identified a general representation or algorithm for the optimal ranks. However, Section A.3 develops an optimization procedure for the case $p = q = 2, -\infty < c < \infty$. Note

that 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.

5.5 Optimizing $L^{\ddagger}$

We have not identified a general representation or algorithm for the optimal ranks. However, Section A.4 develops a helpful relation. The estimator $\hat{\tilde{P}}_k(\gamma)$, motivated by loss function (9), is very effective though not optimal for $L^{\ddagger}$.

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

Since the (above γ)/(below γ) groups are formed by $\tilde{P}_k(\gamma)$, $\hat{\tilde{P}}_k(\gamma)$ also minimizes $L_{0/1}$. In addition, we have the following.

Theorem 3. For any (above γ)/(below γ) grouping, ordering within the groups by the $\hat{P}_k$ produces the constrained minimum for $\hat{L}$.

Proof. 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,

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

Nothing can be done to reduce the variance terms and 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 $R_k^{est} = \mathrm{rank}(\overline{R}_k) = k, k = 1, \ldots, \gamma K$ 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 $\hat{R}_k = \text{rank}(\overline{R}_k)$ produces the result. □

Consequence: For the convex combination loss function $\hat{L}^w_{0/1}(\gamma)$ (formula 9), 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 (broadening its potential use).

5.6 Relations among estimators

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

$$R^*_k(\nu) = \frac{K(K+1)}{2(K-\nu+1)} pr(R_k \geq \nu) = \frac{K(K+1)}{2(K-\nu+1)} \pi_k(\gamma).$$

The rank of the $R^*_k(\nu)$ is same as the rank of the $\pi_k(\gamma)$ and so each generate the $\tilde{R}_k(\gamma)$. Note that,

$$\sum_k \frac{K(K+1)}{2(K-\nu+1)} \pi_k(\gamma) = \sum_k R^*_k(\nu) = \frac{K(K+1)}{2} = \sum_k R_k.$$

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 $\overline{R}_k$ outperforms $R^*_k(\nu)$ 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 $\text{rank}(R^*_k(\nu))$ 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

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

□

Relation (12) can be used to show that when the posterior distributions are stochastically ordered, $\hat{R}_k \equiv \tilde{R}_k(\gamma)$. In this case, for all γ the $\pi_k(\gamma)$ have the same order and so the $\overline{R}_k$ inherit this order. More generally, if the posterior distributions $g_k(t \mid Y_k)$ are stochastically ordered (the $G_k(t \mid Y_k)$ never cross), then for a broad class of loss functions (including many that we consider) the $\hat{R}_k$ are optimal.

6 Simulation Scenarios and Results

6.1 Scenarios

6.1.1 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,

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

The variances $\{\sigma_k^2\}$ form an ordered, geometric sequence with specified ratio of geometric mean to τ^2. Sine we use $\tau^2 = 1$ in our prior, this ratio is then equal to the geometric mean of the variance sequence $gmv = GM(\sigma_1^2, \ldots, \sigma_K^2)$. Furthermore, we use $rls = \sigma_K^2/\sigma_1^2$, the ratio of the largest σ_k^2 to the smallest as an indicator of range of variance.

Each result is based on $K = 200$ with $nrep = 2000$ simulation replications. We compute the $p_{k\ell}$ (needed to compute $\tilde{P}_k(\gamma)$) using an independent sample Monte Carlo with $n = 2000$ draws. All simulations are for loss functions with $p = q = 2;\ c = 1$.

6.1.2 Mixture model

We consider the situation that θ comes from a two-component mixture of Gaussian distribution which has mean 0 and variance 1:

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

where

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

This form comes from standardizing $\theta \overset{iid}{\sim} (1-\epsilon)N(0,1)+\epsilon N(\Delta, \xi^2)$ to mean 0 and variance 1.

We then calculate the preposterior risk for estimators in percentiles in two way: assuming that we know exactly $\theta's$ are from mixture prior (part II of the table); incorrectly assuming that $\theta's$ are from N(0,1) (part III of the table). Like the Gaussian-Gaussian scenario, we selected $K = 200, n = 2000, \text{nrep} = 2000$; $p = q = 2$; $c = 1$. We use several sets of extra parameters values and here only present the results for $\epsilon = .1, \Delta = 3.40, \xi^2 = .25, \gamma = .9$.

6.2 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. We do not include these estimates in subsequent evaluations.

Table 2 reports results for the remaining competitors. Again, when $rls = 1$, $\tilde{P}_k(\gamma) \equiv \hat{P}_k \equiv \hat{\tilde{P}}_k(\gamma)$ and so the SEL results in the first and seventh rows show the small residual variation in the simulation and the Monte Carlo evaluation of the posterior distribution needed to

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 $gmv = 1$.

produce $\tilde{P}_k(\gamma)$. Results for other values of rls show that for $\hat{L}$, as must be the case, $\hat{P}$ outperforms $\tilde{P}_k(\gamma)$ and $\hat{\tilde{P}}_k(\gamma)$. Since $\tilde{P}_k(\gamma)$ optimizes $L_{0/1}$ and $\tilde{L}$, for $rls \neq 1$ it performs better for these loss function than do $\hat{P}$ and $\hat{\tilde{P}}_k(\gamma)$. Even though $\hat{\tilde{P}}_k(\gamma)$ optimizes $\hat{L}^w$ for sufficiently small w, it performs relatively poorly for $L^{\ddagger}$. $\tilde{P}_k(\gamma)$ appears to dominate and $\hat{P}_k$ performs well. The poor performance of $\hat{\tilde{P}}_k(\gamma)$ shows that a unit-specific combining of a misclassification penalty with squared-error loss is essentially different from using them in an overall convex combination. Preliminary results show that similar relations among the estimators hold for a 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 to these, but a look at extremes for gmv is instructive. Results (not shown) indicate that for $rls = 1$, the risk associated with $L_{0/1}$ is of the form $const(gmv)\gamma(1-\gamma)$. Furthermore, when $gmv = 0$, the data are fully informative, $Y_k \equiv \theta_k$ and all risks are 0. When $\sigma_k^2 \equiv \infty$, $gmv = \infty$ and the Y_k provide no information on the θs or on the P_k. Table 3 displays the preposterior risk for the no information ($\sigma^2 = \infty$) case. These values provide an upper bound of $rls = 1$ situation to which we can compare the values in Table 2.

Under $L_{0/1}$ percentile estimates $\hat{P}$ and $\tilde{P}$ perform similarly. Many of the tabulated values are identical up to the rounding resolution.

Figure 1 to 3 are based on Gaussian-Gaussian model and no figures based on mixture models are presented here. 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

		$L_{0/1}$		$\hat{L}$			$\tilde{L}$			$L^{\ddagger}$		
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 risks for $gmv = 1$. $L_{0/1}$ is the percent; other values are 10000×(Loss). The first block is the situation of Gaussian-Gaussian model, the second block is mixture model with mixture form known and the third block is mixture model with mixture form unknown.

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 $rls = 1$ when $gmv = \infty$. $L_{0/1}$ is the percentage; other values are 10000×Risk.

loss equals the overall average risk and so the box plots summarize the sampling distribution of unit-specific risk.

6.2.1 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 have not pursued mathematical results for this non-exchangeable situation, but study it 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 $\gamma = 0.8$. The plots for $\hat{L}$ ($gmv = 3$) and for $L_{0/1}$ ($gmv = 1$) correspond to intuition 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 as a function of σ_k^2 the risk increases and then decreases. Similar patterns emerge for other values of γ, gmv and rls with the presence of a downturn depending on the proximity of γ to 0.5 and rls or gmv being sufficiently large.

These results appear anomalous; however they can be explained. If γ 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 of the percentile estimates for high-variance units towards 50%. 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 20% as $rls \to \infty$. Low variance units have small misclassification probabilities. Percentiles for units with moderate variances are not shrunken sufficiently toward 0.5 to produce a low $L_{0/1}$.

6.2.2 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 jumps from 0 to 1 at γ.

7 Analysis of the Community Tracking Study

We compare $\hat{P}_k, \tilde{P}_k(\gamma)$ and $\hat{\tilde{P}}_k(\gamma)$ using income data from The Community Tracking Study (CTS). CTS is a national survey designed to monitor changes in the health care system and their effects on insurance coverage, access to care, service use and delivery, cost, and quality of care (Kemper et al. 1996). Since 1996, data have been collected every two years on the 60 communities in the 48 contiguous states enrolled into the study. A total of 56,343 persons were sampled as part of the 60 site samples. A random subset of 12 of these sites are selected to be "high-intensity" sites that could be studied individually in greater detail. We examine data from the third wave of the study, which was conducted during 2000-2001 (Center for Studying Health System Change 2003).

Figure 4 displays the community-specific mean incomes and 95% confidence intervals, ordered by the community means. (Incomes greater than \$150,000 were set aside before computing community-specific means). There is considerable variation in mean income and the σ_k^2 vary substantially.

The Gaussian distribution provides a good approximation to the sampling distribution of community-specific average income. We assume that the true community-specific mean incomes are also Gaussian, a less tenable assumption, but sufficiently accurate for our illustrative purpose. We used the marginal likelihood to estimate the prior mean ($\hat{\mu} = \$46,118$) and standard deviation ($\hat{\tau} = \$6834$) and conduct a plug-in (naive) empirical Bayes analysis. We calculate estimates based on posterior distributions from the plug-in prior.

When the posterior distributions are stochastically ordered, all three estimators are identical.

Figure 5 shows that though the posterior cdfs do cross, they are "almost" stochastically ordered.

Figure 6 shows predictive performance and is similar to Figure 3. The data set is reasonably informative, but as is usually the case, the predictive value for the (above γ)/(below γ) classification is only moderate. For example, the posterior probability is only about 0.50 that a community with estimated percentile = 0.8 truly is in the upper 20% of the percentiles and the estimated percentile must exceed 0.9 before this probability is greater than 0.8.

Figure 7 displays percentiles comparing the three estimators (for the CTS income data, $\hat{P}_k \equiv \hat{\tilde{P}}_k(\gamma)$ and so only two estimators are displayed). All three sets of percentiles produce the same categorization above and below the $\gamma = 0.8$ cut point (none of the lines cross the $\gamma = 0.8$ vertical line). Interestingly, because $\pi_k(0.8) = 0$ for 26 of the communities, $\tilde{P}_k(\gamma)$ produces identical percentiles for them relative to the $\gamma = 0.8$ cut point. The $\hat{\tilde{P}}_k(\gamma)$ percentiles break these ties by optimizing SEL subject to maintaining the (above γ)/(below γ) categorization. The pattern of ties changes with γ. For example, when $\gamma = 0.5$, only 6 of the original 26 are tied, but now 5 other communities are tied at the largest $\tilde{P}_k(\gamma)$ with $\pi_k(0.5) = 1$.

The ties result from our calculating the posterior probabilities via a finite ($n = 4000$) Monte Carlo sample from the posterior distribution. As $n \to \infty$, the estimated posterior for each community would indicate a non-zero chance to be in each (above γ)/(below γ) region. However, the differences in actual probabilities are so small that it is appropriate to treat them as tied.

8 Discussion

Table 1 clearly shows that percentiles based on the Ys or on the posterior means of the target parameter can perform well in general but should not be used to rank or percentile. Effective approaches must be invariant with respect to a monotone transform of the target parameter(e.g., the θs). Basing inferences on the representation (2) ensures this invariance.

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. In some scenarios the benefits of using alternatives to it are notable and in other cases quite small, so the decision to replace $\hat{P}_k$ by other candidates must depend on the trade-off between the benefits of reporting general purpose percentiles and specific targeting of a loss function other than $\hat{L}$.

For $\hat{L}$, $\hat{P}$ is optimal, $\tilde{P}_k(\gamma)$ can perform poorly, especially for extreme γ and $gmv \neq 1$; $\hat{\tilde{P}}_k(\gamma)$ is almost identical to $\hat{P}_k$. For $\tilde{L}(\gamma)$, $\tilde{P}_k(\gamma)$ is optimal, $\hat{P}$ and $\hat{\tilde{P}}_k(\gamma)$ perform reasonably well. For $L^{\ddagger}(\gamma)$, $\tilde{P}_k(\gamma)$ is best, $\hat{P}_k$ performs reasonably well; $\hat{\tilde{P}}_k(\gamma)$ is worst, but not very far from the optimal. Recall that $\hat{\tilde{P}}_k(\gamma)$ will be optimal for $\hat{L}^w_{0/1}$ with a suitably small w and should perform well for a broad range of w values. Though by no means optimal, $\hat{\tilde{P}}_k(\gamma)$ performs very well for Ł‡.

Importantly, as do Liu et al. (2004) and Lockwood et al. (2002), we show that even the optimal estimates can perform poorly, unless the data are highly informative (gmv is small). Therefore, assessments such as those in Figures 3 and 6 should be part of any analysis.

Though the estimation approach applies to general models, we have only studied performance for the fully parametric, Gaussian/Gaussian model with a known prior distribution. Results will be essentially identical for empirical Bayes (or Bayes empirical Bayes) for K sufficiently large. Additional studies for a two component Gaussian mixture prior, a "frequentist scenario" with a fixed set of parameters and repeated sampling only from the Gaussian sampling distribution conditional on these parameters, and a robust Bayes analysis based on flexible priors such as the Dirichlet Process (Escobar 1994) will address efficiency and robustness. Broadening performance evaluations to other sampling distributions (e.g., Poisson) is important.

The new loss functions we consider address classification into the (above γ)/(below γ) categories. Extensions to three categories (below γ_1, between γ_1 and γ_2; above γ_2) will address the goal of identifying the top, middle and bottom performers. Characterizing and comput-

ing the optimizers for these loss functions is challenging, though we predict that, as for the (above γ)/(below γ) context we have studied, $\hat{P}_k$ will perform well.

A Appendix

A.1 Optimizing weighted squared error loss (WSEL)

Theorem 5. *Under weighted squared error loss:*

$$\sum_k \omega_k \left(R_k^{est} - R_k\right)^2, \tag{13}$$

the optimal rank estimates are

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

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

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

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

When all $w_k \equiv w$,

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

optimizes (13) subject to the R_k exhausting the integers $(1, \ldots, K)$. To see this, if $0 \leq \mathrm{E}(R_i) = m_i \leq \mathrm{E}(R_j) = m_j$, $r_i < r_j$, then

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

and the $\hat{R}_k$ are optimal. □

For general w_k there is no closed form solution, but the following sorting-based algorithm (letting $\omega_{ji} = \frac{\omega_j}{\omega_i}$) based on:

$$\begin{aligned} & (m_i - r_i)^2 + \omega_{ji}(m_j - r_j)^2 < (m_i - r_j)^2 + \omega_{ji}(m_j - r_i)^2, \text{if } r_j > r_i \\ \iff & (r_j - r_i)((1 - \omega_{ji})(r_i + r_j - 2m_j) + 2(m_j - m_i)) > 0, \text{if } r_j > r_i \\ \iff & (1 - \omega_{ji})(r_i + r_j - 2m_j) + 2(m_j - m_i) > 0, \text{if } r_j > r_i. \end{aligned} \tag{14}$$

will help to solve the problem.

Theorem 6. *Starting from any initial ranks, iteratively switch the position of unit i and unit j, $i, j = 1, ...K$, if inequality (14) is satisfied will lead to optimal ranking under the weighted square loss function (13).*

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 (14) will not be satisfied. On the other hand, the optimal ranking result exist 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 (14). This means that the optimal ranking result has the exact same order as the ranking result at stop. □

However, the convergence can be very slow. After unit i and unit j are compared and ordered, if unit i is compared to some other unit k and switch happens, then i should be compared to j again. This conclusion tell us that pairwise switch sorting based optimization algorithm is impractical in our case.

A.2 Proof of Theorem 2

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.

$$\begin{aligned} a_1b_1 + a_2(1-b_2) \leq a_1b_2 + a_2(1-b_1) &\Leftrightarrow a_1b_1 - a_2b_2 \leq a_1b_2 - a_2b_1 \\ &\Leftrightarrow (a_1+a_2)b_1 \leq (a_1+a_2)b_2. \end{aligned}$$

□

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:

$$a_1b_2 + a_2b_1 \leq a_1b_1 + a_2b_2 \Leftrightarrow (a_2 - a_1)(b_2 - b_1) \geq 0.$$

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. □

Continuing the proof, let

$$\Phi = \{R^{est} = (R_1^{est}, ...R_K^{est}) : \text{ any permutation of 1 to K }\}$$

and denote by

$$\{j_1, j_2, ...j_K\} = \underset{R^{est} \in \Phi}{\operatorname{argmin}} \; E(L_{R_K^{est}}(\gamma, p, q, c))$$

the optimum ranking. Let $R_{(i)} = R_k$ if $j_k = i$ (i.e., i is the optimum rank of the k^{th} unit). Then,

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

For optimum ranking, the following conditions are necessary.

1. By Lemma 1, for any (i_1, i_2) satisfying $(1 \le i_1 \le [\gamma(K+1)]$, $[\gamma(K+1)]+1 \le i_2 \le K)$, it is required that $pr(R_{(i_1)} \ge \gamma(K+1)) \le pr(R_{(i_2)} \ge \gamma(K+1))$. To satisfy this condition, divide the units into two groups by picking the largest $K - [\gamma(K+1)]$ from $\{pr(R_k \ge \gamma(K+1)) : k = 1, ...K\}$ 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 $pr(R_{(i_1)} \ge \gamma(K+1)) \ge pr(R_{(i_2)} \ge \gamma(K+1))$ if $i_1 > i_2$. Therefore, for the units with ranks $(1, \ldots \gamma K)$, the ranks should be determined by ranking the $pr(R_k \ge \gamma(K+1))$.

   (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 $pr(R_{(i_1)} \ge \gamma(K+1)) \ge pr(R_{(i_2)} \ge \gamma(K+1))$ if $i_1 > i_2$. Therefore, for the units with ranks $(\gamma K + 1, \ldots, K)$, the ranks should be determined by ranking the $pr(R_k \ge \gamma(K+1))$.

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 $pr(R_k \ge \gamma(K+1))$.

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 coordinates whose ranks are estimated as i_1, i_2, where $i_1 < \gamma(K+1), i_2 > \gamma(K+1)$. Let

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

For the index selection to be optimal,

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

The following is equivalent to the foregoing.

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

Therefore, with $p_k = pr(R_k \geq \gamma(K+1))$ 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 impractical because quantity for comparison depends on the position and thus in each iteration, all pairwise relation have to been check again. We have not identified a general representation or 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,

$$|P_k - P_k^{est}| = |P_k - \gamma| + |P_k^{est} - \gamma|.$$

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 $L^\ddagger$.

References

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

Center for Studying Health System Change (2003). *Community Tracking Study Household Survey, 2000-2001: ICPSR version.* Ann Arbor, MI: Inter-university Consortium for Political and Social Research.

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

Conlon, E. and T. Louis (1999). Addressing multiple goals in evaluating region-specific risk using Bayesian methods. In A. Lawson, A. Biggeri, D. Böhning, E. Lesaffre, J.-F. Viel, and R. Bertollini (Eds.), *Disease Mapping and Risk Assessment for Public Health*, Chapter 3, pp. 31–47. Wiley.

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

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

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.

Escobar, M. (1994). Estimating normal means with a Dirichlet process prior. *Journal of the American Statistical Association 89*, 268–277.

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

Goldstein, H. and D. Spiegelhalter (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., J. E. Littlewood, and G. Polya (1967). *Inequalitites* (2nd ed.). Cambridge University Press.

Kemper, P., D. Blumenthal, J. M. Corrigan, P. J. Cunningham, S. M. Felt, J. M. Grossman, L. Kohn, C. E. Metcalf, R. F. St. Peter, R. C. Strouse, and P. B. Ginsburg (1996, Summer). The design of the community tracking study: A longitudinal study of health system change and its effects on people. *INQUIRY 33*, 195–206.

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

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

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

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

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

McClellan, M. and D. Staiger (1999). The quality of health care providers. Technical Report 7327, National Bureau of Economic Research, Working Paper.

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

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

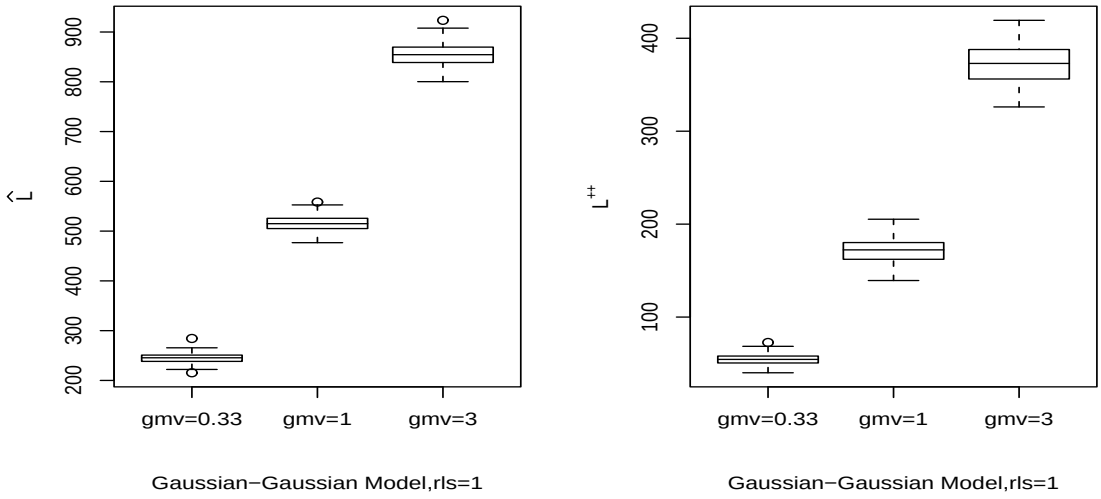

Figure 1: Unit-specific, $\hat{L}$ and $L^{\ddagger}$ performance as a function of gmv for $K = 200$, $\gamma = 0.8$ and $rls = 1$. The box plots summarize the sampling distribution of unit-specific risk.

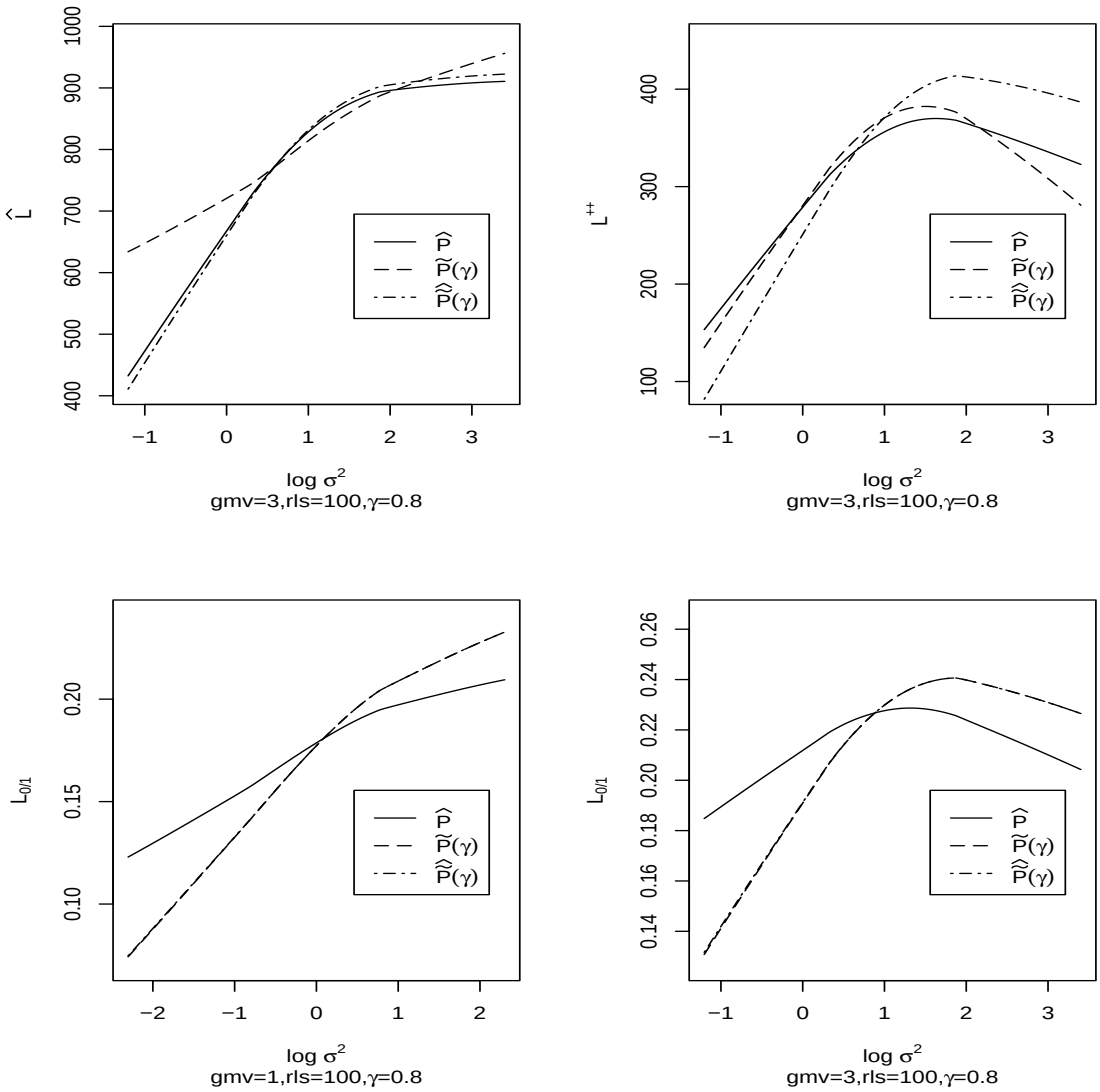

Figure 2: Loess smoothed, unit-specific $\hat{L}$, $L^{\ddagger}$, and $L_{0/1}$ performance for $\hat{P}_k$, $\tilde{P}_k(\gamma)$ and $\hat{\tilde{P}}_k(\gamma)$ as a function of unit-specific variance (σ_k^2) for $\gamma = 0.8$, $rls = 100$ and $gmv = 1$ and 3.

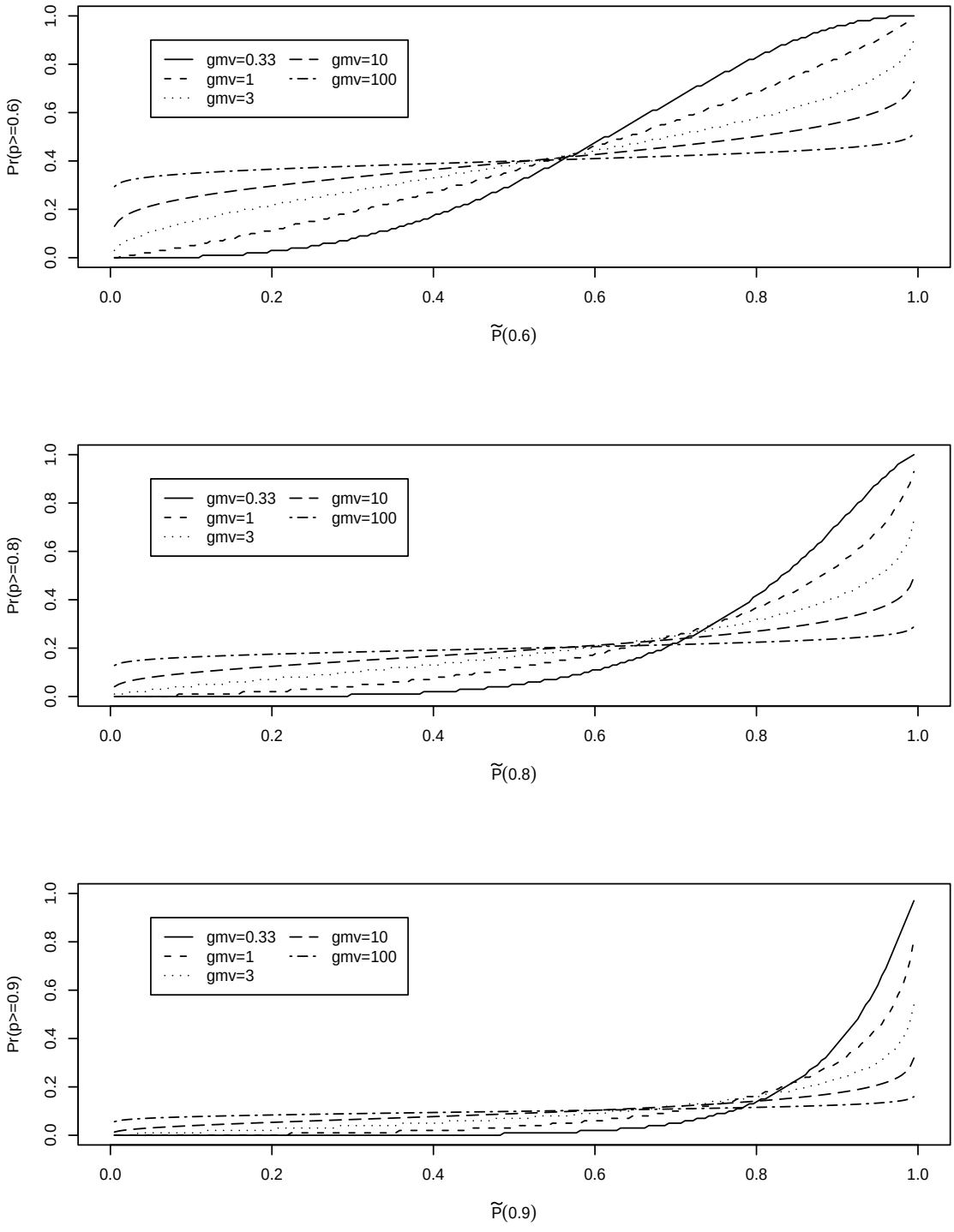

Figure 3: Average posterior classification probabilities as a function of the optimally estimated percentiles for $rls = 1$, $gmv = (0.33, 1, 3, 10, 100)$, $\gamma = (0.6, 0.8, 0.9)$.

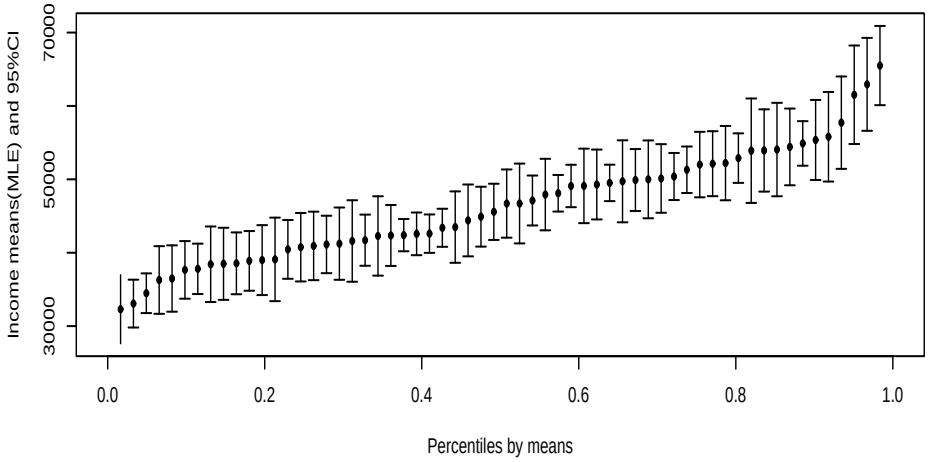

Figure 4: "Caterpillar" plot of mean income and 95% confidence intervals for the 60 CTS communities.

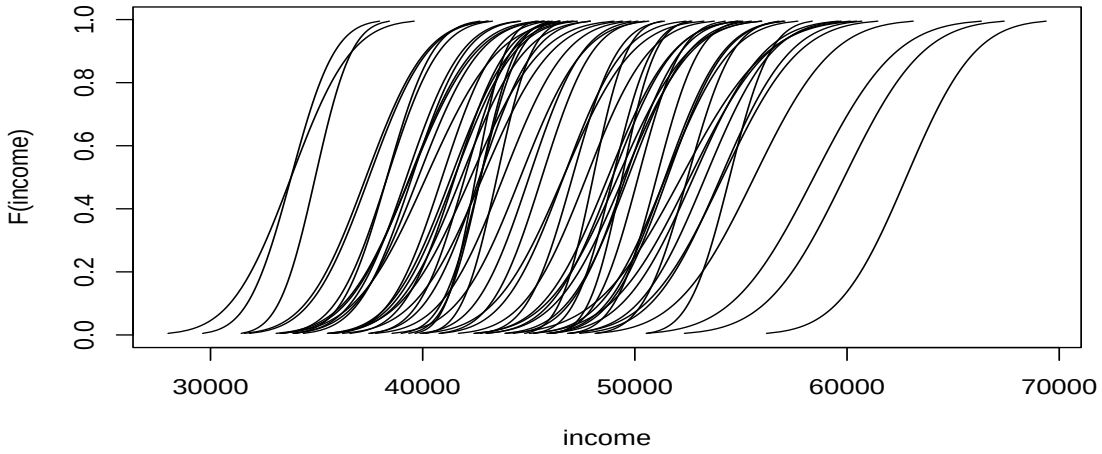

Figure 5: Posterior cumulative income distributions for the 60 CTS communities.

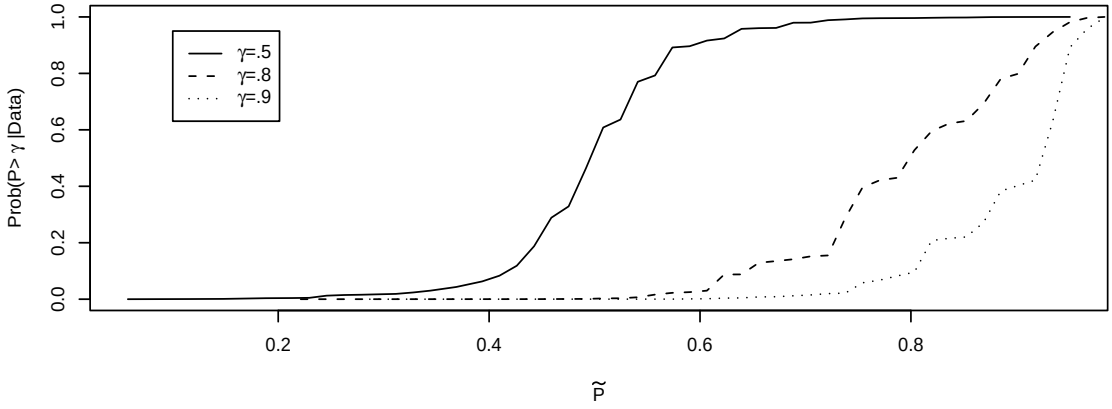

Figure 6: Posterior probability of being in the upper $(1-\gamma)$ income percentile region, $\gamma = 0.5, 0.8, 0.9$ for the 60 CTS communities.

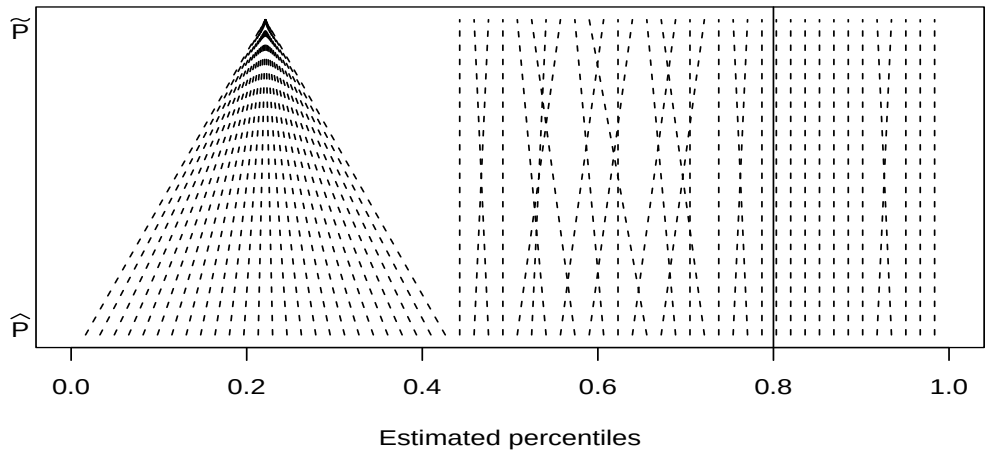

Figure 7: Comparison plot of income-based $\hat{P}_k$, $\tilde{P}_k(\gamma)$ and $\hat{\tilde{P}}_k(\gamma)$ for the 60 CTS communities. $\hat{P}_k$ and $\hat{\tilde{P}}_k(\gamma)$ are identical and plotted together.