Some Curiosities Arising in Objective Bayesian Analysis

Jim Berger
Duke University
Statistical and Applied Mathematical Institute

Yale University
May 15, 2009
Three vignettes related to John’s work

- Some puzzling things concerning invariant priors
- Some puzzling things concerning multiplicity
- What is the effective sample size?
I. Objective Priors: Why Can’t We Have It All?

Figure 1: John at Princeton in 1965.
An Example: Inference for the Correlation Coefficient

The bivariate normal distribution of \((x_1, x_2)\) has mean \((\mu_1, \mu_2)\) and covariance matrix \(\Sigma = \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix}\), where \(\rho\) is the correlation between \(x_1\) and \(x_2\).

For a sample \((x_{11}, x_{21}), (x_{12}, x_{22}), \ldots, (x_{1n}, x_{2n})\), the sufficient statistics are \(\bar{x} = (\bar{x}_1, \bar{x}_2)'\), where \(\bar{x}_i = n^{-1} \sum_{j=1}^{n} x_{ij}\), and

\[
S = \sum_{i=1}^{n} (x_i - \bar{x})(x_i - \bar{x})' = \begin{pmatrix} s_{11} & r \sqrt{s_{11}s_{22}} \\ r \sqrt{s_{11}s_{22}} & s_{22} \end{pmatrix},
\]

where \(s_{ij} = \sum_{k=1}^{n} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)\), \(r = s_{12}/\sqrt{s_{11}s_{22}}\).
Three interesting priors for inference concerning $\rho$:

- The reference prior (Lindley and Bayarri)
  \[
  \pi^R(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{\sigma_1 \sigma_2 (1 - \rho^2)} .
  \]

- The right-Haar prior
  \[
  \pi^{RH1}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{\sigma_2 (1 - \rho^2)} ,
  \]
  which is right-Haar w.r.t. the lower triangular matrix group action
  \[
  (a_1, a_2, T^l) \circ (x_1, x_2)' = T^l(x_1, x_2)' + (a_1, a_2) .
  \]

- The right-Haar prior
  \[
  \pi^{RH2}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{\sigma_1 (1 - \rho^2)} ,
  \]
  which is right-Haar w.r.t. the upper triangular matrix, $T^u$, group action.
Credible intervals for \( \rho \), under either right-Haar prior, can be approximated by

- drawing independent \( Z \sim N(0, 1) \), \( \chi^2_{n-1} \) and \( \chi^2_{n-2} \);
- setting \( \rho = \frac{Y}{\sqrt{1+Y^2}} \), where \( Y = -\frac{Z}{\sqrt{\chi^2_{n-1}}} + \frac{\sqrt{\chi^2_{n-2}}}{\sqrt{\chi^2_{n-1} \sqrt{1-r^2}}} \);
- repeating this process 10,000 times;
- using the \( \frac{\alpha}{2} \)% upper and lower percentiles of these generated \( \rho \) to form the desired confidence limits.

Credible intervals for \( \rho \), under the reference prior, can be found by replacing the first two steps above by

- Generate \( \Sigma = \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix} \sim \text{Inverse Wishart}(S^{-1}, n-1) \).
- Generate \( u \sim \text{unif}(0, 1) \). If \( u \leq \sqrt{1-\rho^2} \), record \( \rho \). Otherwise repeat.
Lemma 1

1. The Bayesian credible set for $\rho$, from either right-Haar prior, has exact frequentist coverage $1 - \alpha$.

2. This credible set is the the fiducial confidence interval for $\rho$ of Fisher 1930.

So we have it all, a simultaneous objective Bayesian, fiducial, and exact frequentist confidence set.

Or do we? In the 60’s, Brillinger showed that, if one starts with the density $f(r | \rho)$ of $r$, there is no prior distribution for $\rho$ whose posterior equals the fiducial distribution.

- Geisser and Cornfield (1963) thus conjectured that fiducial and Bayesian inference could not agree here. (They do.)

- But, since $\pi^{RH}(\rho | x)$ can be shown only to depend on the data through $r$, we have a marginalization paradox (Dawid, Stone and Zidek, 1973): $\pi^{RH}(\rho | x) = g(\rho, r) \neq f(r | \rho)\pi(\rho)$ for any $\pi(\cdot)$. 
Can We Trust Bayesian ‘Truisms’ with Improper Priors?

Two ‘Truisms:’ If considering various priors, either

- “average” (or go hierarchical); or
- choose the empirical Bayes prior that maximizes the marginal likelihood.

1. Consider the symmetrized right-Haar prior

   \[
   \pi^S(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \pi^{RH1}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) + \pi^{RH2}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)
   \]
   \[
   = \frac{1}{\sigma_1^2(1 - \rho^2)} + \frac{1}{\sigma_2^2(1 - \rho^2)}.
   \]

2. Any rotation \( \mathbf{\Gamma} \) of coordinates yields a new right-Haar prior. The empirical Bayes prior, \( \pi^{EB} \), is the right-Haar prior for that rotation for which \( s_{12}^* = 0 \), where

   \[
   \mathbf{S}^* \equiv \begin{pmatrix} s_{11}^* & s_{12}^* \\ s_{12}^* & s_{22}^* \end{pmatrix} = \mathbf{\Gamma} \mathbf{S} \mathbf{\Gamma}'.
   \]
Innovation and Inventiveness in Statistics Methodologies

<table>
<thead>
<tr>
<th>((\sigma_1, \sigma_2, \rho))</th>
<th>(R(\hat{\Sigma}_1))</th>
<th>(R(\hat{\Sigma}_2))</th>
<th>(R(\hat{\Sigma}_S))</th>
<th>(R(\hat{\Sigma}_{EB}))</th>
</tr>
</thead>
<tbody>
<tr>
<td>((1, 1, 0))</td>
<td>.4287</td>
<td>.4288</td>
<td>.4452</td>
<td>.6052</td>
</tr>
<tr>
<td>((1, 2, 0))</td>
<td>.4278</td>
<td>.4270</td>
<td>.4424</td>
<td>.5822</td>
</tr>
<tr>
<td>((1, 5, 0))</td>
<td>.4285</td>
<td>.4287</td>
<td>.4391</td>
<td>.5404</td>
</tr>
<tr>
<td>((1, 50, 0))</td>
<td>.4254</td>
<td>.4250</td>
<td>.4272</td>
<td>.5100</td>
</tr>
<tr>
<td>((1, 1, .1))</td>
<td>.4255</td>
<td>.4266</td>
<td>.4424</td>
<td>.5984</td>
</tr>
<tr>
<td>((1, 1, .5))</td>
<td>.4274</td>
<td>.4275</td>
<td>.4403</td>
<td>.5607</td>
</tr>
<tr>
<td>((1, 1, .9))</td>
<td>.4260</td>
<td>.4255</td>
<td>.4295</td>
<td>.5159</td>
</tr>
<tr>
<td>((1, 1, -.9))</td>
<td>.4242</td>
<td>.4243</td>
<td>.4280</td>
<td>.5119</td>
</tr>
</tbody>
</table>

Table 1: Estimated frequentist risks of various estimates of \(\Sigma\), under Stein’s loss and when \(n = 10\); \(\hat{\Sigma}_i\) are the right-Haar estimates, \(\hat{\Sigma}_S\) is the symmetrized prior estimate, and \(\hat{\Sigma}_{EB}\) is the empirical Bayes estimate.
II. Bayesian Multiplicity Issues in Variable Selection

Figure 2: John Hartigan with one of his multiple grandchildren
Problem: Data $\mathbf{X}$ arises from a normal linear regression model, with $m$ possible regressors having associated unknown regression coefficients $\beta_i, i = 1, \ldots, m$, and unknown variance $\sigma^2$.

Models: Consider selection from among the submodels $\mathcal{M}_i$, $i = 1, \ldots, 2^m$, having only $k_i$ regressors with coefficients $\beta_i$ (a subset of $(\beta_1, \ldots, \beta_m)$) and resulting density $f_i(\mathbf{x} \mid \beta_i, \sigma^2)$.

Prior density under $\mathcal{M}_i$: Zellner-Siow priors $\pi_i(\beta_i, \sigma^2)$.

Marginal likelihood of $\mathcal{M}_i$: $m_i(\mathbf{x}) = \int f_i(\mathbf{x} \mid \beta_i, \sigma^2)\pi_i(\beta_i, \sigma^2) \, d\beta_i d\sigma^2$

Prior probability of $\mathcal{M}_i$: $P(\mathcal{M}_i)$

Posterior probability of $\mathcal{M}_i$:

$$P(\mathcal{M}_i \mid \mathbf{x}) = \frac{P(\mathcal{M}_i)m_i(\mathbf{x})}{\sum_j P(\mathcal{M}_j)m_j(\mathbf{x})}.$$
Common Choices of the $P(M_i)$

Equal prior probabilities: $P(M_i) = 2^{-m}$

Bayes exchangeable variable inclusion:

- Each variable, $\beta_i$, is independently in the model with unknown probability $p$ (called the prior inclusion probability).
- $p$ has a Beta($p \mid a, b$) distribution. (We use $a = b = 1$, the uniform distribution, as did Jeffreys 1961, who also suggested alternative choices of the $P(M_i)$. Probably $a = b = 1/2$ is better.)
- Then, since $k_i$ is the number of variables in model $M_i$,

$$P(M_i) = \int_0^1 p^{k_i} (1 - p)^{m-k_i} \text{Beta}(p \mid a, b) dp = \frac{\text{Beta}(a + k_i, b + m - k_i)}{\text{Beta}(a, b)}.$$ 

Empirical Bayes exchangeable variable inclusion: Find the MLE $\hat{p}$ by maximizing the marginal likelihood of $p$, $\sum_j p^{k_j} (1 - p)^{m-k_j} m_j(x)$, and use $P(M_i) = \hat{p}^{k_i} (1 - \hat{p})^{m-k_i}$ as the prior model probabilities.
Controlling for multiplicity in variable selection

Equal prior probabilities: \( P(M_i) = 2^{-m} \) does not control for multiplicity; it corresponds to fixed prior inclusion probability \( p = 1/2 \) for each variable.

Empirical Bayes exchangeable variable inclusion does control for multiplicity, in that \( \hat{p} \) will be small if there are many \( \beta_i \) that are zero.

Bayes exchangeable variable inclusion also controls for multiplicity (see Scott and Berger, 2008), although the \( P(M_i) \) are fixed.

Note: The control of multiplicity by Bayes and EB variable inclusion usually reduces model complexity, but is different than the usual Bayesian Ockham’s razor effect that reduces model complexity.

- The Bayesian Ockham’s razor operates through the effect of model priors \( \pi_i(\beta_i, \sigma^2) \) on \( m_i(x) \), penalizing models with more parameters.
- Multiplicity correction occurs through the choice of the \( P(M_i) \).
### Table 2: Posterior inclusion probabilities for 10 real variables in a simulated data set.

<table>
<thead>
<tr>
<th>Signal</th>
<th>Equal model probabilities</th>
<th>Bayes variable inclusion</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Number of noise variables</td>
<td></td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>10</td>
</tr>
<tr>
<td>$\beta_1 : -1.08$</td>
<td>.999</td>
<td>.999</td>
</tr>
<tr>
<td>$\beta_2 : -0.84$</td>
<td>.999</td>
<td>.999</td>
</tr>
<tr>
<td>$\beta_3 : -0.74$</td>
<td>.999</td>
<td>.999</td>
</tr>
<tr>
<td>$\beta_4 : -0.51$</td>
<td>.977</td>
<td>.977</td>
</tr>
<tr>
<td>$\beta_5 : -0.30$</td>
<td>.292</td>
<td>.289</td>
</tr>
<tr>
<td>$\beta_6 : +0.07$</td>
<td>.259</td>
<td>.286</td>
</tr>
<tr>
<td>$\beta_7 : +0.18$</td>
<td>.219</td>
<td>.248</td>
</tr>
<tr>
<td>$\beta_8 : +0.35$</td>
<td>.773</td>
<td>.771</td>
</tr>
<tr>
<td>$\beta_9 : +0.41$</td>
<td>.927</td>
<td>.912</td>
</tr>
<tr>
<td>$\beta_{10} : +0.63$</td>
<td>.995</td>
<td>.995</td>
</tr>
<tr>
<td>False Positives</td>
<td>0</td>
<td>2</td>
</tr>
</tbody>
</table>
Comparison of Bayes and Empirical Bayes Approaches

Theorem 1  In the variable-selection problem, if the null model (or full model) has the largest marginal likelihood, \( m(x) \), among all models, then the MLE of \( p \) is \( \hat{p} = 0 \) (or \( \hat{p} = 1 \)). (The naive EB approach, which assigns \( P(M_i) = \hat{p}^{k_i}(1 - \hat{p})^{m - k_i} \), concludes that the null (full) model has probability 1.)

A simulation with 10,000 repetitions to gauge the severity of the problem:

- \( m = 14 \) covariates, orthogonal design matrix
- \( p \) drawn from \( U(0, 1) \); regression coefficients are 0 with probability \( p \) and drawn from a Zellner-Siow prior with probability \( (1 - p) \).
- \( n = 16, 60, \) and \( 120 \) observations drawn from the given regression model.

<table>
<thead>
<tr>
<th>Case</th>
<th>( \hat{p} = 0 )</th>
<th>( \hat{p} = 1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( n = 16 )</td>
<td>820</td>
<td>781</td>
</tr>
<tr>
<td>( n = 60 )</td>
<td>783</td>
<td>766</td>
</tr>
<tr>
<td>( n = 120 )</td>
<td>723</td>
<td>747</td>
</tr>
</tbody>
</table>
Is empirical Bayes at least accurate asymptotically as $m \to \infty$?

Posterior model probabilities, given $p$:

$$P(\mathcal{M}_i \mid \mathbf{x}, p) = \frac{p^{k_i}(1 - p)^{m - k_i} m_i(\mathbf{x})}{\sum_j p^{k_j}(1 - p)^{m - k_j} m_j(\mathbf{x})}$$

Posterior distribution of $p$: $\pi(p \mid \mathbf{x}) = K \sum_j p^{k_j}(1 - p)^{m - k_j} m_j(\mathbf{x})$

This does concentrate about the true $p$ as $m \to \infty$, so one might expect that

$$P(\mathcal{M}_i \mid \mathbf{x}) = \int_0^1 P(\mathcal{M}_i \mid \mathbf{x}, p) \pi(p \mid \mathbf{x}) dp \approx P(\mathcal{M}_i \mid \mathbf{x}, \hat{p}) \propto m_i(\mathbf{x}) \hat{p}^{k_i}(1 - \hat{p})^{m - k_i}.$$ 

This is not necessarily true; indeed

$$\int_0^1 P(\mathcal{M}_i \mid \mathbf{x}, p) \pi(p \mid \mathbf{x}) dp = \int_0^1 \frac{p^{k_i}(1 - p)^{m - k_i} m_i(\mathbf{x})}{\pi(p \mid \mathbf{x})/K} \times \pi(p \mid \mathbf{x}) \, dp$$

$$\propto m_i(\mathbf{x}) \int_0^1 p^{k_i}(1 - p)^{m - k_i} dp \propto m_i(\mathbf{x}) P(\mathcal{M}_i).$$

Caveat: Some EB techniques have been justified; see Efron and Tibshirani (2001), Johnstone and Silverman (2004), Cui and George (2006), and Bogdan et. al. (2008).
III. What is the Effective Sample Size in Generalized BIC?

Figure 3: John Hartigan in Sin City
Data: Independent vectors $x_i \sim g_i(x_i \mid \theta)$, for $i = 1, \ldots, n$.

Unknown: $\theta = (\theta_1, \ldots, \theta_p)$; $\hat{\theta}$ is the MLE

Log-likelihood function: $l(\theta) = \log f(x \mid \theta) = \log (\prod_{i=1}^{n} g_i(x_i \mid \theta))$

where $x = (x_1, \ldots, x_n)$.

Usual BIC: $BIC \equiv 2l(\hat{\theta}) - p \log n$ (Schwarz, 1978)

Generalization of BIC: $2l(\hat{\theta}) - \sum_{i=1}^{p} \log(1 + n_i) + 2 \sum_{i=1}^{p} \log \frac{1-e^{-v_i}}{\sqrt{2} v_i}$,

- $v_i = \frac{\xi_i^2}{d_i(1+n_i)}$,
- the $d_i^{-1}$ are the eigenvalues of the observed information matrix,
- the $\xi_i$ are the coordinates in an orthogonally transformed $\theta$.
- $n_i$ is the effective sample size corresponding to $\xi$. What should these be?
Ex. Group means: For $i = 1, \ldots, p$ and $l = 1, \ldots, r$,

$$X_{il} = \mu_i + \epsilon_{il}, \quad \text{where} \quad \epsilon_{il} \sim N(0, \sigma^2).$$

- It might seem that $n = pr$ but, if one followed Schwarz, one would have (defining $\mu = (\mu_1, \ldots, \mu_p)^t$) that $X_l = (X_{1l}, \ldots, X_{pl})^t \sim N_p(\mu, \sigma^2 I)$, $l = 1, \ldots, r$, so that the ‘sample size’ appearing in BIC should be $r$.

- The ‘effective sample size’ for each $\mu_i$ is $r$, but the effective sample size for $\sigma^2$ is $pr$, so effective sample size is parameter-dependent.

- One could easily be in the situation where $p \to \infty$ but the effective sample size $r$ is fixed.
Ex. Random effects group means: $\mu_i \sim N(\xi, \tau^2)$, with $\xi$ and $\tau^2$ being unknown. What is the number of parameters (see also Pauler (1998))? 

(1) If $\tau^2 = 0$, there is only one parameter $\xi$.

(2) If $\tau^2$ is huge, is the number of parameters $p + 2$?

(3) But, if one integrates out $\mu = (\mu_1, \ldots, \mu_p)$, then 
\[
    f(x | \sigma^2, \xi, \tau^2) = \int f(x | \mu, \xi, \sigma^2) \pi(\mu | \xi, \tau^2) d\mu
\]
\[
    \propto \frac{1}{\sigma^p(r-1)} \exp \left\{ \frac{\hat{\sigma}^2}{2\sigma^2} \right\} \prod_{i=1}^{p} \exp \left\{ -\frac{(\bar{x}_i - \xi)^2}{2(\frac{\sigma^2}{r} + \tau^2)} \right\},
\]
so $p = 3$ if one can work directly with $f(x | \sigma^2, \xi, \tau^2)$.

Note: In this example the effective sample sizes should be $\approx pr$ for $\sigma^2$, $\approx p$ for $\xi$ and $\tau^2$, and $\approx r$ for the $\mu_i$’s.

Ex. Common mean, differing variances: Suppose $n/2$ of the $Y_i$ are $N(\theta, 1)$, while $n/2$ are $N(\theta, 1000)$. Clearly the ‘effective sample size’ is roughly $n/2$. 
**Ex. ANOVA:** $Y = (Y_1, \ldots, Y_n)^t \sim N_n(X\beta, \sigma^2 I)$, where $X$ is a given $n \times p$ matrix of 1’s and -1’s with orthogonal columns, where $\beta = (\beta_1, \ldots, \beta_p)^t$ and $\sigma^2$ are unknown. Then the information matrix for $\theta = (\beta, \sigma^2)$ is

$$\hat{I} = \begin{pmatrix} \frac{n}{\hat{\sigma}^2} I_{p \times p} & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4} \end{pmatrix}$$

so that now the effective sample size appears to be $n$ for all parameters.

Note: The group means problem and ANOVA are linear models, so one can have effective sample sizes from $r = 1$ to $n$ for parameters in the linear model.
Defining the ‘effective sample size’ \( n_j \) for \( \xi_j \):

For the case where no variables are integrated out, a possible general definition for the ‘effective sample size’ follows from considering the information associated with observation \( x_i \) arising from the single-observation expected information matrix \( I_i^* = O'(I_{i,jk}^*)O \), where

\[
I_{i,jk}^* = -E \left[ \frac{\partial^2}{\partial \theta_j \partial \theta_k} \log f_i(x_i | \theta) \right] \bigg|_{\theta = \hat{\theta}}.
\]
Since $I_{jj}^* = \sum_{i=1}^n I_{i,jj}^*$ is the expected information about $\xi_j$, a reasonable way to define $n_j$ is

- define information weights $w_{ij} = I_{i,jj}^*/\sum_{k=1}^n I_{k,jj}^*$;
- define the effective sample size for $\xi_j$ as

$$n_j = \frac{I_{jj}^*}{\sum_{i=1}^n w_{ij} I_{i,jj}^*} = \frac{(I_{jj}^*)^2}{\sum_{i=1}^n (I_{i,jj}^*)^2}.$$  

Intuitively, $\sum w_{ij} I_{i,jj}^*$ is a weighted measure of the information ‘per observation’, and dividing the total information about $\xi_j$ by this information per case seems plausible as an effective sample size.
THANKS ALL
THANKS JOHN