Bayesian Ranking and Ensemble Estimates

Thomas A. Louis
Department of Biostatistics
Johns Hopkins Bloomberg SPH

PROFILING
(League Tables)

- The process of comparing "units" on an outcome measure with relative or normative standards
  - Quality of care, use of services, cost
  - Educational quality
  - Disease rates in small areas
  - Gene expression
  - "Best of breed" livestock

- Developing and implementing performance indices to compare physicians, hospitals, schools, teachers, genes, ...

U. S. RENAL DATA SYSTEM
STANDARDIZED MORTALITY RATIO

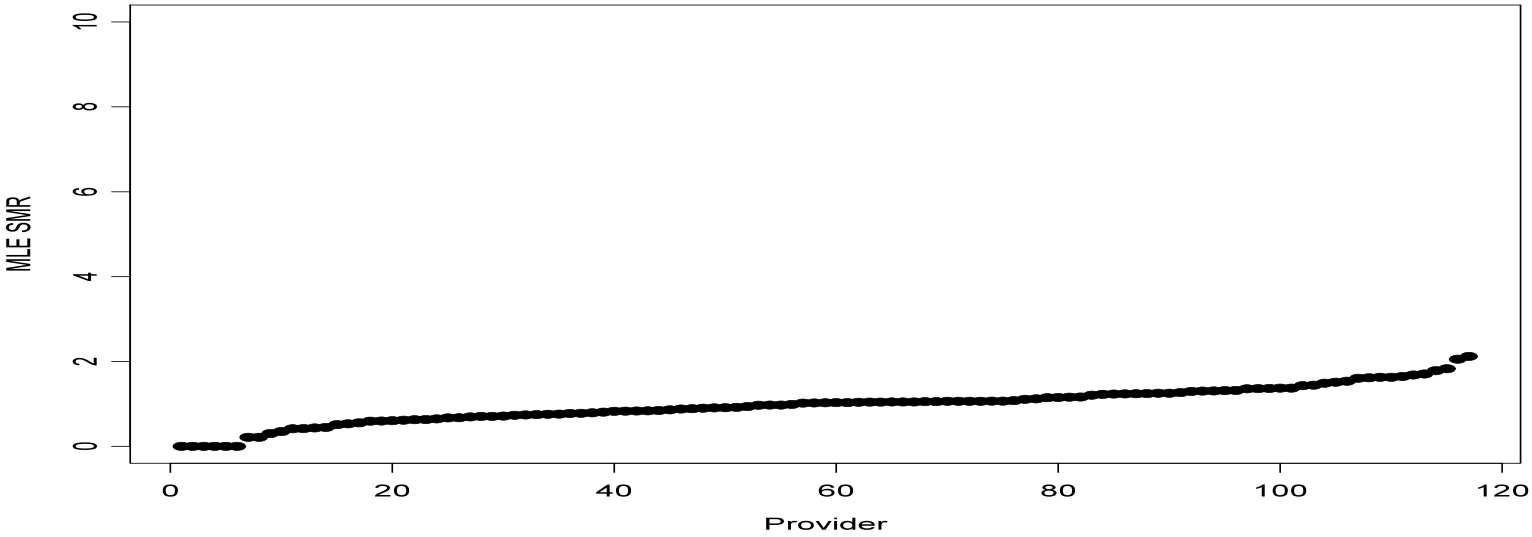
$$\text{SMR} = \frac{\text{observed deaths}}{\text{expected deaths}}$$

- Expecteds from a case mix adjustment model
- Rank 3459 dialysis providers using 1998 USRDS data
- Large and small providers, so standard errors of the estimated SMRs vary considerably

RANKING IS EASY

Just compute estimates & order them

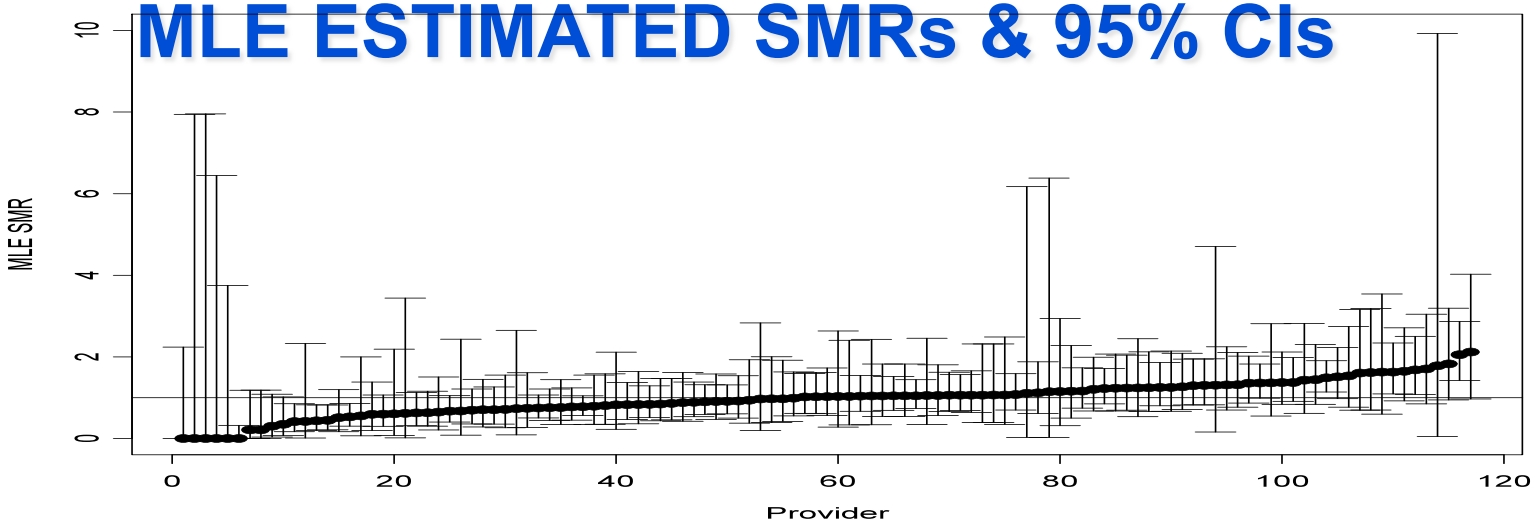
MLE ESTIMATED SMRs



RANKING IS DIFFICULT

Need to trade-off the estimates and uncertainties

MLE ESTIMATED SMRs & 95% CIs



Comparing performance measures

Ranks/percentiles, of:

- Direct estimates (MLEs)
- Shrunken estimates (BLUPs, Posterior Means)
- Z-scores testing H_0 that a unit is just like others
- Optimal (best) ranks or percentiles

THE RANKING CHALLENGE

- Ranking estimated SMRs is inappropriate, if the SEs vary over providers
  - Unfairly penalizes or rewards providers with relatively high variance
- Hypothesis test based ranking

  $$H_0 : \text{SMR}_{unit} = 1$$

  - Unfairly penalizes or rewards providers with relatively low variance
- Need to trade-off signal and noise
- However, even the optimal estimates can perform poorly

USRDS

Table 1: Percentages of providers in small MLE SMR category (the bottom 25%), moderate MLE SMR category (the middle 50%), and large MLE SMR category (the upper 25%) for each provider stratum with strata defined by total follow-up patient-years.

Provider Size	Lower 25%	Middle 50%	Upper 25%	# of providers
< 10	60	10	30	429
10-20	30	32	38	367
20-30	25	43	31	417
30-40	19	54	28	340
40-50	28	57	19	350
50-60	19	62	19	290
60-70	17	60	23	254
70-80	17	59	24	204
80-90	15	66	20	151
90-100	15	68	17	139
100-130	11	74	15	235
130-170	12	66	21	161
170-250	12	74	14	73
≥ 250	0	87	13	15

USRDS

Table 2: Percentages of providers in small Z-score category (the bottom 25%), moderate Z-score category (the middle 50%), and large Z-score category (the upper 25%) for each provider stratum with strata defined by total follow-up patient-years.

Provider Size	Lower 25%	Middle 50%	Upper 25%
< 10	13	68	19
10-20	22	50	29
20-30	26	47	27
30-40	21	52	27
40-50	31	49	20
50-60	25	53	22
60-70	27	47	26
70-80	28	46	26
80-90	26	50	25
90-100	29	47	23
100-130	33	37	29
130-170	34	33	33
170-250	26	48	26
$>= 250$	13	47	40

THE MULTI-LEVEL MODEL

- Repeated sampling from the prior:

$$\theta_1, \ldots \theta_K \sim G(\cdot \mid \eta)$$

- Then,

$$Y_k|\theta_k \sim f_k(y_k|\theta_k)$$

- We can add more stages representing "how we get to see the data"

$$\begin{aligned} \eta &\sim h(\cdot) \\ \theta \mid \eta &\sim g(\cdot \mid \eta) \\ Y \mid \theta &\sim f(\cdot \mid \theta) \end{aligned}$$

- Need Monte-Carlo methods to implement most analyses

Ranking USRDS Standardized Mortality Ratio (SMR)

- Around 3,100 dialysis centers data used from 1998-2001
- Patient numbers of each center from < 10 to 700
- For unit k, Y_k is the observed death number, μ_k is the expected death number, and ρ_k is SMR.

Model:

$$
\begin{aligned}
Y_k|\mu_k, \rho_k &\sim \text{Pois}(\mu_k * \rho_k) \\
\rho_k &= \exp(\theta_k) \\
\theta_k &\sim \text{N}(\xi, \lambda^{-1}) \\
\lambda &\sim \text{Gamma}(\alpha, \beta) \text{ with mean } \alpha\beta^{-1} \text{ and variance } \alpha\beta^{-2}
\end{aligned}
$$

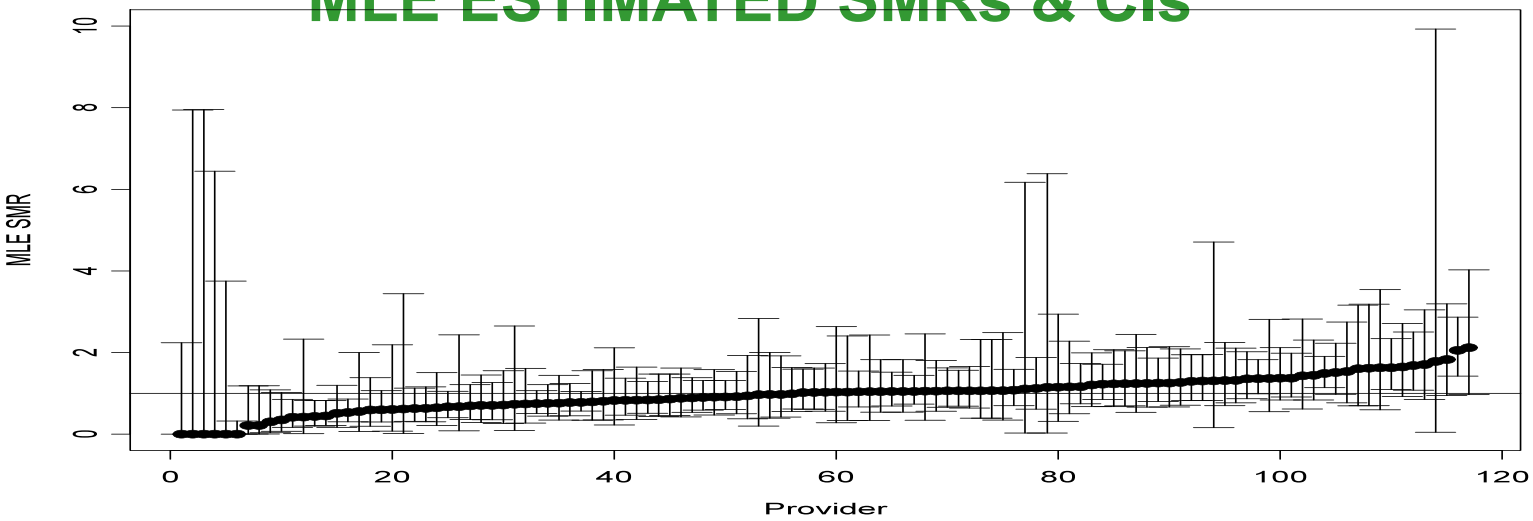
Poisson-Normal Model

(N, Y[k] , emort[k]) are inputs

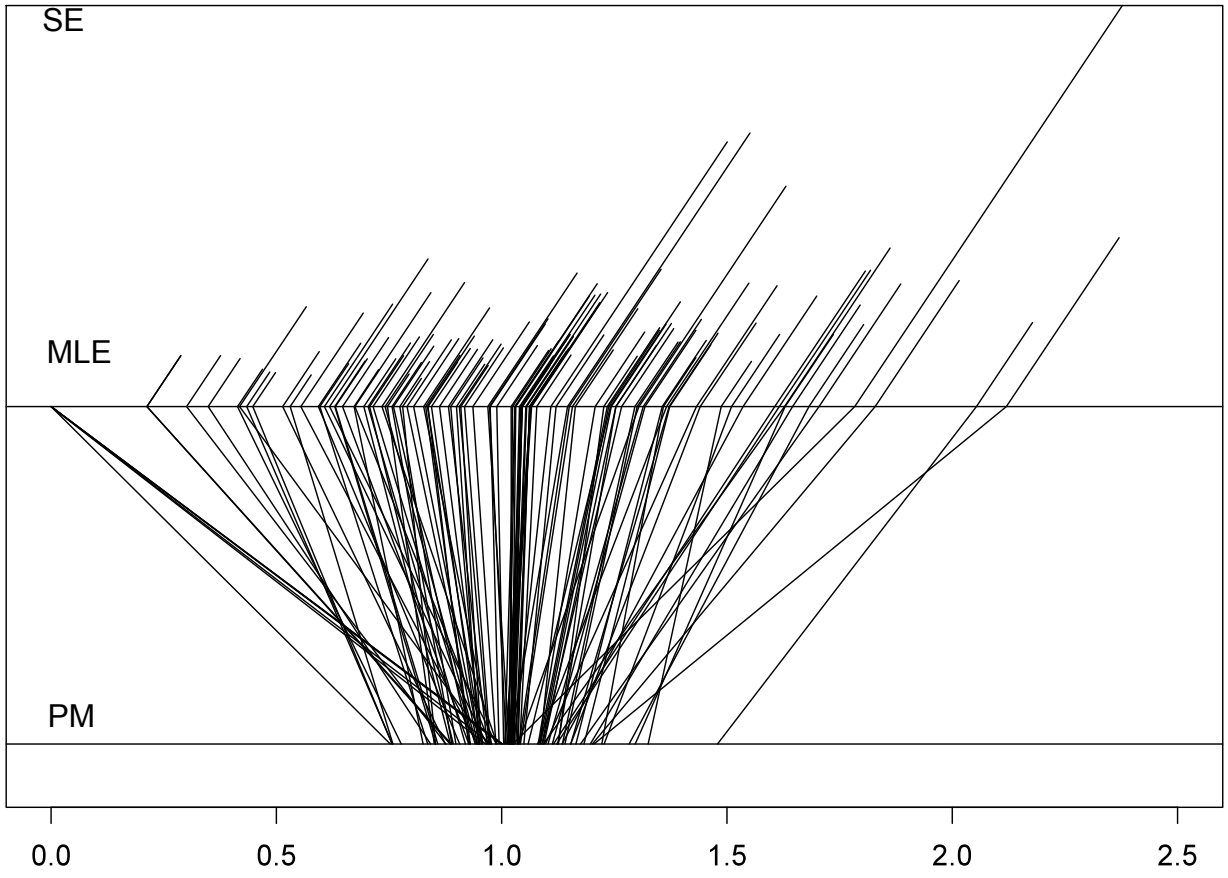
```
model
{
{prec~dgamma(0.00001,0.00001)
for (k in 1:N) {
logsmr[k]~dnorm(0,prec)
smr[k]<-exp(logsmr[k])
rate[k]<-emort[k]*smr[k]
Y[k] ~ dpois(rate[k])
}
}
```

Monitor the SMR[k]

MLE ESTIMATED SMRs & CIs

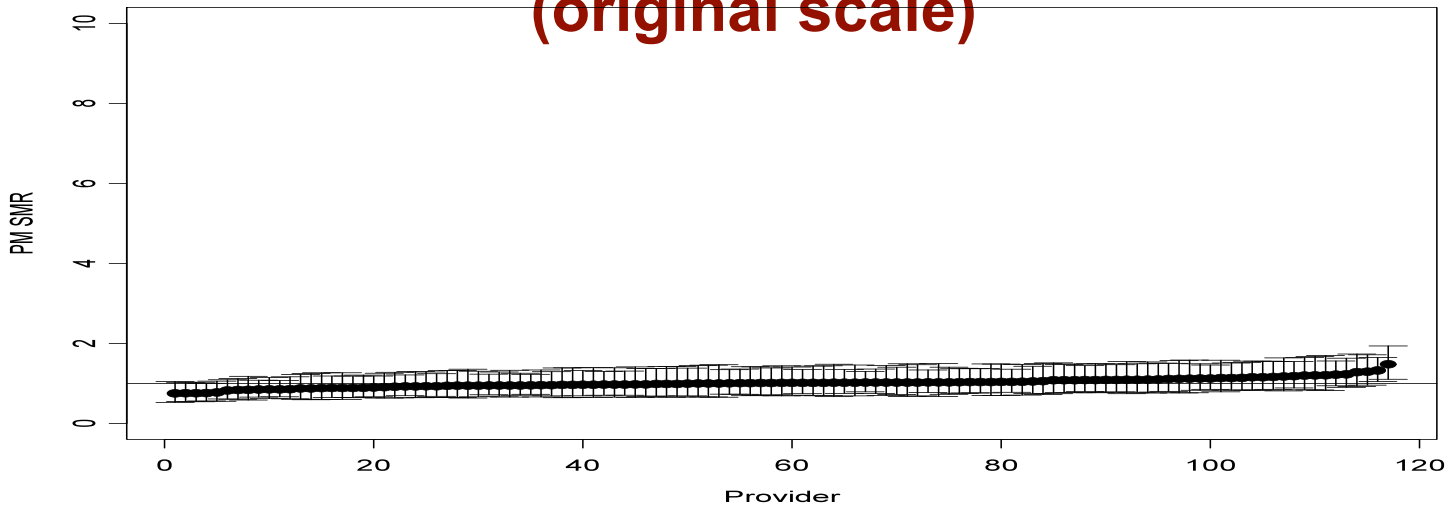


MLE, SE & POSTERIOR MEAN SMRs
(using a log-normal/Poisson model)



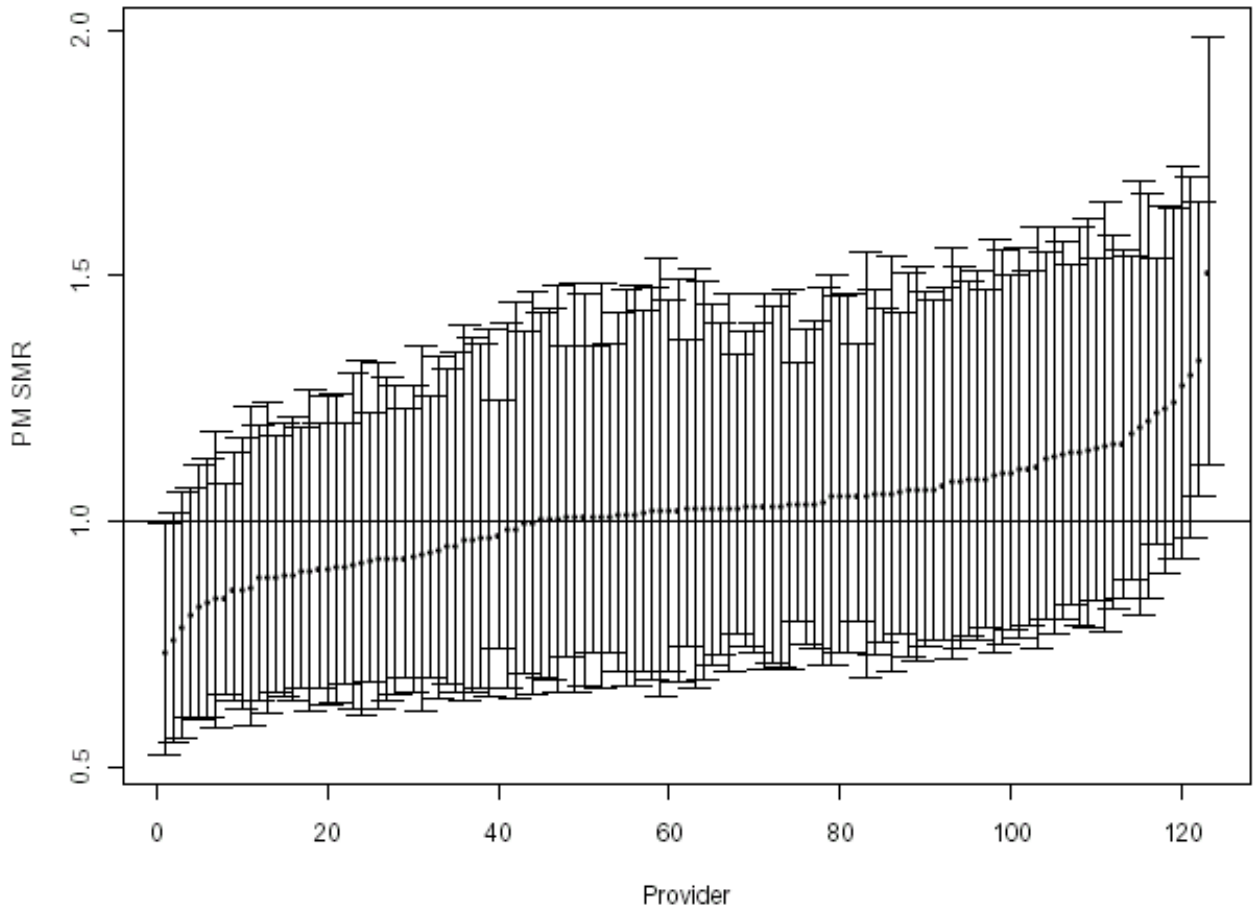
Posterior Mean: estimated SMRs & CIs

using a log-normal/Poisson model (original scale)



Posterior Mean: estimated SMRs & CIs

using a Gamma/Poisson model (expanded scale)



RANKS

Laird & Louis 1989 JES

Shen & Louis 1998 JRSS(B)

$$\theta = (\theta_1, \ldots, \theta_K)$$

$$R_k(\theta) = \sum_{\nu=1}^{K} I_{\{\theta_k \geq \theta_\nu\}}$$

- The smallest θ has rank 1.

- Expected Ranks

$$\bar{R}_k(\mathrm{Y}) = E[R_k(\theta) \mid \mathbf{Y}] = \sum_{\nu} pr[\theta_k \geq \theta_\nu \mid \mathbf{Y}]$$

$$\bar{P} = \bar{R}/(K+1)$$

- The $\bar{R}$ are not integers and move (shrink) the integers towards $(K+1)/2$
- The $\bar{P}$ shrink towards $\frac{1}{2}$
- For integer ranks, rank the $\bar{R}$ producing $\hat{R}$ and $\hat{P}$

Performance of some Ranking/percentiling methods (Gaussian/Gaussian model)

Constant variance: ranks the same for all methods

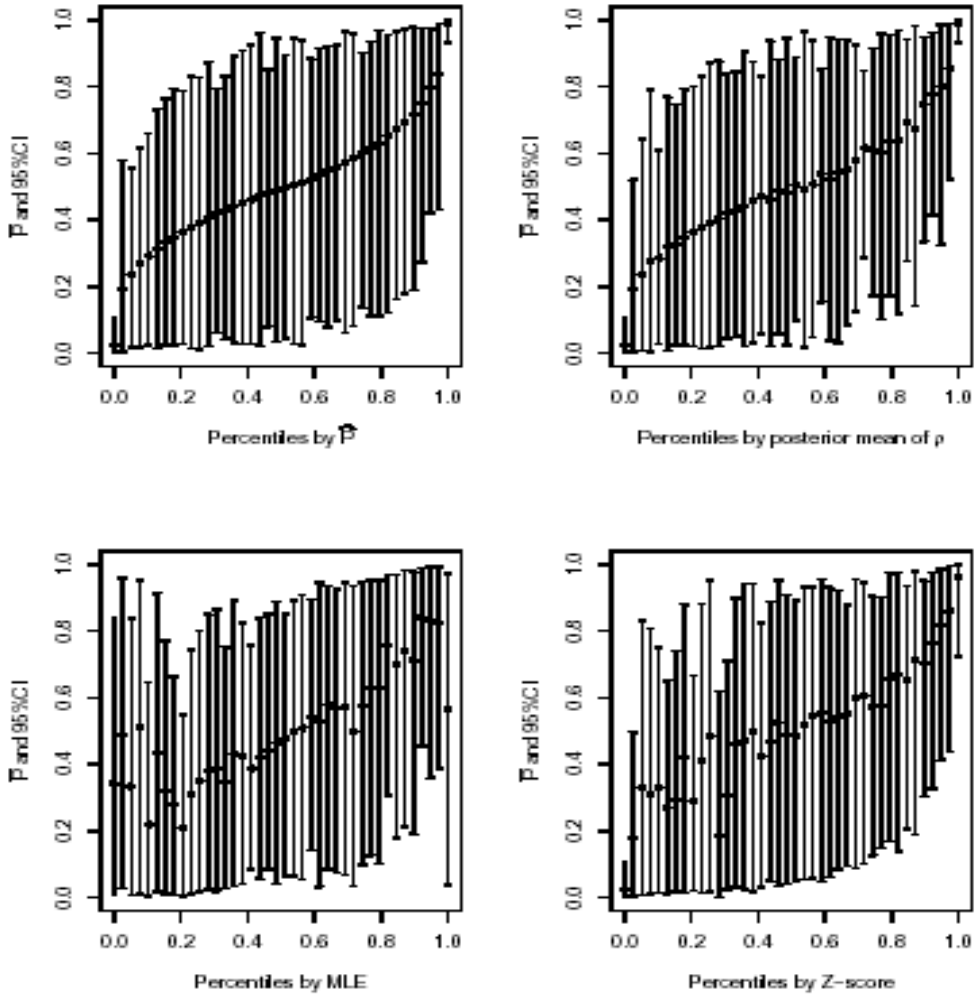
	percentiles based on			
rls	$\hat{P}_k$	θ_k^{pm}	$\exp\left\{\theta_k^{pm} + \frac{(1-B_k)\sigma_k^2}{2}\right\}$	Y_k
1	516	516	516	516
25	517	517	534	582
100	522	525	547	644

Sampling Variances are a geometric series with rls = Largest/Smallest

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

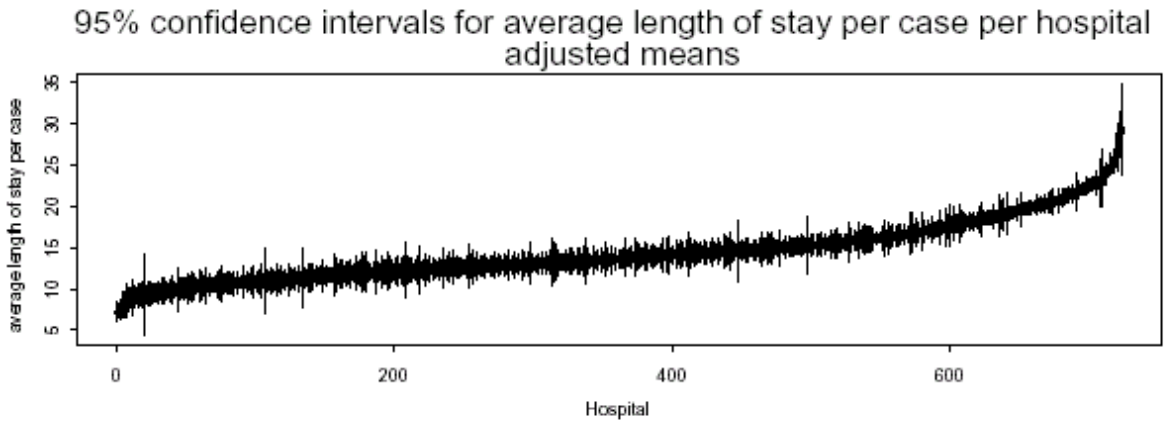
Relations among percentiling methods

1998 USRDS Percentiles



LENGTH OF STAY

- LOS (in days) from 724 rehabilitation hospitals; 255,018 discharges
  - We assume a 10% sample of discharges; 25,502 discharges
- Analysis adjusts for case-level severity of illness covariates
  - A "reason for stay" categorical variable
  - Comorbidity status
- Estimates
  - Within-hospital SE (σ) ranges from 0.55 to 8.74, median = 1.73
  - $\mu = 14.5$; $\tau = 3.1$



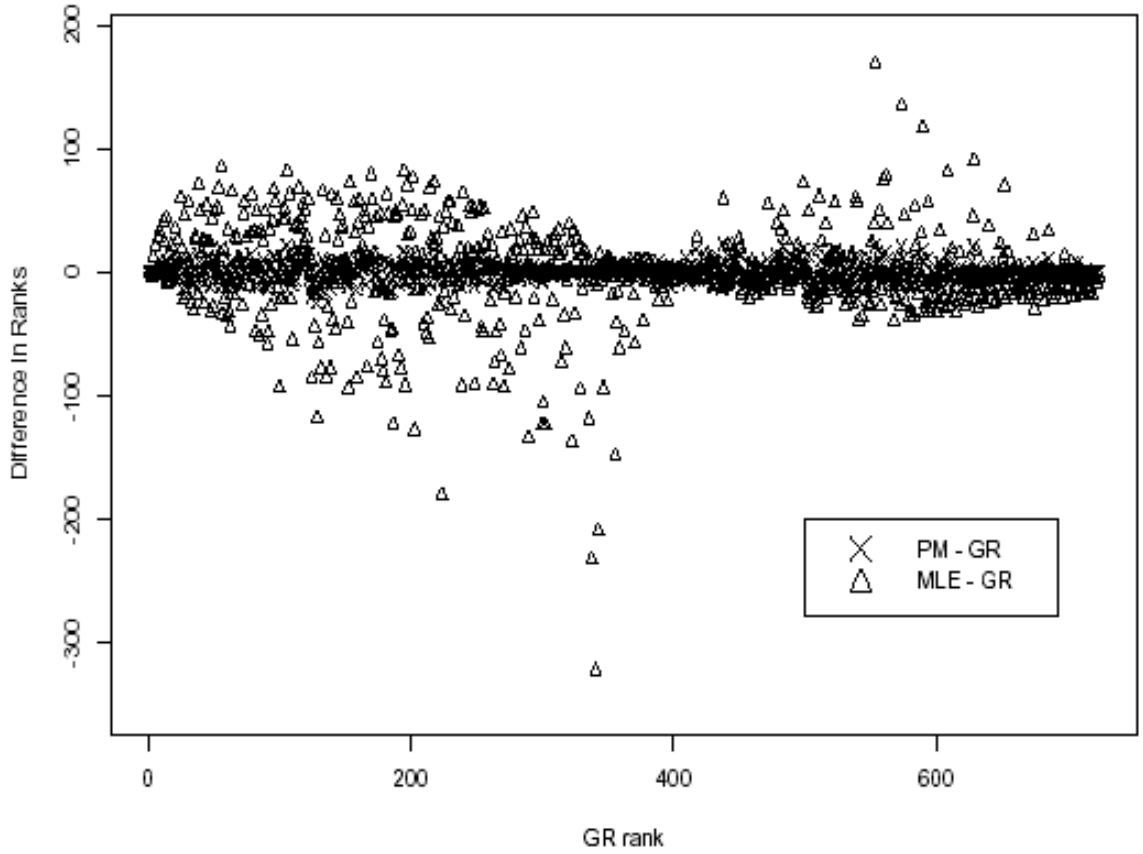
MLE and 95% CI for the LOS data; full sample per hospital

For the 10% sample, CIs would be 3.2 ($= \sqrt{10}$) times wider

LOS

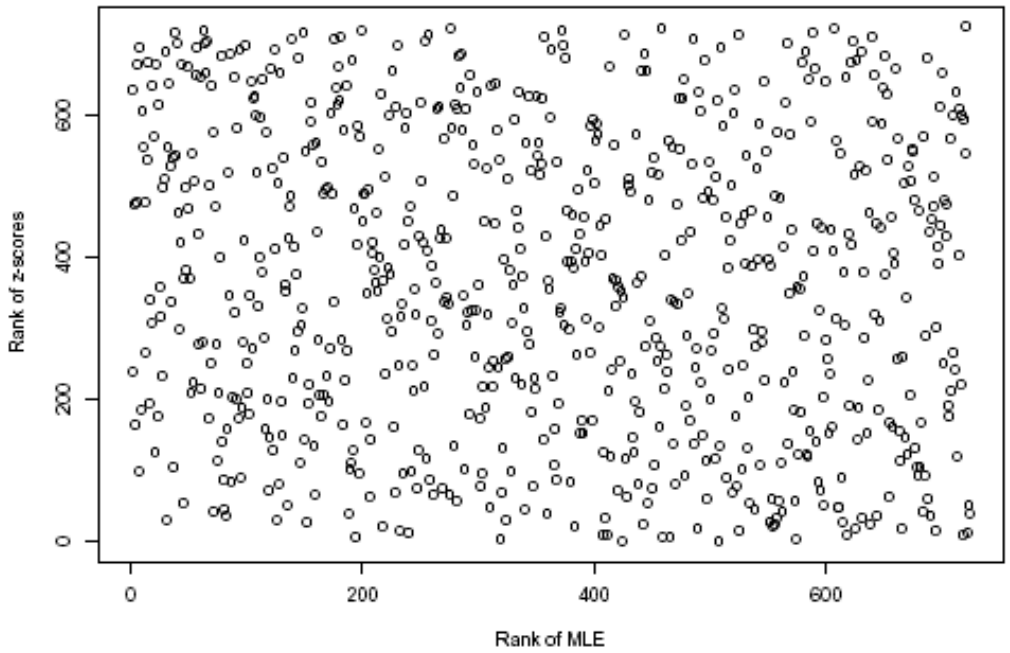
X = (Posterior Mean-Based Ranks) – (Optimal Ranks)

Δ = (DIRECT ESTIMATE RANKS) - (OPTIMAL RANKS)



LOS

Ranks computed from Z-scores versus Ranks computed from MLEs



Bayesian ranking outperforms Z-score based ranks
($K = 2$ example)

Data: $(\hat{\theta}_1, \sigma_1^2), (\hat{\theta}_2, \sigma_2^2)$

Z-score: $Z_k = \hat{\theta}_k / \sigma_k$

Bayes: $[\theta_k \mid \hat{\theta}_k, \sigma_k] \sim N(\hat{\theta}_k, \sigma_k^2)$

- Compute, $pr(\theta_1 > \theta_2 \mid \text{data})$
- $R_1 = 2 \iff$ this probability is $> \frac{1}{2}$
- Distance from $\frac{1}{2}$ measures degree of separation

Z-score Performance

$$\mathbf{Z_1 > Z_2 \iff \sigma_2\hat{\theta}_1 - \sigma_1\hat{\theta}_2 > 0}$$

- If $\hat{\theta}_1 = \hat{\theta}_2$, $\iff$ $\sigma_2 > \sigma_1$
- Can have $\hat{\theta}_1 << \hat{\theta}_2$, but $Z_1 > Z_2$
- "Pre-posterior"

$$pr(Z_1 > Z_2 \mid \theta_1, \sigma_1; \theta_2, \sigma_2) = \Phi\left(\frac{\sigma_2\theta_1 - \sigma_1\theta_2}{\sigma_1\sigma_2\sqrt{2}}\right)$$

- If $\theta_1 = \theta_2 = \theta$,

$$= \Phi\left(\theta\frac{\{\sigma_2 - \sigma_1\}}{\sigma_1\sigma_2\sqrt{2}}\right)$$

Bayes Ranks

$$pr(\theta_1 > \theta_2 \mid \text{data}) = \Phi\left(\frac{\hat{\theta}_1 - \hat{\theta}_2}{(\sigma_1^2 + \sigma_2^2)^{\frac{1}{2}}}\right)$$

$$E\{pr(\theta_1 > \theta_2 \mid \text{data})\} = \Phi\left(\frac{\theta_1 - \theta_2}{(\sigma_1^2 + \sigma_2^2)^{\frac{1}{2}}}\right)$$

- Each equals $1/2$, if the (hat) θs are equal
- Each is $> 1/2$, if (hat) $\theta_1 > \theta_2$

Extreme case: $\sigma_1 <<< \sigma_2$ & $\hat{\theta}$s > 0

- The Z-score approach will produce $R_1 = 2$ **unless $\hat{\theta}_2 >>>> \hat{\theta}_1$**
- The Bayesian approach will produce $R_1 = 2$ $\iff \hat{\theta}_1 > \hat{\theta}_2$
  - And also will provide the degree of separation

False detection and non-detection

OPERATING CHARACTERISTIC

$$OC(\gamma) = pr(P_k < \gamma | P_k^{est} > \gamma, \mathrm{Y}) + pr(P_k > \gamma | P_k^{est} < \gamma, \mathrm{Y})$$

$$OC(\gamma) = \frac{pr(P_k > \gamma | P_k^{est} < \gamma, \mathrm{Y})}{\gamma}$$

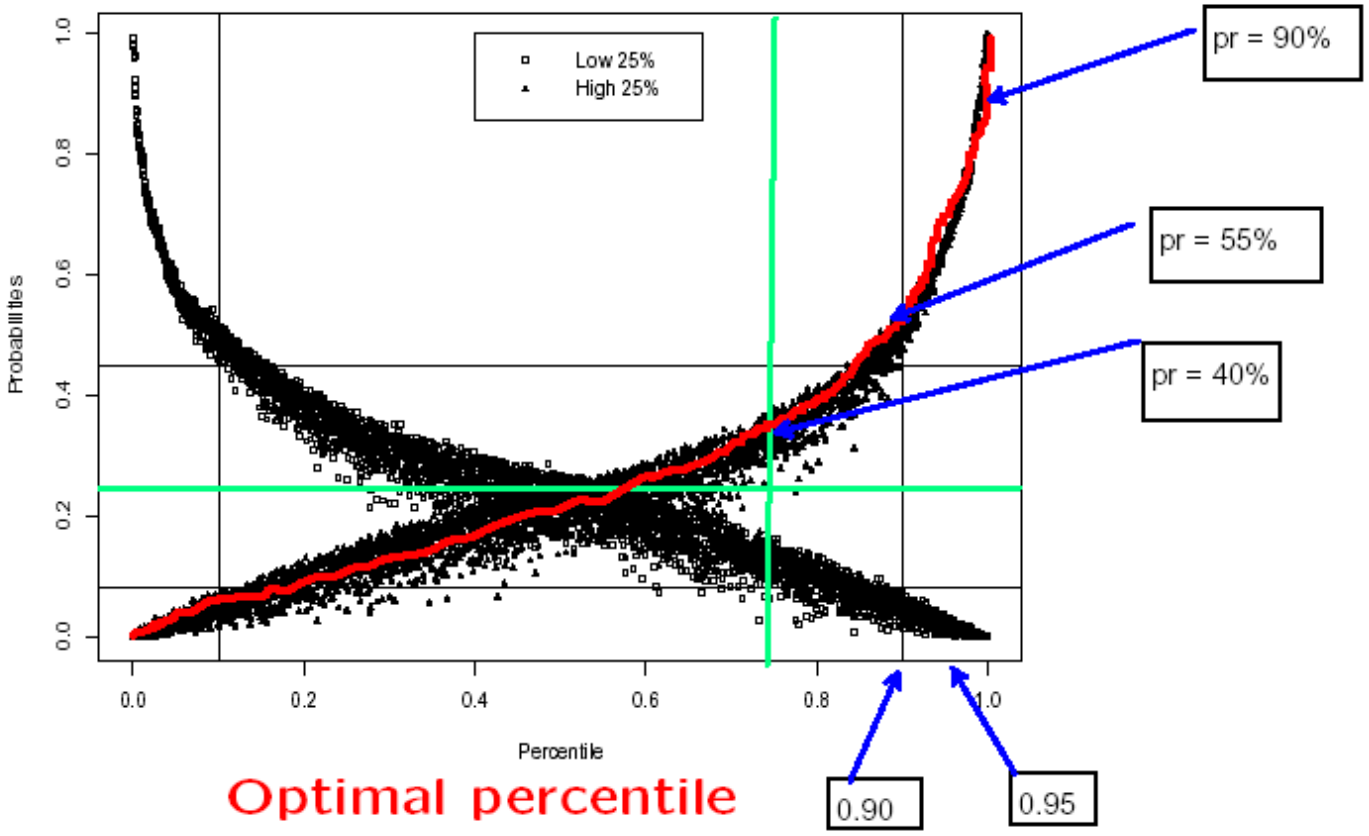
If the data are uninformative: $OC(\gamma) \equiv 1$

If the data are fully informative: $OC(\gamma) \equiv 0$

$OC(\gamma) = pr(P_k < \gamma \mid P_k^{est} > \gamma , Y)/(1 - \gamma)$ = "FDR"/(1 - γ)

BACK TO THE USRDS, SMR EXAMPLE

pr(SMR IN UPPER/LOWER 25% | DATA)



Minimize OC

$OC(\gamma)$ is minimized by:

$$\tilde{P}_k(\gamma) = \text{order based on } pr(P_k > \gamma \mid \mathbf{Y})$$

$$P_k^*(\gamma) = \text{order based on } pr(\theta_k > t(\gamma) \mid \mathbf{Y})$$

$$t(\gamma) = \bar{G}_K^{-1}(\gamma)$$

$$\bar{G}_K(t) = E(G_K(t) \mid \mathbf{Y}) = \frac{1}{K}\sum_k pr(\theta_k > t \mid \mathbf{Y})$$

$$\boxed{\tilde{P}_k(\gamma) \doteq P_k^*(\gamma)}$$

Advantages of P_k^*

- Relates percentiles to a substantive scale
- Far easier to compute than $\widetilde{\mathbf{P}}_k$
- Shows that the Normand et al. (JASA 1997) approach is loss function based

Posterior Classification Performance (γ = 0.80)

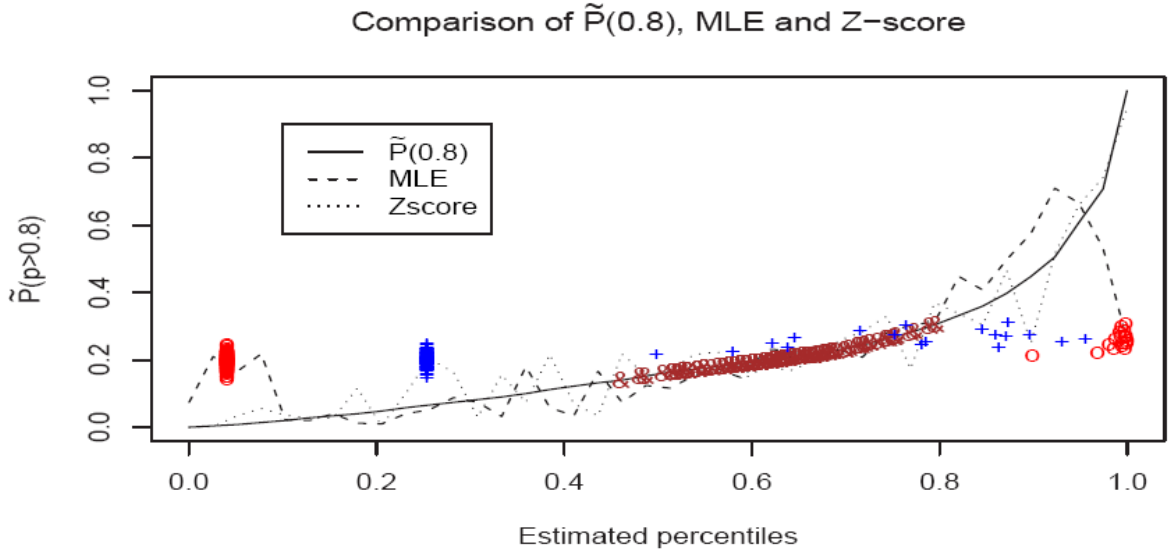


Figure 5: $\pi_k(0.8)$ versus estimated percentiles for three ranking methods using the 1998 data: $\tilde{P}_k(\gamma)$, MLE-based and Z-score-based. For small dialysis centers (fewer than 5 patients), the red "o" denotes MLE-based percentiles, the blue "+" denotes Z-score-based percentiles and the brown "&" denotes $\tilde{P}_k(\gamma)$.

A smoothed, non-parametric prior

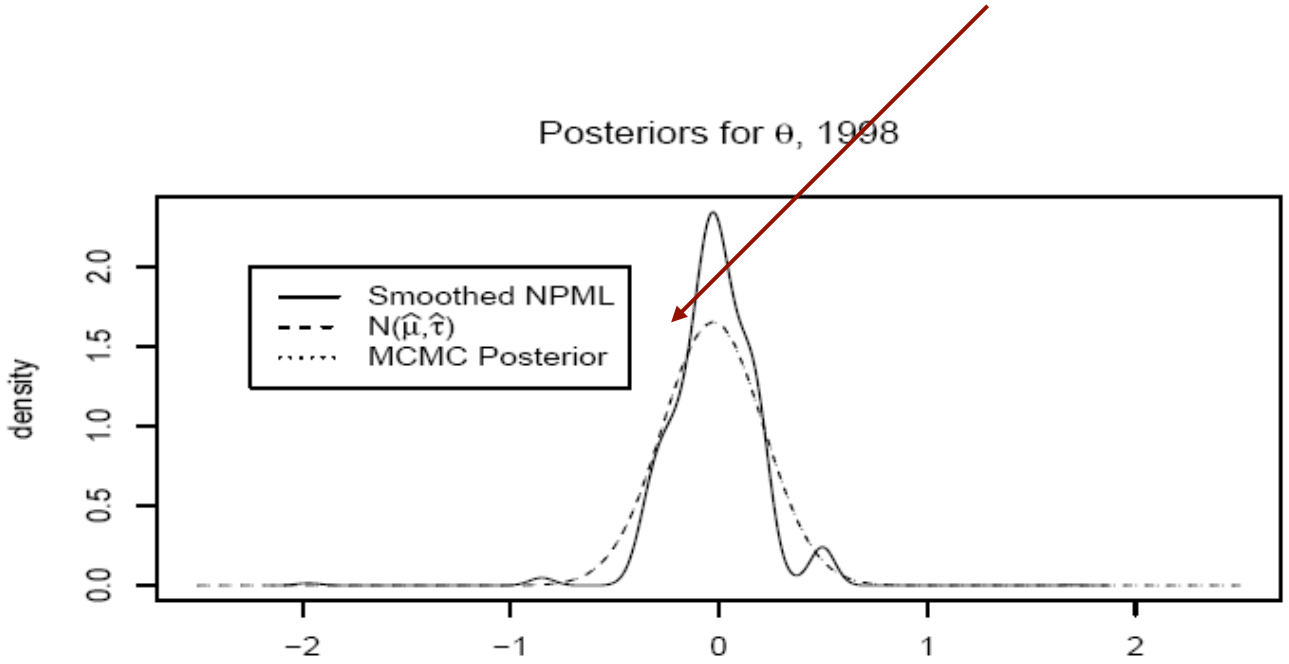


Figure 3. Estimated priors for $\theta = \log(\rho)$ using the 1998 data. The dashed curve is Gaussian with posterior medians for (μ, τ); the dotted curve is a mixture of Gaussians using the posterior distribution of (μ, τ); the solid curve is a smoothed NPML using the "density" function in R with adjustment parameter $= 10$.

An AR(1) Model for the USRDS Data

- (Y_{kt}, m_{kt}) are the observed and expected deaths for provider k in year t, $(k = 1, \ldots 3173;\ t = 0, 1, 2, 3)$
  - Y_{kt}/m_{kt} is the MLE
- With $\rho_{kt} = exp(\theta_{kt})$ the true SMR and $-1 < \phi < 1$ the first-order correlation, we have:

$$\xi_t \ iid\ N(0, V),\ \ \phi \sim h_\phi(\cdot),\ \ \lambda_t = \tau_t^{-2}\ iid\ \text{Gamma}(\alpha, \mu/\alpha)$$

$$\begin{aligned}
[\theta_{10}, \ldots, \theta_{K0} \mid \xi_0, \tau_0] &\quad iid \quad N(\xi_0, \tau_0^2) \\
[\theta_{kt} | \theta_{k(t-1)}, \theta_{k(t-2)}, \ldots; \phi] &\quad = \quad [\theta_{kt} | \theta_{k(t-1)}; \phi] \\
[\theta_{kt} | \theta_{k(t-1)}; \phi] &\quad ind \quad N\left(\xi_t + \phi\tau_t\tau_{t-1}^{-1}\{\theta_{k(t-1)} - \xi_{t-1}\}, \{1 - \phi^2\}\tau_t^2\right) \\
[Y_{kt} \mid m_{kt}, \rho_{kt}] &\quad \sim \quad \text{Poisson}(m_{kt}\rho_{kt})
\end{aligned}$$

- Single-year analyses result from setting $\phi \equiv 0$
- Multi-year analyses put a N(0, 0.2) hyper-prior on the Fisher's-z transformed ϕ

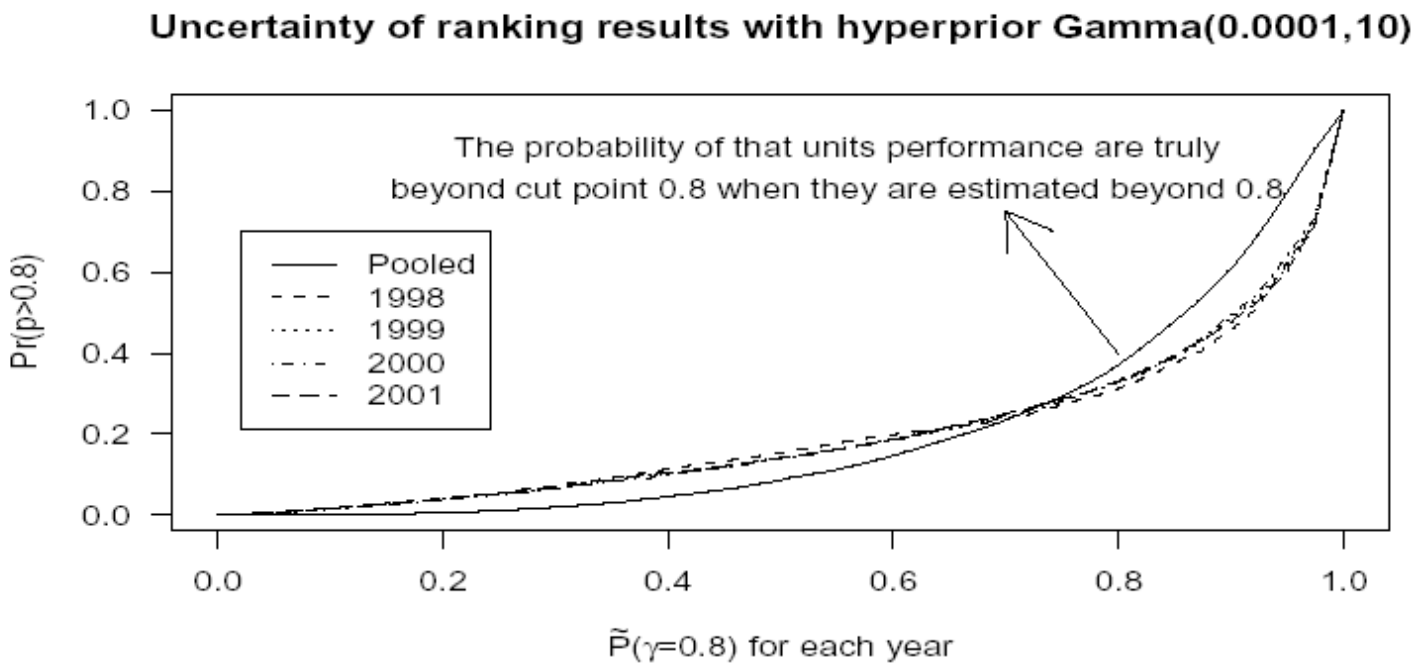
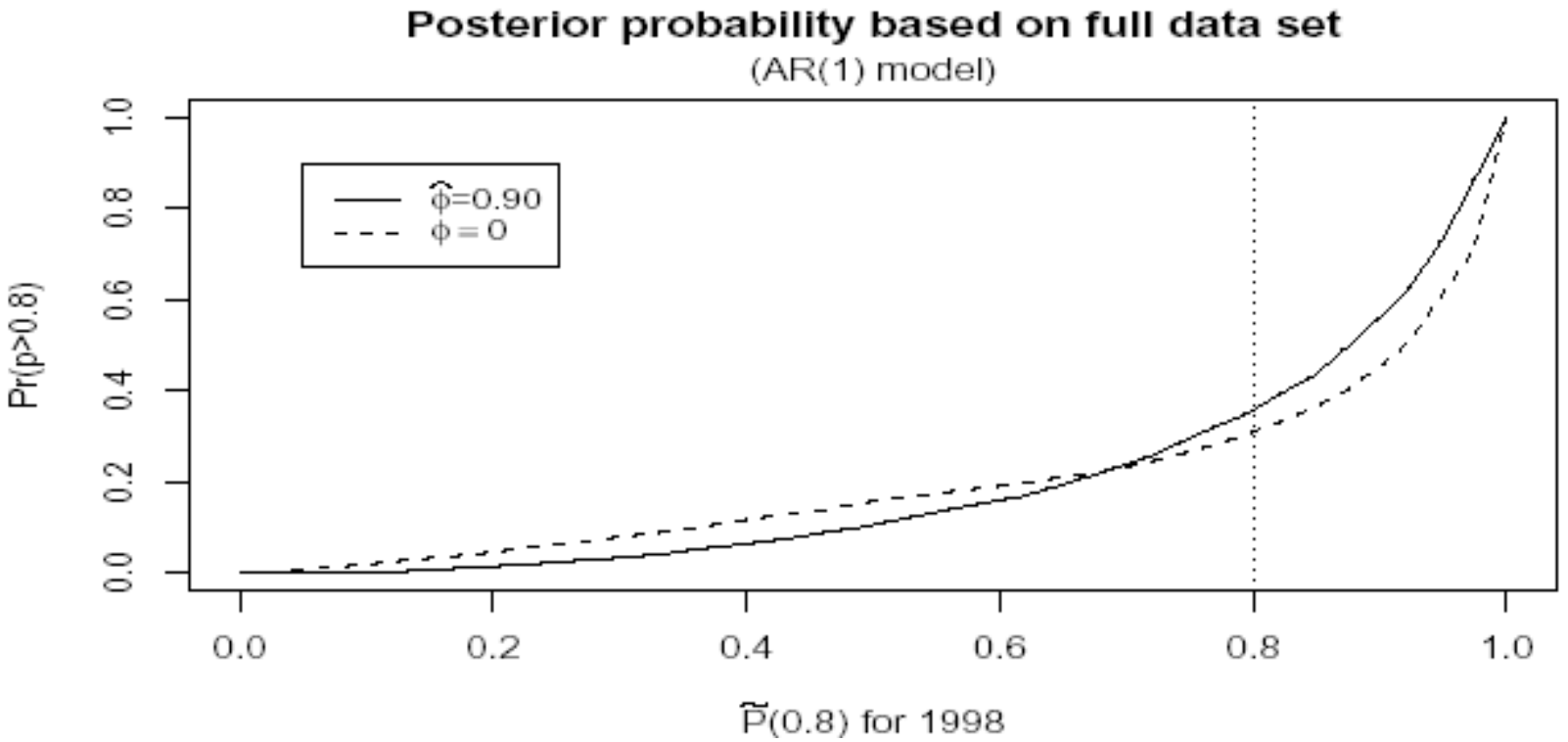


Figure 2: When reporting the ranking results, the uncertainty of the ranking should always be reported together. Current prior does not have much influence on the ranking results.



- **Posterior probability that the true percentile is above $\gamma = 0.8$ vs $\tilde{P}_k(\gamma)$ for 1998**
- **Computed by the single year model $(\phi \equiv 0)$ and the AR(1) model $(\hat{\phi} = 0.90)$**

- To measure longitudinal variation in the ranks/percentile estimates, compute:

$$MSP(\hat{P}) = 100\frac{1}{3K}\sum_{k=1}^{K}\sum_{t=0}^{3}(\hat{P}_{kt} - \hat{P}_{k\cdot})^2$$

$\hat{P}_{kt}$ is the estimated percentile for unit k in year t and $\hat{P}_{k\cdot}$ is the mean over the four years

	Single Year: ($\phi \equiv 0$)				Multi-Year: ($100\phi \sim {}_{88}90_{92}$)			
Parameter	1998	1999	2000	2001	1998	1999	2000	2001
$100 \times OC(0.8)$*	62	61	60	62	49	47	46	50
MSP($\hat{P}$)	62				4			

- In the multi-year section, $100 \times OC(0.8)$ is for the indicated year as estimated from the multi-year model
- ${}_{88}90_{92}$ is the posterior median and 95% credible interval

OC Risk for the exchangeable Gaussian/Gaussian model

$$OC(\gamma) = pr(P_k < \gamma \mid P_k^{est} > \gamma\ ,\ Y)/(1 - \gamma) = \text{“FDR”}/(1 - \gamma)$$

	γ						
$B \downarrow$	0.50	0.60	0.70	0.80	0.90	0.95	0.99
0.01	90	91	93	99	110	122	149
0.25	460	463	474	496	542	589	689
0.50	667	670	682	705	751	796	879
0.75	839	842	850	866	896	923	966
0.95	968	969	971	975	982	988	996
0.99	994	994	994	995	996	998	999

Table 4: Pre-posterior $1000\times$OC for large K in the exchangeable, Gaussian Model.

$$B = \sigma^2/(\sigma^2 + \tau^2)$$

Data are more informative for smaller B

SUMMARY

- A structured approach is necessary
- Substantial information is needed to produce good performance
  - The "best of breed" may still be a dog
- Estimates should always be accompanied by uncertainty assessments

HISTOGRAMS

HISTOGRAMS AND RANKS

- Problems with MLEs
  - The histogram of MLEs is more spread out than that of the true parameters
  - Ranks derived from MLEs are not optimal
- Problems with Posterior Means (PMs)
  - The histogram of PMs is more condensed than that of the true parameters
  - Ranks derived from PMs are not optimal

JUST RIGHT SHRINKAGE

- Shrink toward the prior mean, but less than for the posterior mean

- In the basic, Gaussian model use

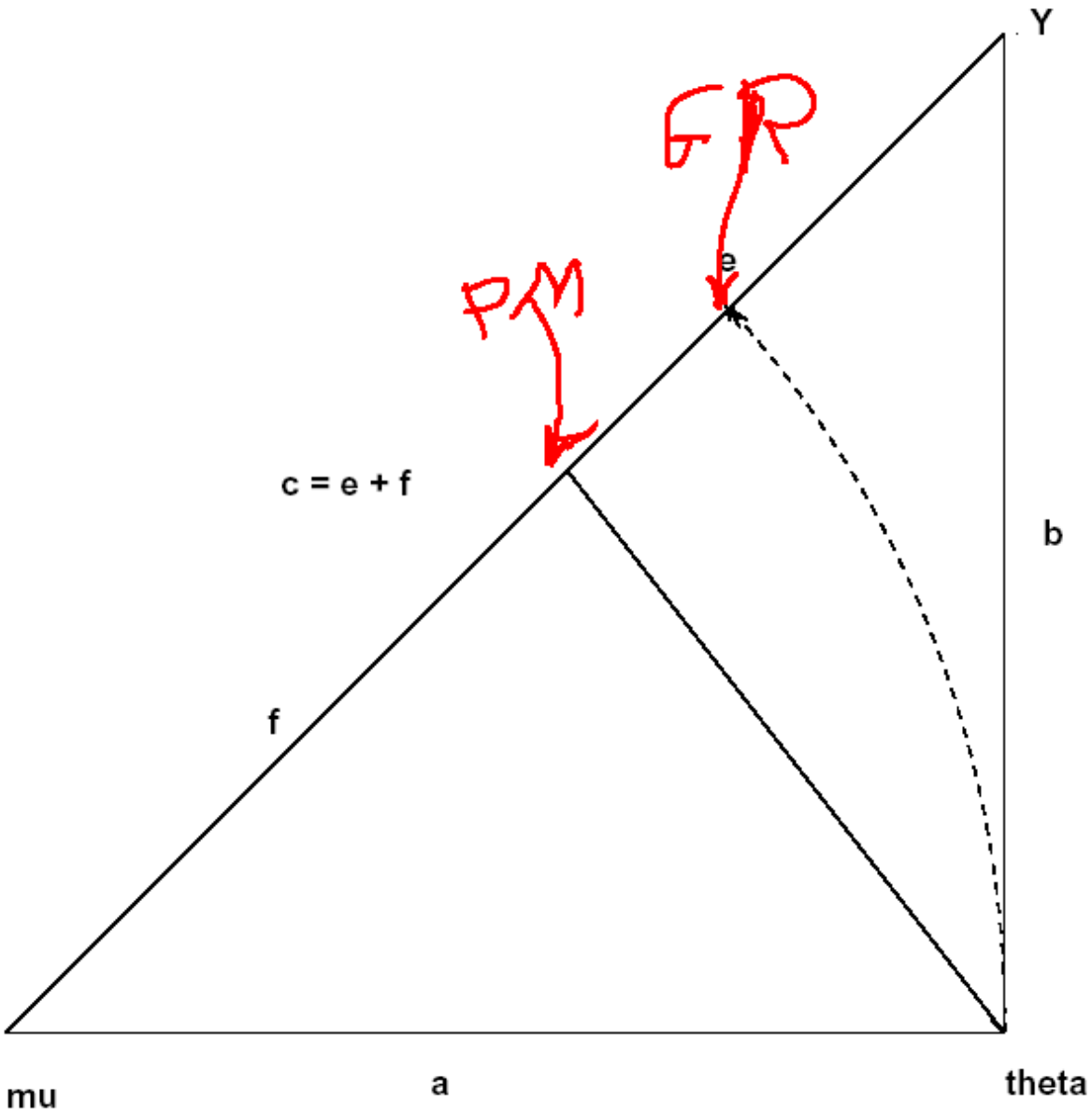
$$\mu + (1 - B)^{\frac{1}{2}}(Y - \mu)$$

- Rather than

$$\mu + (1 - B)(Y - \mu)$$

- Weight on data

(1 - B)	$(1\text{-}B)^{\frac{1}{2}}$
.25	.50
.50	.71
.64	.80
.75	.87


Y
GR
PM
e
c = e + f
b
f
mu
a
theta

HISTOGRAM ESTIMATES
(Shen & Louis 1998)

- The empirical distribution (or the histogram) of the θ_k:

$$G_K(t;\boldsymbol{\theta}) = \frac{1}{K}\sum_{k=1}^{K} I_{\{\theta_k \le t\}}$$

- Applications
  - Histogram estimates
  - Exceedences
  - Subgroups and contrasts
  - "League tables"
  - Prioritization

HISTOGRAM ESTIMATES

- Under Squared error loss (SEL) for G_K use:

$$\bar{G}_K(t \mid \mathbf{Y}) = E_G[G_K(t;\boldsymbol{\theta}) \mid \mathbf{Y}]$$

$$= \frac{1}{K}\Sigma\, P_G[\theta_k \le t \mid \mathbf{Y}]$$

- The optimal discretization ($\hat{G}_K$) has mass $\frac{1}{K}$ at:

$$U_\nu = \bar{G}_K^{-1}\left(\frac{2\nu-1}{2K}\right), \nu = 1, \ldots, K$$

TRIPLE-GOAL ESTIMATES

- Optimal ranks
- Optimal Histogram
- Good performance on estimating individual θs
- "GR" estimates
  - Use $\hat{G}$ for the histogram
  - Use $\hat{R}$ for the ranks
  - Assign the U_ν to coordinates using the $\hat{R}$

Estimated G_K (symmetric case)

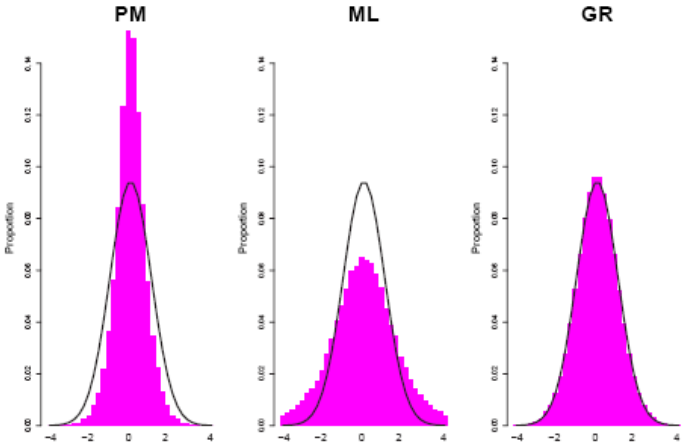


Figure 14: EDF estimates based on PM, ML, and GR when the data-generating and data-analytic distributions are Gaussian. $GM(\{\sigma_k^2\}) = 1,\ rls = 100.$

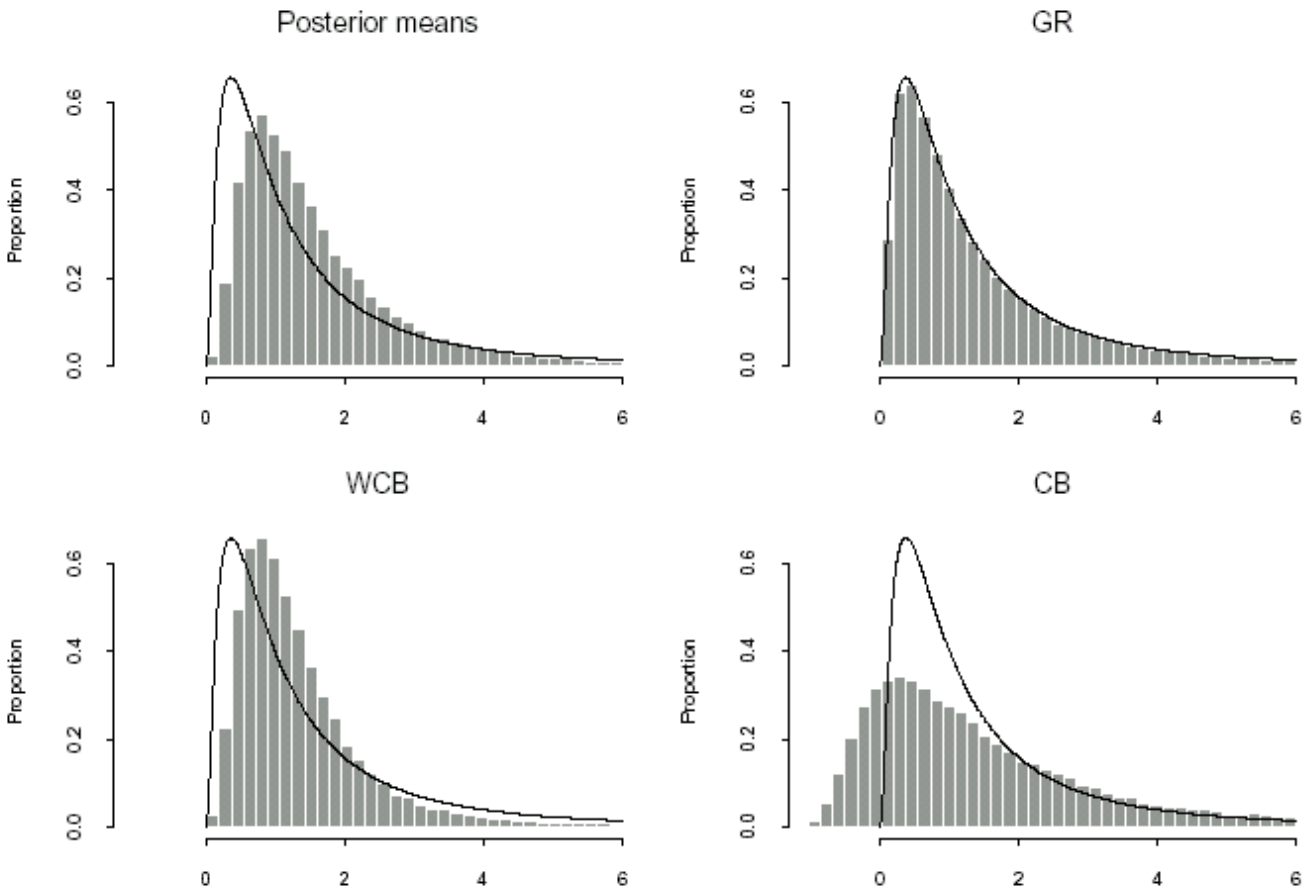


Figure 6: PM, GR and CB for a log-normal prior

Estimated G_K (bimodal case)

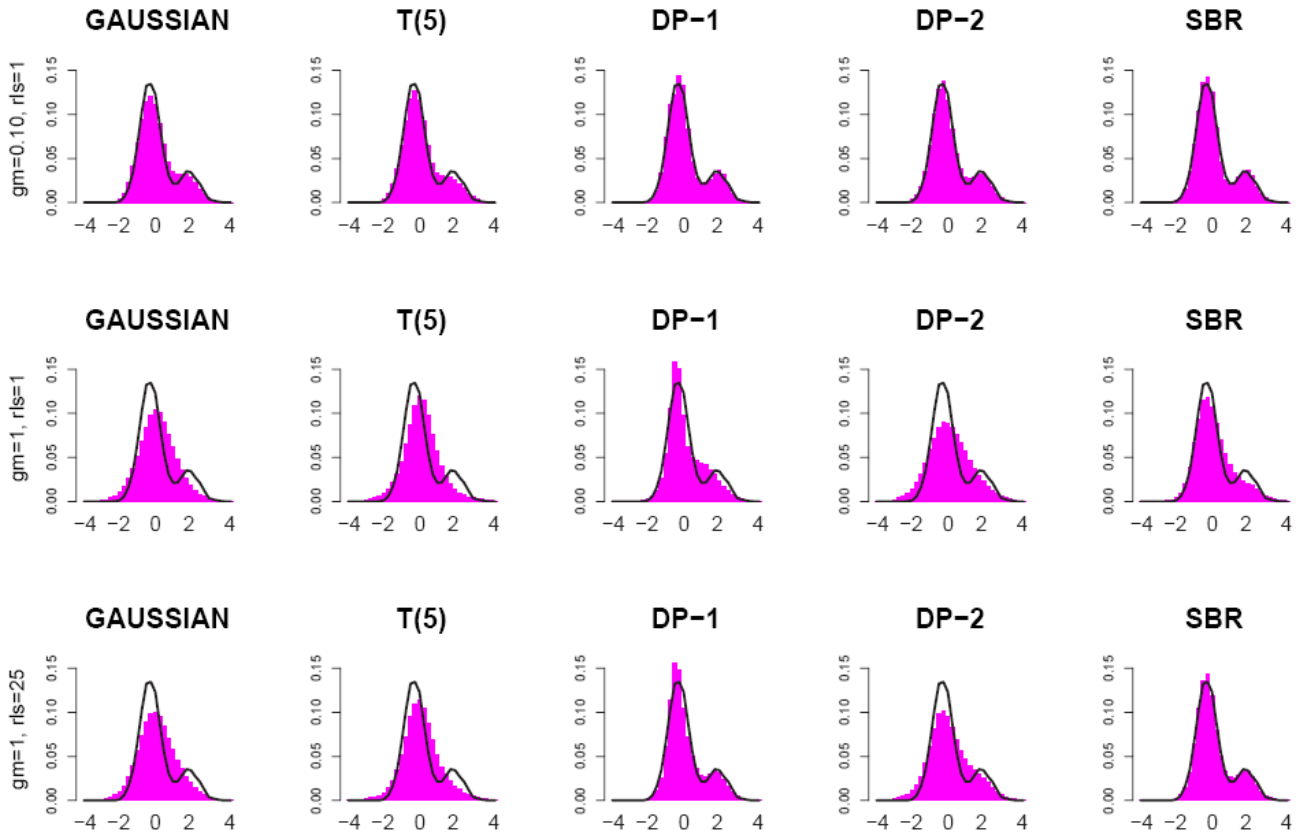


Figure 15: EDF estimates when true G is a mixture of two Gaussians. First row, $gmv = 0.1, rls = 1$; second row, $gmv = 1, rls = 1$; third row, $gm = 1, rls = 25$. Columns display results for different analysis models: column 1, Gaussian; column 2, T_5; column 3, DP-1; column 4, DP-2; column 5, SBR.

Current Research: Percentile-Based Histograms

Replace Gbar by median G
Also do percentile bands

Can also do median quantiles (easier)
and percentile bands on the quantiles

Compare with Gbar

Different Goals → Different Inferences

The Bayesian formalism facilitates goal orientation

Example using BUGS for hospital performance ranking

This example considers mortality rates in 12 hospitals performing cardiac surgery in babies. The data are shown below.

Hospital	No of ops	No of deaths
A	47	0
B	148	18
C	119	8
D	810	46
E	211	8
F	196	13
G	148	9
H	215	31
I	207	14
J	97	8
K	256	29
L	360	24

A Binomial, Multi-level model

For k = 1, . . . , K

$$
\begin{aligned}
n_k &= \#\text{ of patients in hospital k} \\
r_k &= \#\text{ of deaths in hospital k} \\
r_k &\sim \text{Binomial}(n_k, P_k)
\end{aligned}
$$

$$
\begin{aligned}
\text{logit}(P_k) &= \mu + b_k \\
b_k &\sim N(0, \tau^2), \ (\tau^2\text{is the Variance})
\end{aligned}
$$

Population-averaged probability

$$
\bar{P} = E\left[\frac{e^{\mu+b}}{1+e^{\mu+b}}\right] \neq \frac{e^{\mu}}{1+e^{\mu}}
$$

BUGS Model specification

```
model
{
for k in 1:K {
b[k]~dnorm(0, prec)
r[k]~dbin(p[k], n[k])
logit(p[k]) <- mu + b[k]
}
pop.mean<-exp(mu + bb)/(1+exp(mu + bb))
mu~dnorm(0, 1E-6)
prec~dgamma(.0001,.0001)
tausq<-1/prec
add~dnorm(0, prec)
bb<- mu + add
}
```

Monitor the p[k] and ask for ranks

Summary Statistics

node	mean	sd
p[1]	0.05357	0.01959
p[2]	0.1026	0.02203
p[3]	0.07102	0.01701
p[4]	0.05947	0.008078
p[5]	0.05252	0.01354
p[6]	0.06867	0.01401
p[7]	0.06796	0.01597
p[8]	0.1217	0.02196
p[9]	0.06943	0.01435
p[10]	0.07859	0.0193
p[11]	0.1019	0.01745
p[12]	0.06893	0.01185
pop.mean	0.07246	0.0105

Posterior distributions of the ranks