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 Heath
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 Shahian (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 $k = 1, \ldots, K$,

$$\begin{aligned} \theta_k &\overset{iid}{\sim} G; \;\; [Y_k \mid \theta_k] \overset{indep}{\sim} f_k(Y_k \mid \theta_k) \\ [\theta_k \mid Y_k] &\sim \frac{f_k(Y_k \mid \theta_k) g(\theta_k)}{f_k(Y_k \mid u) g(u) du} \end{aligned}$$

The ranks, $\mathbf{R}$ $(= R_1, \ldots, R_K)$ are: $R_k = \text{rank}(\theta_k) = \sum_{j=1}^{K} I_{\{\theta_k \geq \theta_j\}}$, where $I_{\{.\}}$ is the indicator function. The smallest θ 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 θ_k and estimated ranks should also be invariant.

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

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,

$$OC_{P^{est}}(\gamma \mid \mathbf{Y}) \;\; = \;\; \frac{\text{pr}(P \leq \gamma \mid P^{est} > \gamma, \mathbf{Y})}{\gamma}.$$

When the data are completely uninformative, $OC_{P^{est}}(\gamma \mid \mathbf{Y}) = 1.0$ and so is standardized across γ 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 $OC_{P^{est}}(\gamma)$ is minimized by $\tilde{R}_k(\gamma) = \text{rank}(\pi_k(\gamma))$, $\tilde{P}_k(\gamma) = \tilde{R}_k(\gamma)/(K+1)$, but computing the π_k is extremely computer intensive. Fortunately, as Lin et al. (2006) show, the $\tilde{P}_k(\gamma)$ are virtually identical to ranks based on exceedance probabilities $\text{pr}(\theta_k > t_\gamma \mid \mathbf{Y})$ with $t_\gamma = \bar{G}_K^{-1}(\gamma)$, where $\bar{G}_K(t) = \frac{1}{K}\sum_k \text{pr}(\rho_k \le t \mid \mathbf{Y})$. 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, $OC_{P^{est}}(\gamma \mid \mathbf{Y})$ 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 $P^{est} > \gamma$. Plotting the $\pi_k(\gamma \mid \mathbf{Y})$ versus the P_k^{est} 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,

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

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 $rls = \sigma_K^2/\sigma_1^2$ (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 θ_k^{pm}, ranking the posterior mean of e^{θ_k}, or ranking the Y_k.

	percentiles based on			
rls	$\hat{P}_k$	θ_k^{pm}	$E(e^{\theta_k} \mid Y_k)$	Y_k
1	31.0	31.0	31.0	31.0
25	31.0	31.0	32.0	34.9
100	31.3	31.5	32.8	38.6

Table 1: Simulated preposterior $100\times$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 ρ_{kt} be the SMR. Then,

$$\begin{aligned} [Y_{kt} \mid m_{kt},\rho_{kt}] &\sim \text{Poisson}(\rho_{kt} m_{kt}) \\ E(Y_{kt} \mid m_{kt},\rho_{kt}) &= m_{kt}\rho_{kt} = m_{kt} \times e^{\theta_{kt}} \end{aligned} \tag{2}$$

"Average performance" is equivalent to $\rho_{kt} = 1, \theta_{kt} = 0$.

4.1 Single-year Analyses

For year t, $\theta_{kt} \overset{iid}{\sim} G_t, k = 1, \ldots, 3173$, 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 $\phi = \text{cor}(\theta_{k,t}, \theta_{k(t+1)})$, with $-1 < \phi < 1$. Then, use a normal prior on the θ_{kt} and a normal prior on $Z(\phi) = 0.5 \log\{(1+\phi)/(1-\phi)\}$ in the hierarchical model,

$$\begin{aligned} \xi_t \overset{iid}{\sim} N(0, V), \qquad & \lambda_t = \tau_t^{-2} \overset{iid}{\sim} \text{Gamma}(\alpha, \mu/\alpha) \\ Z(\phi) \quad \sim \quad & N(0, V_\phi) \qquad (3) \\ [\theta_{10}, \ldots, \theta_{K0} \mid \xi_0, \tau_0] \quad \overset{iid}{\sim} \quad & N(\xi_0, \tau_0^2) \\ [\theta_{kt} | \theta_{k(t-1)}, \ldots, \theta_{k0}, \xi, \tau, \phi] \quad \overset{ind}{\sim} \quad & N\left(\xi_t + \phi\tau_t\tau_{t-1}^{-1}\{\theta_{k(t-1)} - \xi_{t-1}\}, \{1-\phi^2\}\tau_t^2\right) \quad (4) \\ [Y_{kt} \mid m_{kt}, \rho_{kt}] \quad \sim \quad & \text{Poisson}(m_{kt}\rho_{kt}), \rho_{kt} = \exp(\theta_{kt}). \end{aligned}$$

Marginally, for year t, θ_{kt} iid $N(\xi_t, \tau_t^2)$ 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 $OC_{P^{est}}(0.80)$ and a measure of longitudinal variation (LV) for both single and multiple years, with,

$$LV_{P^{est}} = 1000 \times \frac{1}{3K} \sum_{k=1}^{K} \sum_{t=0}^{3} (P_{kt}^{est} - P_{k\bullet}^{est})^2,$$

where P_{kt}^{est} is the estimated percentile for dialysis center k in year t and $P_{k\bullet}^{est}$ is the mean over the four years.

4.4 Comparisons using the 1998 data

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

$$\rho_k^{mle} = \frac{Y_k}{m_k}, \quad \text{Z}_k = \sqrt{m_k} \log\left(\frac{y_k}{m_k} + 0.25\right), \quad \rho_k^{pm} = E(\rho_k \mid \mathbf{Y}). \qquad (5)$$

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

In the upper left display the $\bar{P}_k$ 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 ξ 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 ρ_{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 $100 \times \tau_t$ 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 ϕ concentrated around 0.90) reduces $100 \times OC_{\tilde{p}}(0.8)$ 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 $100 \times \tilde{P}_k(0.8)$. 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 $\phi \to 1$, $LV_{\hat{P}} \to 0$ and if the data provide no information on the SMRs (the $\tau \to \infty$), then $LV_{\hat{P}} = 83$.

Figure 2 displays two classification curves. In the upper range of $\tilde{P}_k(0.8)$, 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 $\tilde{P}_k(0.8)$, the curves would need to cross at $\tilde{P}_k(\gamma) = 0.8$, but the curves cross at about 0.7. Performance when the NPML replaces the Gaussian distribution for the θ 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 $\{OC_{P^{est}}(\gamma), \pi_k(\gamma \mid \mathbf{Y})\}$ must also be reported. The $\pi_k(\gamma \mid \mathbf{Y})$ can be used to can be used to temper penalties or rewards. The $OC_{\tilde{P}_k(\gamma)}(\gamma)$ minimizing percentiles optimize that goal, but the SEL minimizing percentiles perform well over a broad range of γ 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

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. P. and Louis, T. A. *Bayesian Methods for Data Analysis, 3rd edition*. Boca Raton, FL: Chapman and Hall/CRC Press, 3nd edition (2009).

Diggle, P. J., Thomson, M. C., Christensen, O. F., Rowlingson, B., Obsomer, V., Gardon, J., Wanji, S., Takougang, I., Enyong, P., Kamgno, J., Remme, J. H., Boussinesq, M., and Molyneux, D. H. "Spatial modelling and the prediction of Loa loa risk: decision making under uncertainty." *Ann Trop Med Parasitol*, 101(6):499–509 (2007).

Dyer, J. S. and Owen, A. B. "Correct Ordering in the Zipf?Poisson Ensemble." *Journal of the American Statistical Association*, 107:1510–1517 (2012).

Efron, B. and Tibshirani, R. "Empirical Bayes methods and false discovery rates for microarrays." *Genetic Epidemiology*, 23:70–86 (2002).

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).

Lin, R., Louis, T. A., Paddock, S. M., and Ridgeway, G. "Loss function based ranking in two-stage, hierarchical models." *Bayesian Analysis*, 1(4):915–946 (2006).

—. "Ranking of USRDS, provider-specific SMRs from 1998-2001." *Health Services and Outcomes Reseearch Methodology*, 8:22–38 (2009).

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).

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

Normand, S.-L. T. and Shahian, D. M. "Statistical and clinical aspects of hospital outcomes profiling." *Statistical Science*, 22:206–226 (2007).

Ohlssen, D. I., Sharples, L. D., and Spiegelhalter, D. J. "A hierarchical modelling framework for identifying unusual performance in health care providers." *Journal of the Royal Statistical Society, Series A: Statistics in Society*, 170(4):865–890 (2007).

Paddock, S., Ridgeway, G., Lin, R., and Louis, T. A. "Flexible distributions for triple-goal estimates in two-stage hierarchical models." *Computational Statistics & Data Analysis*, 50/11:3243–3262 (2006).

Pepe, M. S., Feng, Z., Huang, Y., Longton, G., Prentice, R., Thompson, I. M., and Zheng, Y. "Integrating the predictiveness of a marker with its performance as a classifier." *Am J Epidemiol*, 167(3):362–368 (2008).

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).

USRDS. "2005 Annual Data Report: Atlas of end-stage renal disease in the United States." Technical report, Health Care Financing Administration (2005).

	Single Year: ($\phi \equiv 0$)				Multi-Year: ($\boldsymbol{100 \times \phi \sim {}_{88}90_{92}}$)			
Parameter	1998	1999	2000	2001	1998	1999	2000	2001
$100 \times \xi$	-2.8	-1.3	-2.3	-0.7	-3.1	-0.8	-1.7	-0.3
$100 \times \tau$	24.1	23.5	23.1	22.2	25.8	25.0	24.9	24.1
$100 \times OC_{\tilde{P}(0.8)}(0.8)$	62	61	60	62	49	47	46	50
LV($\hat{P}_k$)	62				4			

Table 2: Results for $\hat{P}_k$ 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).

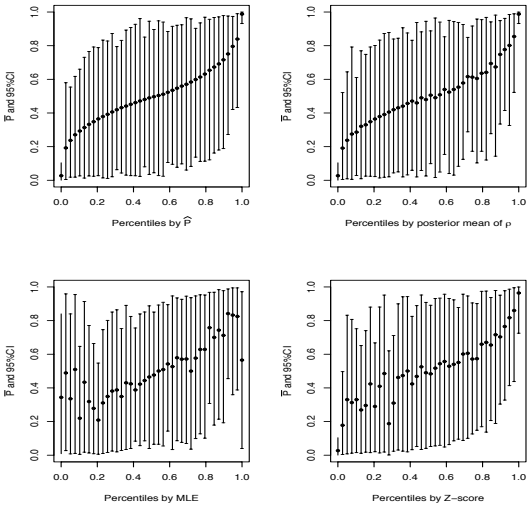

Figure 1: SEL-based percentiles for 1998. For each display, the Y-axis is $100 \times \bar{P}_k$ with its 95% probability interval. The X-axis for the upper left panel is $\hat{P}$, for the upper right is percentiles based on ρ^{pm}, for the lower left is percentiles based on the ρ^{mle} and for the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

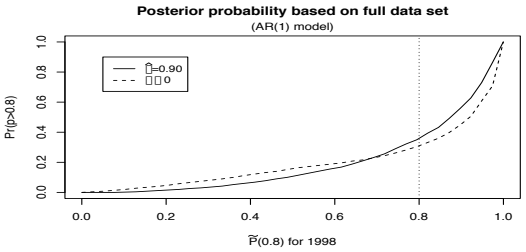

Figure 2: $\pi_k(0.8 \mid \mathbf{Y})$ versus $\tilde{P}_k(0.8)$ for 1998. Optimal percentiles and posterior probabilities computed with the single year model ($\phi \equiv 0$) and the $AR(1)$ model ($\hat{\phi} = 0.90$). Two curves don't cross at $\gamma = 0.8$. The line for fully informative data,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