Ranking Methods that Address Specific Goals

Thomas A. Louis, PhD, Research & Methodology Directorate
U. S. Census Bureau, thomas.arthur.louis@census.gov
Biostatistics, Johns Hopkins Bloomberg School of Public Health
Baltimore MD USA, tlouis@jhsph.edu

Abstract
Allocation of funds to high poverty regions, development of language-specific ballots, prioritization of public health interventions and environmental remediations in small areas, profiling health service providers, comparative evaluation of school effectiveness, gene and SNP identification studies; all depend on the relative position (ranks) of unit-specific parameters. Invalid or sub-optimal ranks can have serious scientific, policy and financial consequences. Ranking is challenging because intuitive approaches tend to perform poorly. Happily, Bayesian structuring coupled with a ranking-specific loss function has proven very effective. We outline the Bayesian approach, discuss simulation evaluations, and analyze Standardized Mortality Ratios.

Key Words: Bayesian Modeling, Rank-targeted Loss, Triple-goal Estimates

1 Introduction
The inferential goal is to rank underlying parameters based on observed data. Ranking is challenging because relatively high variance MLEs tend to be extreme, while Z-scores testing that a unit’s underlying parameter is equal to the typical value tend to be relatively close to zero for these units. The Bayesian approach guided by a ranking-specific loss function provides the necessary structure to strike an effective compromise. Recent articles on ranking include, Dyer and Owen (2012); Gelman and Price (1999); Lockwood et al. (2002); Normand and Shaian (2007); Ohlssen et al. (2007); Shen and Louis (1998).

2 Loss Function Based Ranking
Consider the two-stage, hierarchical model with a continuous prior distribution. For \( k = 1, \ldots, K \),

\[
\begin{align*}
\theta_k & \sim \text{iid } G; \quad [Y_k | \theta_k] \sim f_k(Y_k | \theta_k) \\
[\theta_k | Y_k] & \sim \int f_k(Y_k | \theta_k)g(\theta_k) f_k(Y_k | \theta_k) g(u) du
\end{align*}
\]

The ranks, \( R = (R_1, \ldots, R_K) \) are: \( R_k = \text{rank}(\theta_k) = \sum_{j=1}^{K} I_{\{\theta_k \geq \theta_j\}} \), where \( I_{\{\cdot\}} \) is the indicator function. The smallest \( \theta \) has rank 1. To keep all estimates in the interval (0, 1) we use percentiles, with \( P_k = R_k/(K+1) \). The \( R_k \) and \( P_k \) are invariant under a monotone transform of the \( \theta_k \) and estimated ranks should also be invariant.

Squared-error loss (SEL) minimizing ranks
The posterior mean of the target quantity minimizes SEL, producing \( \hat{R}_k = E[R_k | Y] = \sum_{j=1}^{K} P[\theta_k \geq \theta_j | Y] \), \( \hat{P}_k = \hat{R}_k/(K+1) \). The \( \hat{R}_k \) are not integers and do not span the full range \([1, K]\). To obtain optimal integer ranks, rank the \( \hat{R}_k \), producing: \( \hat{R}_k = \text{rank}(\hat{R}_k) \), \( \hat{P}_k = \hat{R}_k/(K+1) \).

2.1 A threshold-specific loss function
Assume we want to correctly identify the top \((1 - \gamma)\) units, \( 0 < \gamma < 1 \). Lin et al. (2006) propose using a loss function that penalizes for misclassification,

\[
OC_{\text{post}}(\gamma | Y) = \frac{\text{pr}(P \leq \gamma | \hat{P}_{\text{post}} > \gamma, Y)}{\gamma}.
\]
When the data are completely uninformative, \( OC_{\text{pest}}(\gamma \mid \mathbf{Y}) = 1.0 \) and so is standardized across \( \gamma \) values. If the goal is to identify units with the largest percentiles, then the numerator is similar to the False Discovery Rate (Benjamini and Hochberg [1995], Efron and Tibshirani [2002]).

Let, \( \pi_k(\gamma \mid \mathbf{Y}) = \text{pr}(P_k > \gamma \mid \mathbf{Y}) \), then \( OC_{\text{pest}}(\gamma) \) is minimized by \( \tilde{R}_k(\gamma) = \text{rank}(\pi_k(\gamma)) \), \( \hat{P}_k(\gamma) = \tilde{R}_k(\gamma)/(K + 1) \), but computing the \( \pi_k \) is extremely computationally intensive. Fortunately, as Lin et al. (2006) show, the \( \tilde{R}_k(\gamma) \) are virtually identical to ranks based on exceedance probabilities \( \text{pr}(\theta_k > t \mid \mathbf{Y}) \) with \( t_\gamma = \bar{G}_K^{-1}(\gamma) \), where \( \bar{G}_K(t) = \frac{1}{K} \sum_k \text{pr}(\rho_k \leq t \mid \mathbf{Y}) \). Normand et al. (1997) rank providers using exceedance probabilities, and Diggle et al. (2007) use them to identify the areas with elevated disease rates.

For any percentiling method, \( OC_{\text{pest}}(\gamma \mid \mathbf{Y}) \) provides a data analytic performance evaluation, computed by summing \( \pi_k(\gamma \mid \mathbf{Y}) = \text{pr}(P_k > \gamma \mid \mathbf{Y}) \) over the set of indices for which \( P_{\text{pest}} > \gamma \). Plotting the \( \pi_k(\gamma \mid \mathbf{Y}) \) versus the \( P_{\text{pest}} \) displays percentile-specific classification performance. This plot is similar to that proposed by Pepe et al. (2008).

### 3 Gaussian-Gaussian Simulation

We use the model,

\[
\theta_k | G \sim \text{iid } N(0, 1); \quad [Y_k | \theta_k, \sigma_k^2] \sim \text{indep } N(\theta_k, \sigma_k^2)
\]

and set scenarios using variances \( \{\sigma_k^2\} \) that form an ordered, geometric sequence with \( gmv = \text{GM}(\sigma_1^2, \ldots, \sigma_K^2) \) (informativeness of the data) and \( rls = \sigma_K^2 / \sigma_k^2 \) (heterogeneity of the variances). Table 1 reports SEL performance (see Lin et al., 2006); when \( rls = 1 \) \( (\sigma_k^2 \equiv \sigma^2) \) all ranking methods have identical performance, but for \( rls > 1 \) the optimal percentiles (\( \hat{P}_k \)) perform notably better than ranking the posterior means \( \theta_k^{\text{pm}} \), ranking the posterior mean of \( e^{\theta_k} \), or ranking the \( Y_k \).

<table>
<thead>
<tr>
<th>( rls )</th>
<th>( \hat{P}_k )</th>
<th>( \theta_k^{\text{pm}} )</th>
<th>( E(e^{\theta_k} \mid Y_k) )</th>
<th>( Y_k )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>31.0</td>
<td>31.0</td>
<td>31.0</td>
<td>31.0</td>
</tr>
<tr>
<td>25</td>
<td>31.0</td>
<td>31.0</td>
<td>32.0</td>
<td>34.9</td>
</tr>
<tr>
<td>100</td>
<td>31.3</td>
<td>31.5</td>
<td>32.8</td>
<td>38.6</td>
</tr>
</tbody>
</table>

Table 1: Simulated preposterior \( 100 \times \text{SEL/1666} \) for \( gmv = 1 \).

For uninformative data \( (\sigma_k^2 \equiv \infty) \) SEL = 1666.

### 4 Ranking Standardized Mortality Ratios

The Standardized Mortality Ratio (SMR) is the ratio of observed to expected deaths. The United States Renal Data System (USRDS) produces annual estimated SMRs for several thousand dialysis centers and uses these as a quality screen (ESRD, 2000, USRDS (2005)). We report selected results for the \( K = 3173 \) dialysis centers that reported data for the four years 1998-2001 (see Lin et al., 2009, for full details). Let \( (Y_{kt}, m_{kt}) \) be the observed and case-mix adjusted, expected deaths for provider \( k \) in year \( t, k = 1, \ldots, 3173, t = 0, 1, 2, 3 \) and \( \rho_{kt} \) be the SMR. Then,

\[
[Y_{kt} \mid m_{kt}, \rho_{kt}] \sim \text{Poisson}(\rho_{kt}m_{kt})
\]

\[
E(Y_{kt} \mid m_{kt}, \rho_{kt}) = m_{kt}\rho_{kt} = m_{kt} \times e^{\theta_{kt}}
\]

“Average performance” is equivalent to \( \rho_{kt} = 1, \theta_{kt} = 0 \).
4.1 Single-year Analyses
For year \( t \), \( \theta_{kt} \stackrel{\text{iid}}{\sim} G_t, k = 1, \ldots, 3173 \), with \( G_t \) either a normal distribution or the non-parametric maximum likelihood (NPML) prior (see Carlin and Louis, 2009; Paddock et al., 2006).

4.2 The Longitudinal, AR(1) Model
Let \( \phi = \text{cor}(\theta_{kt}, \theta_{k(t+1)}) \), with \(-1 < \phi < 1\). Then, use a normal prior on the \( \theta_{kt} \) and a normal prior on \( Z(\phi) = 0.5 \log\{(1 + \phi)/(1 - \phi)\} \) in the hierarchical model,

\[
\xi_t \stackrel{\text{iid}}{\sim} N(0, V), \quad \lambda_t = \tau_t^{-2} \stackrel{\text{iid}}{\sim} \text{Gamma}(\alpha, \mu/\alpha), \quad Z(\phi) \sim N(0, V_\phi)
\]

\[
[\theta_{t0}, \ldots, \theta_{t0}] \stackrel{\text{iid}}{\sim} N(\xi_0, \tau_0^2)
\]

\[
[\theta_{kt}|\theta_{k(t-1)}, \ldots, \theta_{kt}, \xi, \tau, \phi] \stackrel{\text{iid}}{\sim} N(\xi_t + \phi \tau_t \tau_{t-1}^{-1}\{\theta_{k(t-1)} - \xi_{t-1}\}, (1 - \phi^2)\tau_t^2)
\]

\[
[Y_{kt} | m_{kt}, \rho_{kt}] \sim \text{Poisson}(m_{kt}\rho_{kt}), \rho_{kt} = \exp(\theta_{kt})
\]

Marginally, for year \( t \), \( \theta_{kt} \stackrel{\text{iid}}{\sim} N(\xi_t, \tau_t^2) \) and setting \( \phi = 0 \) produces four, single-year analyses, each using model 2 with no borrowing of information over time. For \( \phi > 0 \), in addition to combining evidence across centers in a single year, an AR(1) model combines evidence within a dialysis center across years.

4.3 Performance measures
Using the 1998 data, we compute the relations between optimal percentiles and other candidate approaches, and compute \( OC_{\text{pest}}(0.80) \) and a measure of longitudinal variation (LV) for both single and multiple years, with,

\[
LV_{\text{pest}} = 1000 \times \frac{1}{3K} \sum_{k=1}^{K} \sum_{t=0}^{3} (P^\text{pest}_{kt} - P^\text{est}_{k*})^2,
\]

where \( P^\text{pest}_{kt} \) is the estimated percentile for dialysis center \( k \) in year \( t \) and \( P^\text{est}_{k*} \) is the mean over the four years.

4.4 Comparisons using the 1998 data
We computed ranks and percentiles based on the MLEs (\( \rho^\text{mle}_k \)), the posterior means (\( \rho^\text{pm}_k \)), Z-scores testing the hypothesis \( H_0 : \rho = 1 \) for 1998) and the optimal \( \hat{P}_k \), \( \hat{P}_k(\gamma) \). The MLEs and Z-scores are,

\[
\rho^\text{mle}_k = \frac{Y_k}{m_k}, \quad Z_k = \sqrt{m_k \log\left(\frac{y_k}{m_k} + 0.25\right)}, \quad \rho^\text{pm}_k = E(\rho_k | Y).
\]

Figure 1 displays estimates for the 40 providers at the 1/3174, 82/3174, 163/3174, \ldots, 3173/3174 percentiles as determined by \( \hat{P}_k \). For each display, the Y-axis is 100 \( \times \hat{P}_k \) with its 95% posterior interval. The X-axis for the upper left panel is \( \hat{P} \), for the upper right is percentiles based on \( \rho^\text{pm} \), for the lower left is percentiles based on \( \rho^\text{mle} \), and for the lower right is percentiles based on Z-scores testing \( \rho_k = 1 \).

In the upper left display the \( \hat{P}_k \) do not fill out the (0, 1.0) percentile range; they are shrunken toward 0.50 by an amount that reflects estimation uncertainty. Also, the posterior probability intervals are very wide, indicating considerable uncertainty in estimating ranks/percentiles. The plotted points in the upper left display are monotone because the X-axis is the percentile based on ranking of Y-axis values. Plotted points in the upper right display, which are based on posterior mean, are almost monotone and close to the best attainable. The lower left and lower right panels show considerable departure from monotonicity, indicating that MLE-based
ranks and hypothesis test-based ranks are very far from optimal. Note also that the pattern of departures is quite different in the two panels, showing that these methods produce quite different ranks. Similar comparisons for SMRs estimated from the pooled 1998-2001 data would be qualitatively similar, but the departures from monotonicity would be less extreme.

4.5 Multi-year analyses

Using model (3) we estimated single-year based and AR(1) model based percentiles. Table 2 reports that the $\xi$ are near 0, as should be the case since we have used internal standardization (the typical log(SMR) = 0). The within year, between provider variation in 100 $\times$ log(SMR) is essentially constant at approximately 100 $\times$ $\tau$ = 24, producing a 95% a priori interval for the $\rho_{kt}$ (0.62, 1.60). While we have a prior centering around 1000 for 100 $\times$ $\tau$, the data likelihood dominates the prior information and the posterior 95% credible interval of 100 $\times$ $\tau$ for all 4 years is (22.8, 26.8). Use of the AR(1) model to combine evidence over years (with the posterior distribution for $\phi$ concentrated around 0.90) reduces 100 $\times$ $OC_{\hat{P}}$ from around 61 to around 48, a 20% decrease. Classification performance comparison using the $\hat{P}_k$ is very close to that for the optimal 100 $\times$ $\hat{P}$ (0.8). Longitudinal variation in ranks/percentiles ($LV_P$est) is dramatically reduced for the AR(1) model going from 62 for the year-by-year analysis to 4 for the multi-year. As a basis for comparison, if $\phi \rightarrow 1$, $LV_P \rightarrow 0$ and if the data provide no information on the SMRs (the $\tau \rightarrow \infty$), then $LV_P = 83$.

Figure 2 displays two classification curves. In the upper range of $\hat{P}$ (0.8), the curve for the AR(1) model lies above that for the single year, in the lower range it lies below. For the AR(1) model to dominate the single year at all values of $\hat{P}$ (0.8), the curves would need to cross at $\hat{P}$ (0.8), but the curves cross at about 0.7. Performance when the NPML replaces the Gaussian distribution for the $\theta$ very similar. Interestingly, the NPML is bimodal with a secondary mode at $\theta = 0.5; \rho = 1.65$, possibly indicating two populations of dialysis centers.

5 Discussion

We have outlined an approach to optimal ranking, but caution that “even the best of the breed can still be a dog.” Therefore, uncertainties and performance measures {OC$_{\text{pext}}$(\$), $\pi_k(\gamma | Y)$} must also be reported. The $\pi_k(\gamma | Y)$ can be used to can be used to temper penalties or rewards. The $OC_{\hat{P}_k(\gamma)}(\gamma)$ minimizing percentiles optimize that goal, but the SEL minimizing percentiles perform well over a broad range of $\gamma$ values and so are general-purpose. The AR1 model and generalizations are very effective in combining evidence over time and can help stabilize estimates and ranks reported by the American Community Survey and other data products in which the direct information is low.

Percentiles are prima facie relative comparisons in that it is possible that all providers are doing well or that all are doing poorly; percentiles will not pick this up. In situations where normative values are available (e.g., death rates), percentiles that have a normative interpretation are attractive and those based on posterior probabilities of exceeding some threshold ($P^*(\gamma)$) provide an excellent link to a substantively relevant scale. Finally, in a policy setting it is important to report one set of estimates that can be used for a variety of purposes. The Shen-Louis (Shen and Louis, 1998) “triple goal” estimates have ranks that are optimal, produce a histogram that is optimal, and the point estimates perform nearly as well as the posterior mean. These should be given serious attention in a policy context.
References


Table 2: Results for $\tilde{P}_k$ and $\tilde{P}(0.8)$. In the multi-year section, $100 \times OC_{\tilde{P}}(0.8)$ is for the indicated year as estimated from the multi-year model and $88\,90\,92$ is a notation for posterior median 90 and 95% credible interval (88, 92).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Single Year: ($\phi \equiv 0$)</th>
<th>Multi-Year: ($100 \times \phi \sim 88,90,92$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>$100 \times \xi$</td>
<td>-2.8 -1.3 -2.3 -0.7</td>
<td>-3.1 -0.8 -1.7 -0.3</td>
</tr>
<tr>
<td>$100 \times \tau$</td>
<td>24.1 23.5 23.1 22.2</td>
<td>25.8 25.0 24.9 24.1</td>
</tr>
<tr>
<td>$100 \times OC_{\tilde{P}}(0.8)$</td>
<td>62 61 60 62</td>
<td>49 47 46 50</td>
</tr>
<tr>
<td>$\text{LV}(\tilde{P}_k)$</td>
<td>62</td>
<td>4</td>
</tr>
</tbody>
</table>

Figure 1: SEL-based percentiles for 1998. For each display, the Y-axis is $100 \times \tilde{P}_k$ with its 95% probability interval. The X-axis for the upper left panel is $\tilde{P}$, for the upper right is percentiles based on $\rho^{pm}$, for the lower left is percentiles based on the $\rho^{mle}$ and for the lower right is percentiles based on Z-scores testing $\rho_k = 1$.

Figure 2: $\pi_k(0.8 \mid Y)$ versus $\tilde{P}_k(0.8)$ for 1998. Optimal percentiles and posterior probabilities computed with the single year model ($\phi \equiv 0$) and the $AR(1)$ model ($\hat{\phi} = 0.90$). Two curves don’t cross at $\gamma = 0.8$. The line for fully informative data, i.e., when there is no uncertainty associated with ranking results is given as reference.