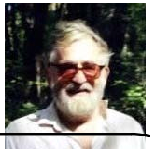


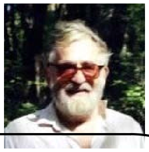
Some Curiosities Arising in Objective Bayesian Analysis

Jim Berger

Duke University

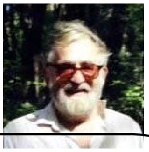
Statistical and Applied Mathematical Institute

Yale University
May 15, 2009



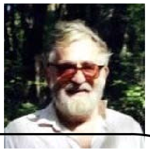
Three vignettes related to John's work

- 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.



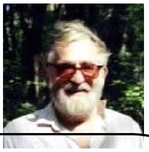
An Example: Inference for the Correlation Coefficient

The bivariate normal distribution of (x_1, x_2) has mean (μ_1, μ_2) and covariance matrix $\mathbf{\Sigma} = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix}$, where ρ 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 $\overline{\boldsymbol{x}} = (\overline{x}_1, \overline{x}_2)'$, where $\overline{x}_i = n^{-1}\sum_{j=1}^n x_{ij}$, and

$$\boldsymbol{S} \;=\; \sum_{i=1}^n (\boldsymbol{x}_i - \overline{\boldsymbol{x}})(\boldsymbol{x}_i - \overline{\boldsymbol{x}})' = \begin{pmatrix} s_{11} & r\sqrt{s_{11}s_{22}} \\ r\sqrt{s_{11}s_{22}} & s_{22} \end{pmatrix},$$

where $s_{ij} = \sum_{k=1}^n (x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j)$, $\quad r = s_{12}/\sqrt{s_{11}s_{22}}$.



Three interesting priors for inference concerning ρ:

- 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

$$\pi^{RH1}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{\sigma_2^2 (1-\rho^2)} \, ,$$

  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

$$\pi^{RH2}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{\sigma_1^2 (1-\rho^2)} \, ,$$

  which is right-Haar w.r.t. the upper triangular matrix, T^u, group action.

Credible intervals for ρ, under either right-Haar prior, can be approximated by

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

Credible intervals for ρ, under the reference prior, can be found by replacing the first two steps above by

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

Lemma 1

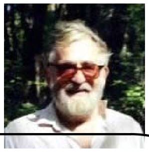
1. The Bayesian credible set for ρ, from either right-Haar prior, has exact frequentist coverage $1-\alpha$.

2. This credible set is the the fiducial confidence interval for ρ 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 $f(r \mid \rho)$ of r, there is no prior distribution for ρ 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 $\pi^{RH}(\rho \mid \boldsymbol{x})$ can be shown only to depend on the data through r, we have a *marginalization paradox* (Dawid, Stone and Zidek, 1973): $\pi^{RH}(\rho \mid \boldsymbol{x}) = g(\rho, r) \neq f(r \mid \rho)\pi(\rho)$ 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

$$\begin{aligned}\pi^S(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho) &= \pi^{RH1}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)+\pi^{RH2}(\mu_1,\mu_2,\sigma_1,\sigma_2,\rho)\\ &= \frac{1}{\sigma_1^2(1-\rho^2)}+\frac{1}{\sigma_2^2(1-\rho^2)}.\end{aligned}$$

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

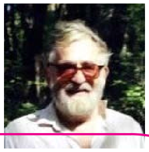
$$\boldsymbol{S}^* \equiv \begin{pmatrix} s_{11}^* & s_{12}^* \\ s_{12}^* & s_{22}^* \end{pmatrix} = \mathbf{\Gamma}\boldsymbol{S}\mathbf{\Gamma}'\,.$$

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

Table 1: Estimated frequentist risks of various estimates of $\mathbf{\Sigma}$, under Stein's loss and when $n = 10$; $\widehat{\mathbf{\Sigma}}_i$ are the right-Haar estimates, $\widehat{\mathbf{\Sigma}}_S$ is the symmetrized prior estimate, and $\widehat{\mathbf{\Sigma}}_{EB}$ 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 σ^2.

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

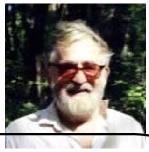
Prior density under $\mathcal{M}_i$: Zellner-Siow priors $\pi_i(\boldsymbol{\beta}_i, \sigma^2)$.

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

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 $P(\mathcal{M}_i)$

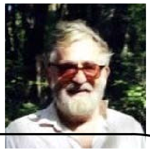
Equal prior probabilities: $P(\mathcal{M}_i) = 2^{-m}$

Bayes exchangeable variable inclusion:

- Each variable, β_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 $P(\mathcal{M}_i)$. Probably $a = b = 1/2$ is better.)
- Then, since k_i is the number of variables in model $\mathcal{M}_i$,

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

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



Controlling for multiplicity in variable selection

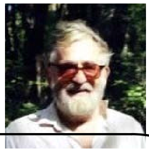
Equal prior probabilities: $P(\mathcal{M}_i) = 2^{-m}$ does *not* control for multiplicity; it corresponds to fixed prior inclusion probability $p = 1/2$ for each variable.

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

Bayes exchangeable variable inclusion also controls for multiplicity (see Scott and Berger, 2008), although the $P(\mathcal{M}_i)$ are fixed.

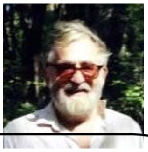
Note: The control of multiplicity by Bayes and EB variable inclusion usually reduces model complexity, but is *different* than the usual Bayeisan Ockham's razor effect that reduces model complexity.

- The Bayesian Ockham's razor operates through the effect of model priors $\pi_i(\boldsymbol{\beta}_i, \sigma^2)$ on $m_i(\mathbf{x})$, penalizing models with more parameters.
- Multiplicity correction occurs through the choice of the $P(\mathcal{M}_i)$.



	Equal model probabilities				Bayes variable inclusion			
	Number of noise variables				Number of noise variables			
Signal	1	10	40	90	1	10	40	90
$\beta_1 : -1.08$	.999	.999	.999	.999	.999	.999	.999	.999
$\beta_2 : -0.84$	.999	.999	.999	.999	.999	.999	.999	.988
$\beta_3 : -0.74$	.999	.999	.999	.999	.999	.999	.999	.998
$\beta_4 : -0.51$	.977	.977	.999	.999	.991	.948	.710	.345
$\beta_5 : -0.30$	.292	.289	.288	.127	.552	.248	.041	.008
$\beta_6 : +0.07$	.259	.286	.055	.008	.519	.251	.039	.011
$\beta_7 : +0.18$	.219	.248	.244	.275	.455	.216	.033	.009
$\beta_8 : +0.35$	.773	.771	.994	.999	.896	.686	.307	.057
$\beta_9 : +0.41$	.927	.912	.999	.999	.969	.861	.567	.222
$\beta_{10} : +0.63$	.995	.995	.999	.999	.996	.990	.921	.734
False Positives	0	2	5	10	0	1	0	0

Table 2: Posterior inclusion probabilities for 10 real variables in a simulated data set.



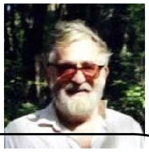
Comparison of Bayes and Empirical Bayes Approaches

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

Case	$\hat{p} = 0$	$\hat{p} = 1$
$n = 16$	820	781
$n = 60$	783	766
$n = 120$	723	747



Is empirical Bayes at least accurate asymptotically as $m \to \infty$?

Posterior model probabilities, given p:

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

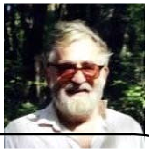
Posterior distribution of p: $\pi(p \mid \mathbf{x}) = K \sum_j p^{k_j}(1-p)^{m-k_j} m_j(\mathbf{x})$

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

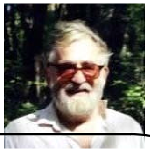
$$\begin{aligned} \int_0^1 P(\mathcal{M}_i \mid \mathbf{x}, p)\pi(p \mid \mathbf{x})dp &= \int_0^1 \frac{p^{k_i}(1-p)^{m-k_i} m_i(\mathbf{x})}{\pi(p \mid \mathbf{x})/K} \times \pi(p \mid \mathbf{x})\ dp \\ &\propto m_i(\mathbf{x}) \int_0^1 p^{k_i}(1-p)^{m-k_i} dp \propto m_i(\mathbf{x}) P(\mathcal{M}_i)\,.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

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 $\mathbf{x}_i \sim g_i(\mathbf{x}_i \mid \boldsymbol{\theta})$, for $i = 1, \ldots, n$.

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

Log-likelihood function: $l(\boldsymbol{\theta}) = \log f(\mathbf{x} \mid \boldsymbol{\theta}) = \log \left(\prod_{i=1}^{n} g_i(\mathbf{x}_i \mid \boldsymbol{\theta}) \right)$ where $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$.

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

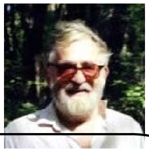
Generalization of BIC: $2l(\hat{\boldsymbol{\theta}}) - \sum_{i=1}^{p} \log(1 + n_i) + 2 \sum_{i=1}^{p} \log \frac{(1 - e^{-v_i})}{\sqrt{2}\, v_i}$,

- $v_i = \frac{\hat{\xi}_i^2}{d_i(1+n_i)}$,
- the d_i^{-1} are the eigenvalues of the observed information matrix,
- the ξ_i are the coordinates in an orthogonally transformed $\boldsymbol{\theta}$.
- n_i is the *effective sample size* corresponding to ξ. *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 $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)^t$) that $\boldsymbol{X}_l = (X_{1l}, \ldots, X_{pl})^t \sim N_p(\boldsymbol{\mu}, \sigma^2 \boldsymbol{I})$, $l = 1, \ldots, r$, so that the 'sample size' appearing in BIC should be r.
- The 'effective sample size' for each μ_i is r, but the effective sample size for σ^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 ξ and τ^2 being unknown. What is the number of parameters (see also Pauler (1998))?

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

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

(3) But, if one integrates out $\boldsymbol{\mu} = (\mu_1, \ldots, \mu_p)$, then

$$f(\boldsymbol{x} \mid \sigma^2, \xi, \tau^2) = \int f(\boldsymbol{x} \mid \boldsymbol{\mu}, \xi, \sigma^2)\pi(\boldsymbol{\mu} \mid \xi, \tau^2) d\boldsymbol{\mu}$$

$$\propto \frac{1}{\sigma^{-p(r-1)}} \exp\left\{ \frac{\hat{\sigma^2}}{2\sigma^2} \right\} \prod_{i=1}^{p} \exp\left\{ -\frac{(\bar{x}_i - \xi)^2}{2(\frac{\sigma^2}{r} + \tau^2)} \right\},$$

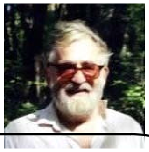
so $p = 3$ if one can work directly with $f(\boldsymbol{x} \mid \sigma^2, \xi, \tau^2)$.

Note: In this example the effective sample sizes should be $\approx p\, r$ for σ^2, $\approx p$ for ξ and τ^2, and $\approx r$ for the μ_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: $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^t \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I})$, where $\boldsymbol{X}$ is a given $n \times p$ matrix of 1's and -1's with orthogonal columns, where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^t$ and σ^2 are unknown. Then the information matrix for $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ is $\hat{\boldsymbol{I}} = \begin{pmatrix} \frac{n}{\hat{\sigma}^2} I_{p \times p} & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4} \end{pmatrix}$ so that now the effective sample size appears to be n for all parameters.

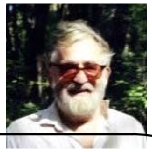
Note: The group means problem and ANOVA are linear models, so one can have effective sample sizes from $r = 1$ to n for parameters in the linear model.



Defining the 'effective sample size' n_j for ξ_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 $\boldsymbol{x}_i$ arising from the single-observation expected information matrix $\boldsymbol{I}_i^* = \boldsymbol{O}'(I_{i,jk}^*)\boldsymbol{O}$, where

$$I_{i,jk}^* = -\mathrm{E}\left[\frac{\partial^2}{\partial\theta_j\partial\theta_k}\log f_i(\boldsymbol{x}_i \mid \boldsymbol{\theta})\right]\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}.$$

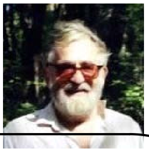


Since $I^*_{jj} = \sum_{i=1}^n I^*_{i,jj}$ is the expected information about ξ_j, a reasonable way to define n_j is

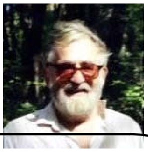
- define information weights $w_{ij} = I^*_{i,jj} / \sum_{k=1}^n I^*_{k,jj}$;
- define the effective sample size for ξ_j as

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

Intuitively, $\sum w_{ij} I^*_{i,jj}$ is a weighted measure of the information 'per observation', and dividing the total information about ξ_j by this information per case seems plausible as an effective sample size.



THANKS ALL



THANKS JOHN