Lecture 6: Data-Driven Priors

Jim Berger

Duke University

CBMS Conference on Model Uncertainty and Multiplicity
July 23-28, 2012

Outline

- Overview and an illustration
- Fractional Bayes factors
- Intrinsic Bayes factors and intrinsic priors
- Expected posterior priors

I. Overview and an Illustration

- Use of imaginary data to construct priors
- Use of actual data to construct priors

(Good, 1950, Smith and Spiegelhalter, 1980, de Voss, 1993, Gelfand and Dey, 1994, O'Hagan, 1995, 1997, Varshavsky, 1995, Berger and Pericchi, 1996,..., 2002, De Santis and Spezzaferri, 1996,1997, Dmochowski, 1996, Sansó, Pericchi and Moreno, 1996, Bertolino and Racugno, 1997, Iwaki, 1997, Gelfand and Ghosh, 1998, Lingham and Sivaganesan, 1997, 1999, Moreno, Bertolino and Racugno, 1998, 1999, Perez, 1998, Ghosh and Samanta, 1999, Key, Pericchi and Smith, 1999, Nadal, 1999, Schluter, Deely and Nicholson, 1999, Rodriguez and Pericchi, 2000, Beattie, Fong, and Lin, 2001, Berger and Perez, 2002, Neal, 2002, Casella and Moreno, ...)

Use of imaginary data to construct priors

Two Approaches:

- In constructing intrinsic and expected posterior priors (discussed later).
- In choosing normalization constants for improper objective priors (Spiegelhalter and Smith, 1982; Ghosh, 1997).

Recall: Improper objective priors π_i^O and π_j^O for parameters of models M_i and M_j yield indeterminate Bayesian answers because they can be multiplies by arbitrary constants c_i and c_j.

Proposed solution: Choose an imaginary training sample, $\boldsymbol{y}_0^*$,

1. of minimal size, such that the marginal likelihoods $m_l(\boldsymbol{y}_0^*) = \int f_l(\boldsymbol{y}_0^* \mid \boldsymbol{\theta}_l)\pi_l^O(\boldsymbol{\theta}_l)d\boldsymbol{\theta}_l < \infty\,, l = i, j\,;$
2. providing maximum possible support to the simpler model, M_i.

The authors argued that, for such a training sample, the Bayes factor of M_j to M_i should be equal to one. For $c_i \pi_i^O$ and $c_j \pi_j^O$, this means

$$1 = B_{ij} = \frac{\int f_i(\boldsymbol{y}_0^* \mid \boldsymbol{\theta}_i) c_i \pi_i^O(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i}{\int f_j(\boldsymbol{y}_0^* \mid \boldsymbol{\theta}_j) c_j \pi_i^O(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j} \, ,$$

so choose c_i and c_j so that

$$\frac{c_i}{c_j} = \frac{m_j(\boldsymbol{y}_0^*)}{m_i(\boldsymbol{y}_0^*)} \, ,$$

and then use $c_i \pi_i^O$ and $c_j \pi_j^O$ as the priors for the full data.

Notes: The choice of $\boldsymbol{y}_0^*$ depends on the models under comparison, so there is no guarantee of coherency across models, i.e., that the resulting Bayes factors satisfy

$$B_{ij} \times B_{jk} = B_{ik} \, .$$

Use of actual data to construct priors or procedures

- Through the likelihood function
  - *The absurd:* choose the prior to be the posterior arising from an improper objective prior
  - *The common but highly questionable:* choose the prior to 'span the range of the likelihood'
  - *The good:* fractional Bayes factors (discussed later)
- Through training samples
  - Intrinsic Bayes factors and intrinsic priors
  - Expected posterior priors

An Illustration of Use of Training Samples: *Intrinsic Median Posterior Probability*

(Schluter, Deely and Nicholson, 1998, and Berger and Pericchi, 1998)

Data: $X_1, X_2, \ldots, X_n$ are $N(\theta, 1)$

Models: $M_1 : \theta = 0, \quad M_2 : \theta \neq 0$

Standard objective prior:

$Pr(M_1) = Pr(M_2) = \frac{1}{2}$; under M_2, $\pi_2(\theta) = 1$.

Formal (illegitimate) Bayes factor:

$$B_{12}^O = \frac{f(\mathbf{x} \mid 0)}{\int f(\mathbf{x} \mid \theta)\,(1) d\theta} = \sqrt{n}\, e^{-\frac{n}{2}\bar{x}^2} \,.$$

Formal (illegitimate) posterior probability of M_1:

$$\Pr(M_1 \mid \mathbf{x}) = \left(1 + \frac{1}{\sqrt{n}}\, e^{\frac{n}{2}\bar{x}^2}\right)^{-1} .$$

Obtaining a proper prior by use of a training sample:

Choose one observation, say x_i, and compute

$$\pi_2(\theta \mid x_i) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(\theta - x_i)^2} \text{ (proper).}$$

Use this prior on the remaining data,
$\mathbf{x}^{(i)} \equiv (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n)$, to compute the Bayes factor

$$B_{12}^O(x_i) = \frac{f(\mathbf{x}^{(i)} \mid 0)}{\int f(\mathbf{x}^{(i)} \mid \theta)\, \pi_2(\theta \mid x_i) d\theta} = \sqrt{n}\, e^{-\frac{n}{2}\bar{x}^2} e^{\frac{1}{2}x_i^2},$$

and the posterior model probabilities

$$\Pr(M_1 \mid \mathbf{x}^{(i)}; x_i) = 1 - \Pr(M_2 \mid \mathbf{x}^{(i)}; x_i) = \left[1 + \frac{1}{\sqrt{n}}\, e^{\frac{n}{2}\bar{x}^2}\, e^{-\frac{1}{2}x_i^2}\right]^{-1}.$$

The median intrinsic posterior probability:

- Find $\Pr(M_1 \mid \mathbf{x}^{(i)}; x_i)$ for all training samples $\{x_i, i = 1, \ldots, n\}$;
- Use the median of $\Pr(M_1 \mid \mathbf{x}^{(i)}; x_i)$ (and $\Pr(M_2 \mid \mathbf{x}^{(i)}; x_i)$) over all training samples,

$$P_1^{\text{med}} = 1 - P_2^{\text{med}} = \left[1 + \frac{1}{\sqrt{n}} \, e^{\frac{n}{2}\bar{x}^2} \, e^{-\frac{1}{2} med\{x_i^2\}} \right]^{-1},$$

as the recommended conventional posterior probabilities of M_1 and M_2.

General Algorithm:

- Begin with standard objective priors, π_i^O, for the parameters θ_i in the model M_i ($\pi_i^O(\theta_i) = 1$ is okay).
- Define a "minimal training sample," $\mathbf{x}^* = (x_1^*, \ldots, x_l^*)$, as any subset of the data which is as small as possible such that the posterior distributions, $\pi_i^O(\theta_i \mid \mathbf{x}^*)$, are all proper, i.e., $m_i^O(\mathbf{x}^*) = \int f_i(\mathbf{x}^* \mid \theta_i)\pi_i^O(\theta_i)d\theta_i < \infty$. (Usually, $l = \#$ parameters in largest model.)
- Compute the Bayes factor of each model to a 'base' model M_0 (often the simplest or most complex), using the remaining data $\mathbf{x}_*$ (through the conditional likelihood $f_i(\mathbf{x}_* \mid \theta_i, \mathbf{x}^*)$) with the $\pi_i^O(\theta_i \mid \mathbf{x}^*)$ as priors.
- Do this for every possible minimal training sample, $\mathbf{x}^*$, (or a large subset) and take the median of the results.

Formula:

$$\begin{aligned} B_{i0}^{\mathrm{med}} &= \text{intrinsic median Bayes factor of } M_i \text{ to } M_0 \\ &= \underset{(\text{all } \mathbf{x}^*)}{\text{Median}} \left\{ \frac{m_i^O(\mathbf{x})\; m_0^O(\mathbf{x}^*)}{m_i^O(\mathbf{x}^*)\; m_0^O(\mathbf{x})} \right\}, \end{aligned}$$

where $m_i^O(\mathbf{x}) = \int f_i(\mathbf{x} \mid \theta_i)\pi_i^O(\theta_i)d\theta_i$. Then

$$\begin{aligned} P_i^{\mathrm{med}} &= \text{‘intrinsic median’ posterior probability of} M_i \\ &= \frac{B_{i0}^{\mathrm{med}}}{\sum_j B_{j0}^{\mathrm{med}}} \qquad \left(or \frac{\Pr(M_i)\; B_{i0}^{\mathrm{med}}}{\sum_j \Pr(M_j)\; B_{j0}^{\mathrm{med}}}\right). \end{aligned}$$

Note: When the number of possible training samples is large, one need only sample from them and take the median posterior probability over those sampled. Indeed, if n is the sample size of the data, it usually suffices to draw just ln (sets) of training samples.

Example: *Hald regression data*

Possible regressors: X_1, X_2, X_3, X_4

Full Model: is

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \epsilon, \quad \epsilon \sim N(0, \sigma^2)$$

Models under consideration: Subsets of regressors.

Notation: Model {1,3,4} means

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \beta_4 X_4 + \epsilon$$

Initial objective priors: $\pi_i(\boldsymbol{\beta}, \sigma) = 1/\sigma$

Minimal training samples: $\mathbf{y}^*$ consists of any subset of six distinct observations (since there are a maximum of six unknown parameters, including σ^2) and their covariates

Formula for m_i^O:

$$m_i^O(\mathbf{y}) = \frac{\pi^{k_i/2}\Gamma((n-k_i)/2)}{\sqrt{det(X_{(i)}^t X_{(i)})}R_i^{(n-k_i)/2}},$$

where, for model M_i, k_i is the number of regressors plus one, $X_{(i)}$ is the design matrix, and R_i is the residual sum of squares.

Answers:

model	posterior probability
{1,2,3,4}	0.049
{1,2,3}	0.171
{1,2,4}	0.190
{1,3,4}	0.160
{2,3,4}	0.041
{1,2}	0.276
{1,4}	0.108
{3,4}	0.004
others	< 0.0003

II. Fractional Bayes factors (O'Hagan 1995, 1997)

Idea: Instead of using a fraction of the data as a training sample, use a fraction of the likelihood

Algorithm:

- Choose some "fraction" $0 < b < 1$; a reasonable choice is $b = p_{\max}/n$, where n is the sample size and $p_{\max}$ is the dimension of the largest model.
- For model M_j, choose the prior
$$\pi_j^*(\boldsymbol{\theta}_j) \propto [f_j(\boldsymbol{x} \mid \boldsymbol{\theta}_j)]^b \cdot \pi_j^O(\boldsymbol{\theta}_j)$$
- Compute Bayes factors using these priors and the "remaining likelihoods " $f_j(\boldsymbol{x} \mid \boldsymbol{\theta}_j)^{(1-b)}$, yielding

$$
\begin{aligned}
B_{ji}^{FBF} &= \frac{\displaystyle\int [f_j(\boldsymbol{x} \mid \boldsymbol{\theta}_j)]^{(1-b)}\, \pi_j^*(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j}{\displaystyle\int [f_i(\boldsymbol{x} \mid \boldsymbol{\theta}_i)]^{(1-b)}\, \pi_i^*(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i} \\
&= B_{ji}^O \cdot \frac{\displaystyle\int [f_i(\boldsymbol{x} \mid \boldsymbol{\theta}_i)]^{b}\, \pi_i^O(\boldsymbol{\theta}_i) d\boldsymbol{\theta}_i}{\displaystyle\int [f_j(\boldsymbol{x} \mid \boldsymbol{\theta}_j)]^{b}\, \pi_j^O(\boldsymbol{\theta}_j) d\boldsymbol{\theta}_j}
\end{aligned}
$$

- Computationally often comparatively simple.
- Broadly applicable, except it doesn't work in irregular problems, especially problems where the sample space depends on the parameter (e.g., $X \sim U(0, \theta)$).
- Specification of different fractions, b, for different parts of the likelihood can be necessary.

III. Intrinsic Bayes Factors and Intrinsic Priors

The intrinsic Bayes Factor Approach

(Berger and Pericchi, others, ...)

Data: $\mathbf{x} = (x_1, \ldots, x_n)$

Models: $M_1, \ldots, M_q$ with densities $f_i(\mathbf{x} \mid \boldsymbol{\theta}_i), \quad i = 1, \ldots, q$

Objective priors: (usually improper) $\pi_i^O(\boldsymbol{\theta}_i)$, $i = 1, \ldots, q$

Posterior distribution for $\boldsymbol{\theta}_i$: $\pi(\boldsymbol{\theta}_i \mid \boldsymbol{x})$

Marginal likelihoods for M_i: $m_i^O(\mathbf{x}) = \int f_i(\mathbf{x} \mid \boldsymbol{\theta}_i)\pi_i^O(\boldsymbol{\theta}_i)\, d\boldsymbol{\theta}_i$.

Definition 1 A training sample, $\mathbf{x}(l)$, is called *proper* if $0 < m_i^N(\mathbf{x}(l)) < \infty$ for all M_i, and *minimal* if it is proper and no subset is proper.

Basic idea: For a minimal training sample $\mathbf{x}(l)$, use the (proper) posteriors $\pi(\boldsymbol{\theta}_i \mid \mathbf{x}(l))$ as priors, to compute Bayes factors for the rest of the data, denoted by $\mathbf{x}(-l)$.

The resulting Bayes factors:

$$B_{ji}(l) = \frac{\int f_j(\boldsymbol{x}(-l) \mid \theta_j, \boldsymbol{x}(l))\ \pi_j^O(\theta_j \mid \boldsymbol{x}(l))\, d\theta_j}{\int f_i(\boldsymbol{x}(-l) \mid \theta_i, \boldsymbol{x}(l))\ \pi_j^O(\theta_i \mid \boldsymbol{x}(l))\, d\theta_i} = B_{ji}^O(\mathbf{x}) \cdot B_{ij}^O(\mathbf{x}(l)),$$

where

$$B_{ji}^O = B_{ji}^O(\mathbf{x}) = \frac{m_j^O(\mathbf{x})}{m_i^O(\mathbf{x})} \qquad \text{and} \qquad B_{ij}^O(l) = B_{ij}^O(\mathbf{x}(l)) = \frac{m_i^O(\mathbf{x}(l))}{m_j^O(\mathbf{x}(l))}.$$

Now ‘average over all possible training samples. Possible averages:

Arithmetic IBF (AIBF):

$$B_{ji}^{AI} = B_{ji}^O(\boldsymbol{x}) \cdot \tfrac{1}{L} \textstyle\sum_{l=1}^{L} B_{ij}^O(\boldsymbol{x}(l)),$$

Median IBF (MIBF):

$$B_{ji}^{MI} = B_{ji}^O(\boldsymbol{x}) \cdot \text{Median}\left\{B_{ij}^O(\boldsymbol{x}(l))\right\}$$

Geometric IBF (GIBF):

$$B_{ji}^{GI} = B_{ji}^O(\boldsymbol{x}) \cdot \left[\textstyle\prod_{l=1}^{L} B_{ij}^O(\boldsymbol{x}(l))\right]^{1/L}$$

Notes:

1. Averages can be based on a random subset of all minimal training samples; indeed $n \times p_{\max}$ minimal training samples, where n is the sample size and $p_{\max}$ is the dimension of the largest model, typically suffices.
2. Computation of the $m_j^O(\mathbf{x}(l))$ can be challenging, because Laplace approximations do not work. Hence IBF's are most used when the $m_j^O(\mathbf{x}(l))$ are closed form.
3. With large model spaces and large n, even closed form marginals leave a challenging computation.

Example: Normal Mean

$$M_1 : x \sim N(x \mid 0, \sigma_1^2) \qquad M_2 : x \sim N(x \mid \mu, \sigma_1^2)$$

Objective priors: $\pi_1^O(\sigma_1) = 1/\sigma_1$ and $\pi_2^O(\mu, \sigma_2) = 1/\sigma_2^2$.

(Note that π_2^O is not the usual prior, but gives simpler expressions.)

Minimal training samples: $\mathbf{x}(l) = (x_i, x_j)$ (distinct)

Then

$$m_1^O(\mathbf{x}(l)) = \frac{1}{2\pi(x_i^2 + x_j^2)}, \quad m_2^O(\mathbf{x}(l)) = \frac{1}{\sqrt{\pi}(x_i - x_j)^2}.$$

The formal Bayes factor for full data $\mathbf{x}$, when using π_1^O and π_2^O directly:

$$B_{21}^O = \sqrt{\frac{2\pi}{n}} \cdot \left(1 + \frac{n\bar{x}^2}{s^2}\right)^{n/2},$$

where $s^2 = \sum_{i=1}^{n}(x_i - \overline{x})^2$. Thus the AIBF is

$$B_{21}^{AI} = B_{21}^{O} \cdot \frac{1}{L} \sum_{l=1}^{L} \frac{(x_1(l) - x_2(l))^2}{2\sqrt{\pi}[x_1^2(l) + x_2^2(l)]} \,.$$

Intrinsic priors (for AIBF's)

Key Question: Does the AIBF correspond (for large sample sizes) to an actual Bayes factor; if so, the priors associated with the actual Bayes factor are called the *intrinsic priors* for the AIBF.

Finding intrinsic priors: Suppose

(i) Under M_j, $\hat{\boldsymbol{\theta}}_j \to \boldsymbol{\theta}_j$, $\hat{\boldsymbol{\theta}}_i \to \psi_i(\boldsymbol{\theta}_j)$, $\sum_{l=1}^{L} B_{ij}^{O}(\boldsymbol{x}(l)) \to B_j^*(\boldsymbol{\theta}_j)$

(ii) Under M_i, $\hat{\boldsymbol{\theta}}_i \to \boldsymbol{\theta}_i$, $\hat{\boldsymbol{\theta}}_j \to \psi_j(\boldsymbol{\theta}_i)$, $\sum_{l=1}^{L} B_{ij}^{O}(\boldsymbol{x}(l)) \to B_i^*(\boldsymbol{\theta}_i)$

When dealing with the AIBF, it will typically be the case that, for $k = i$ or $k = j$,

$$B_k^*(\boldsymbol{\theta}_k) = \lim_{L\to\infty} E_{\boldsymbol{\theta}_k}^{M_k}\left[\frac{1}{L}\sum_{l=1}^{L} B_{ij}^{N}(l)\right];$$

if the $\mathbf{X}(l)$ are exchangeable, then the limits and averages over L can be removed.

Then it can be shown that the *intrinsic prior* (π_j^I, π_i^I) is given by the solutions to the equations

$$\begin{aligned} \frac{\pi_j^I(\boldsymbol{\theta}_j)\pi_i^N(\psi_i(\boldsymbol{\theta}_j))}{\pi_j^N(\boldsymbol{\theta}_j)\pi_i^I(\psi_i(\boldsymbol{\theta}_j))} &= B_j^*(\boldsymbol{\theta}_j), \\ \frac{\pi_j^I(\psi_j(\boldsymbol{\theta}_i))\pi_i^N(\boldsymbol{\theta}_i)}{\pi_j^N(\psi_j(\boldsymbol{\theta}_i))\pi_i^I(\boldsymbol{\theta}_i)} &= B_i^*(\boldsymbol{\theta}_i). \end{aligned} \tag{1}$$

Normal Example: Computation and solution of the equations yields

$$\begin{aligned} \pi_1^I(\sigma_1) &= \frac{1}{\sigma_1} \\ \pi_2^I(\mu, \sigma_2) &= \frac{1}{\sigma_2} \times \frac{1 - \exp[-\mu^2/\sigma_2^2]}{2\sqrt{\pi}(\mu^2/\sigma_2)}. \end{aligned}$$

This last conditional distribution is proper (integrating to one over μ) and, furthermore, is very close to the Cauchy$(0, \sigma_2)$ choice of $\pi_2(\mu|\sigma_2)$ suggested by Jeffreys (1961).

George Casella, 1951-2012

- Forefront of development of Intrinsic priors
  - application to many scenarios
  - first proofs of consistency
  - robust IP bounds
- p-values and Bayes
- Conditional frequentist theory
- Many computational innovations

No Need to Spend α in Interim Analysis:

Data: d_i is the observed treatment difference for subject i treated with two hypotensive agents (Robertson and Armitage, 1959; Armitage, 1975). (Here t_i [s_i] denotes the t-statistic [sample standard deviation] after observation i.)

Model: The d_i are i.i.d $Normal(\theta, \sigma^2)$, $i = 1, \ldots$.

To Test: $H_1 : \theta = 0$ versus $H_2 : \theta < 0$ versus $H_3 : \theta > 0$.

Frequentist analysis:

- Choose a stopping rule and decision rule; e.g., if doing a two-sided test, the Siegmund (1977) sequential t-test *stops the experiment when* $|t_i| > c(i)$ *and rejects* H_1 .
- Controls the associated Type I error probability.

Objective Bayesian analysis: $\Pr(H_j) = 1/3$; noninformative prior for (θ, σ^2) appropriately 'trained' ('Encompassing Intrinsic Bayes Factors': Berger & Mortera, 1999 JASA).

Objective Posterior Probabilities $\Pr_j(i)$ of H_j at observation i:

$$\begin{aligned}
\Pr_1(i) &= \left[1 + \frac{s_1(i)}{\tau_{i-1}(t_i)}\left(\frac{1 - T_{i-1}(t_i)}{s_2} + \frac{T_{i-1}(t_i)}{s_3(i)}\right)\right]^{-1}, \\
\Pr_2(i) &= \left[1 + \frac{s_2}{1 - T_{i-1}(t_i)}\left(\frac{\tau_{i-1}(t_i)}{s_1(i)} + \frac{T_{i-1}(t_i)}{s_3(i)}\right)\right]^{-1},
\end{aligned}$$

and $\Pr_3(i) = 1 - \Pr_1(i) - \Pr_2(i)$, where τ_{i-1} and T_{i-1} are the density and c.d.f. of the standard t-distribution with $(i-1)$ degrees of freedom, $s_3(i) = \pi i(i-1) - s_2$,

$$s_1(i) = \frac{s_i}{\sqrt{i}} \sum_{k \neq l} \frac{|d_k - d_l|}{{d_k}^2 + {d_l}^2 + \epsilon}, \;\; s_2 = \sum_{k \neq l} \left(\frac{\pi}{2} - \arctan(\frac{-(d_k + d_l)}{|d_k - d_l + \epsilon|})\right).$$

($\epsilon \approx 0$ is introduced to avoid numerical indeterminacy)

Pair	Difference	t-statistic	Posterior Probabilities		
n	d_i	t	$\Pr_1$	$\Pr_2$	$\Pr_3$
1	95	-	-	-	-
2	-20	0.652	0.333	0.333	0.333
3	41	1.16	0.357	0.237	0.407
4	-10	1.00	0.431	0.157	0.412
5	1	1.01	0.360	0.148	0.492
6	12	1.15	0.342	0.142	0.516
7	11	1.26	0.348	0.132	0.519
8	-2	1.23	0.276	0.136	0.589
9	6	1.30	0.283	0.130	0.587
10	14	1.44	0.295	0.115	0.590
11	19	1.63	0.291	0.095	0.615
12	71	2.05	0.203	0.058	0.739
13	-9	1.92	0.229	0.058	0.713
14	7	1.97	0.225	0.054	0.721
15	-19	1.74	0.294	0.061	0.646
20	-9	1.51	0.387	0.056	0.557
25	0	1.35	0.465	0.060	0.475
30	-3	0.831	0.620	0.079	0.301
35	0	0.339	0.669	0.112	0.219
40	0	0.056	0.698	0.134	0.168
45	-13	0.099	0.714	0.125	0.162
50	-3	-0.202	0.736	0.141	0.123
53	-37	-0.396	0.740	0.157	0.103

Comments

(i) Neither multiple hypotheses nor the sequential aspect caused difficulties. There is no penalty (e.g., 'spending α') for looks at the data.

(ii) Quantification of the support for $H_1 : \theta = 0$ is direct. At the 12th observation, $t = 2.05$ but $\Pr_1 = 0.203$. At the end, $\Pr_1 = 0.740$.

(iii) At the 12th observation, $\Pr_2 = 0.058$, so H_2 can be effectively ruled out.

(iv) For testing $H_1 : \theta = 0$ versus $H_2 : \theta \neq 0$, the $\Pr_i$ are conditional frequentist error probabilities.

IV. Expected posterior priors

Expected Posterior Priors (Perez, 1998, Perez and Berger, 2001, 2002, Neal, 2002)

Initial priors: $\pi_i^O(\boldsymbol{\theta}_i)$, typically improper

Initial marginals: $m_i^O(\boldsymbol{y}) = \int f_i(\boldsymbol{y} \mid \boldsymbol{\theta}_i)\pi_i^O(\boldsymbol{\theta}_i)d\boldsymbol{\theta}_i$

Training sample posteriors: Consider a training sample, $\boldsymbol{y}^*$, such that the posterior distributions

$$\pi_i^O(\boldsymbol{\theta}_i \mid \boldsymbol{y}^*) = \frac{f_i(\boldsymbol{y}^* \mid \boldsymbol{\theta}_i)\,\pi_i^O(\boldsymbol{\theta}_i)}{m_i^O(\boldsymbol{y}^*)}$$

exist, for $i = 1, \ldots, k$.

Definition: The prior densities

$$\pi_i^*(\boldsymbol{\theta}_i) = \int \pi_i^O(\boldsymbol{\theta}_i \mid \boldsymbol{y}_{(i)}^*)\, m^*(\boldsymbol{y}^*) d\boldsymbol{y}^*,$$

where $\boldsymbol{y}_{(i)}^*$ is a minimal random subsample of $\boldsymbol{y}^*$ such that the $\pi_i^O(\boldsymbol{\theta}_i \mid \boldsymbol{y}_{(i)}^*)$ exist, will be called the *expected posterior priors* (or EP priors) for the $\boldsymbol{\theta}_i$, with respect to m^*.

Note: The EP priors, $\pi_i^*(\boldsymbol{\theta}_i)$, will not be proper unless m^* itself is proper, but are always properly 'calibrated' across models.

Choices of m^* :

- A subjectively elicited marginal distribution
- If M_0 is a model nested in all others, set $m^*(\boldsymbol{y}^*) = m_0^O(\boldsymbol{y}^*)$.
  - Then the EP prior is identical to the 'intrinsic prior.'
- The empirical distribution

Choosing m^* to be the empirical distribution: Given observations $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n$, let

$$m^*(\boldsymbol{y}^*) = \frac{1}{L} \sum_l I_{\{\boldsymbol{y}(l)\}}(\boldsymbol{y}^*),$$

where $\boldsymbol{y}(l) = (y_{l_1}, \ldots, y_{l_m})$ is a subsample of size $0 < m < n$ such that $\pi_i^O(\boldsymbol{\theta}_i \mid \boldsymbol{y}(l))$ exists for all models M_i, and L is the number of such subsamples of size m.

Computation: Introduce $\boldsymbol{y}^*$ as latent variables, effectively replacing $\pi_i^*(\boldsymbol{\theta}_i) = \int \pi_i^O(\boldsymbol{\theta}_i \mid \boldsymbol{y}^*_{(i)})\, m^*(\boldsymbol{y}^*) d\boldsymbol{y}^*$ by

$$\pi_i^O(\boldsymbol{\theta}_i \mid \boldsymbol{y}^*_{(i)})\, m^*(\boldsymbol{y}^*) = \frac{\pi_i^O(\boldsymbol{\theta}_i) f_i(\boldsymbol{y}^*_{(i)} \mid \boldsymbol{\theta}_i)}{m_i^O(\boldsymbol{y}^*_{(i)})}\, m^*(\boldsymbol{y}^*)\,.$$

A default prior for testing a point null

This uses the *intrinsic* or *expected posterior* prior construction. For i.i.d. observations $\boldsymbol{x} = (x_1, \ldots, x_n)$ from a density $f(x \mid \theta)$, and for testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta \neq \theta_0$,

- let $\pi^O(\theta)$ be a good estimation objective prior, so that $\pi^O(\theta \mid \boldsymbol{x}) = f(\boldsymbol{x} \mid \theta)\pi^O(\theta)/m^O(\boldsymbol{x})$ is the resulting posterior, and $m^O(\boldsymbol{x}) = \int f(\boldsymbol{x} \mid \theta)\pi^O(\theta)\, d\theta$;

- then the intrinsic prior (which will be proper) is

$$\pi^I(\theta) = \int \pi^O(\theta \mid \boldsymbol{x}^*) f(\boldsymbol{x}^* \mid \theta_0)\, d\boldsymbol{x}^* \,,$$

  with $\boldsymbol{x}^* = (x_1^*, \ldots, x_q^*)$ being (unobserved) data of the minimal sample size q such that $m^O(\boldsymbol{x}^*) < \infty$.

- The resulting Bayes factor is

$$B_{01}(\boldsymbol{x}) = \frac{f(\boldsymbol{x} \mid \theta_0)}{\int f(\boldsymbol{x} \mid \theta)\pi^I(\theta)d\theta} = \frac{f(\boldsymbol{x} \mid \theta_0)}{\int m^O(\boldsymbol{x} \mid \boldsymbol{x}^*)f(\boldsymbol{x}^* \mid \theta_0)d\boldsymbol{x}^*} .$$

Example: Test $H_0 : \theta = 0$ versus $H_0 : \theta > 0$, based on $X_i \sim f(x_i \mid \theta) = (\theta + b)\exp\{-(\theta + b)x_i\}$, where b is known;

- Suppose we choose $\pi^O(\theta) = 1/(\theta + b)$ (the more natural square root is harder to work with).
- A minimal sample size for the resulting posterior to be proper is $q = 1$.
- Computation then yields $\pi^I(\theta) = \int \pi^O(\theta \mid x_1^*)f(x_1^* \mid 0)dx_1^* = b/(\theta + b)^2$.

Application: In the search for the Higgs boson, we observe $N = \text{Poisson}(s+b)$, where s is the count rate from 'signal' events and b is the known 'background' count rate.

To Test: $H_0 : s = 0$ versus $H_1 : s > 0$.

Intrinsic prior: To obtain a minimal sample corresponding to a single Poisson observation, Berger and Pericchi (2004 AOS) suggest using a single observation from the equivalent exponential inter-arrival time process, here $X^* \sim (\theta + b)e^{-x^*(\theta+b)}$. Then $\pi^I(\theta) = \int \pi^O(\theta \mid x^*) f(x^* \mid 0) dx^* = b/(\theta+b)^2$.

Bayes factor of H_0 to H_1:

$$B_{01} = \frac{b^O \ e^{-b}}{\int_0^\infty (s+b)^O \ e^{-(s+b)} \pi^I(s) \ ds} = \frac{b^{(n-1)} \ e^{-b}}{\Gamma(n-1, b)} \,,$$

where Γ is the incomplete gamma function.

Application to mixture models with an unknown number of bivariate normal components

The model is given by

$$p(k, \boldsymbol{w}, \boldsymbol{z}, \boldsymbol{\theta}, \boldsymbol{y}) = p(k)p(\boldsymbol{w} \mid k)p(\boldsymbol{z} \mid \boldsymbol{w}, k)p(\boldsymbol{\theta} \mid k)f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{z}) \,,$$

- k represents the unknown number of components;
- $\boldsymbol{w} = (w_1, \ldots, w_k)$, where w_j is the probability of an observation coming from component i;
- $\boldsymbol{z} = (z_1, \ldots, z_n)$, where z_i indicates that observation $\boldsymbol{y}_i$ comes from component z_i;
- $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k)$, with $\boldsymbol{\theta}_i$ the parameter for component i.

The distributions are given by

- $p(k)$ is the prior probability of k components (default is uniform over some range).
- $p(\boldsymbol{w} \mid k)$ is a Dirichlet distribution with known parameter $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_0)$ (default is $\alpha_0 = 1/2$).
- $\boldsymbol{z}_i$ are i.i.d. with $p(z_i = j \mid \boldsymbol{w}, k) = w_j$.
- The likelihood is $f(\boldsymbol{y} \mid \boldsymbol{\theta}, \boldsymbol{z}) = \prod_1^O f(\boldsymbol{y}_i \mid \boldsymbol{\theta}_{z_i})$.
- The initial (non-trained) prior for the parameters is $p(\boldsymbol{\theta} \mid k) = \prod_1^k \pi^O(\boldsymbol{\theta}_j)$, with improper priors $\pi^O(\cdot)$.

To avoid problems with identifying the components, we order the first coordinate of the means in the application.

Based on minimal training samples $\boldsymbol{Y}^*$ for a single component, the expected posterior priors are given by

$$\pi^*(\boldsymbol{\theta} \mid k) = \int \prod_1^k \pi^O(\boldsymbol{\theta}_j \mid \boldsymbol{y}^*) m^*(\boldsymbol{y}^*) d\boldsymbol{y}^*$$

The Reversible Jump MCMC method described in Richardson and Green 96 can be used for this model with the following modifications for generating from the posterior of each $\boldsymbol{\theta}_j$:

- Define $u^*(\boldsymbol{y}^* \mid \boldsymbol{y}, \boldsymbol{z}, \ldots) \propto m^*(\boldsymbol{y}^*) \prod_1^k m^O(\boldsymbol{y}_j, \boldsymbol{y}^*)/m^O(\boldsymbol{y}^*)$. Here $m^O(\cdot)$ is the marginal for $f(\cdot \mid \boldsymbol{\theta})\pi^O(\boldsymbol{\theta})$.

- Generate a new $\boldsymbol{y}^*_{(new)}$ using a Metropolis-Hastings algorithm. For generating from the transition probabilities we use
  1. Generate $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ from $\prod_1^k \pi^O(\theta_j \mid \boldsymbol{y}_j, \boldsymbol{y}^*_{(t)})$.
  2. Generate $\boldsymbol{y}^*_{(t+1)}$ from $\sum_1^k w_j f(\cdot \mid \boldsymbol{\theta}_j)$.
- Generate $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_k$ from $\prod_1^k \pi^O(\theta_j \mid \boldsymbol{y}_j, \boldsymbol{y}^*_{(new)})$.

With this approach, $m^*(\cdot)$ in fact acts as a hierarchical common improper prior for all components. A nice property of this approach is that we do not need to restrict the number of observations per component, as for example in Diebolt and Robert 94. Hence the allocations $\boldsymbol{z}$ are independent a posteriori, making the inference much easier.

BATSE gamma ray burst data set: We analyze 745 measurements taken by the Compton Gamma Ray Observatory between 1991 and 1994 (third catalogue). Of interest is the relationship of the duration (T90) and hardness ratio (HR) of the bursts. Thus it is bivariate data

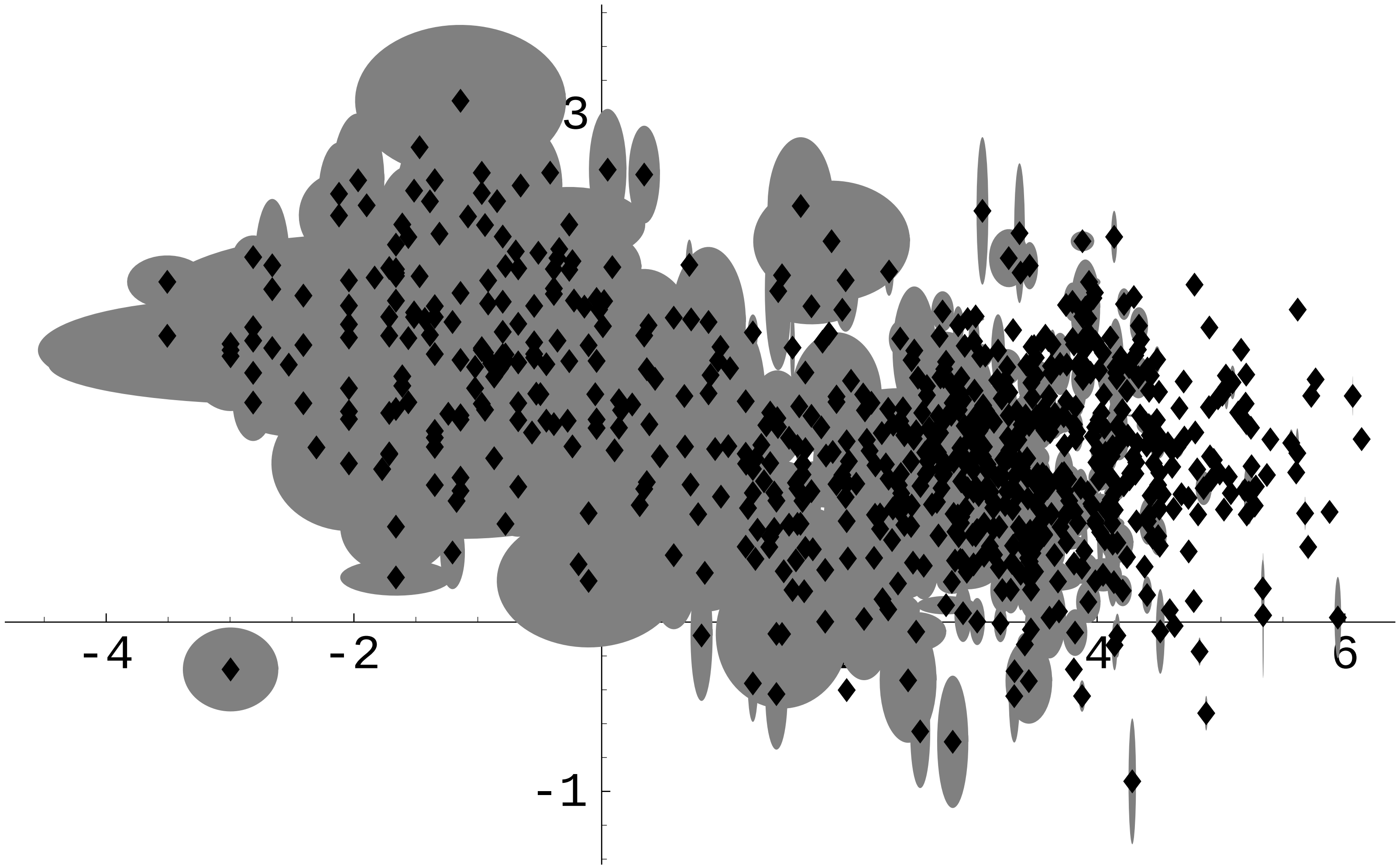
$$\boldsymbol{x}_i = (x_{i1}, x_{i2}) = (\log(\mathrm{T90})_i, \log(\mathrm{HR})_i)$$

with standard errors $\boldsymbol{\sigma}_i = (\sigma_{i1}, \sigma_{i2}) = (\sigma_{T90_i}, \sigma_{HR_i})$.

The true gamma ray burst values, $\boldsymbol{y}_i = (y_{i1}, y_{i2})$, are assumed to arise from a mixture of k bivariate normal distributions, so we have

$$\boldsymbol{x}_i \sim N(\boldsymbol{x}_i \mid \boldsymbol{y}_i, \boldsymbol{\sigma}_i) \quad \text{and} \quad \boldsymbol{y}_i \sim \sum_{j=1}^{k} w_j N(\boldsymbol{y}_i \mid \boldsymbol{\mu}_j, \Sigma_j).$$

Standard initial objective priors were used to develop the expected posterior priors.


3
-4
-2
4
6
-1

MCMC:

- An additional step was added to generate $\boldsymbol{y}_i$ from

$$p(\boldsymbol{y}_i \mid \cdots) \propto N(\boldsymbol{x}_i \mid \boldsymbol{y}_i, \boldsymbol{\sigma}_i) \times N(\boldsymbol{y}_i \mid \boldsymbol{\mu}_{z_i}, \boldsymbol{\Sigma}_{z_i}).$$

- 100,000 iterations, with convergence judged informally.

Results:

- $P(k = 2 \mid \boldsymbol{y}) = .99$.
- Table 1 gives the corresponding estimates of the location and covariance matrices for two components.
- Figure 1 shows the allocation distribution for the gamma ray bursts, along with predictive confidence sets of levels 90%, 95% and 99% for the two components.

Base model EP priors		Empirical EP priors	
Component 1		Group 1	
$\hat{w}_1 = 0.24$		$\hat{w}_1 = 0.24$	
$\hat{\mu}_{T90} = -0.85$	$\hat{\mu}_{HR} = 1.61$	$\hat{\mu}_{T90} = -0.92$	$\hat{\mu}_{HR} = 1.62$
$\hat{\sigma}_{T90} = 1.04$	$\hat{\sigma}_{HR} = 0.50$	$\hat{\sigma}_{T90} = 0.98$	$\hat{\sigma}_{HR} = 0.50$
$\hat{\rho} = -0.03$		$\hat{\rho} = -0.02$	
Component 2		Group 2	
$\hat{w}_2 = 0.76$		$\hat{w}_2 = 0.76$	
$\hat{\mu}_{T90} = 3.31$	$\hat{\mu}_{HR} = 0.95$	$\hat{\mu}_{T90} = 3.31$	$\hat{\mu}_{HR} = 0.95$
$\hat{\sigma}_{T90} = 1.10$	$\hat{\sigma}_{HR} = 0.49$	$\hat{\sigma}_{T90} = 1.10$	$\hat{\sigma}_{HR} = 0.49$
$\hat{\rho} = 0.01$		$\hat{\rho} = 0.01$	

Table 1: BATSE: Estimates for log(T90) and log(HR).

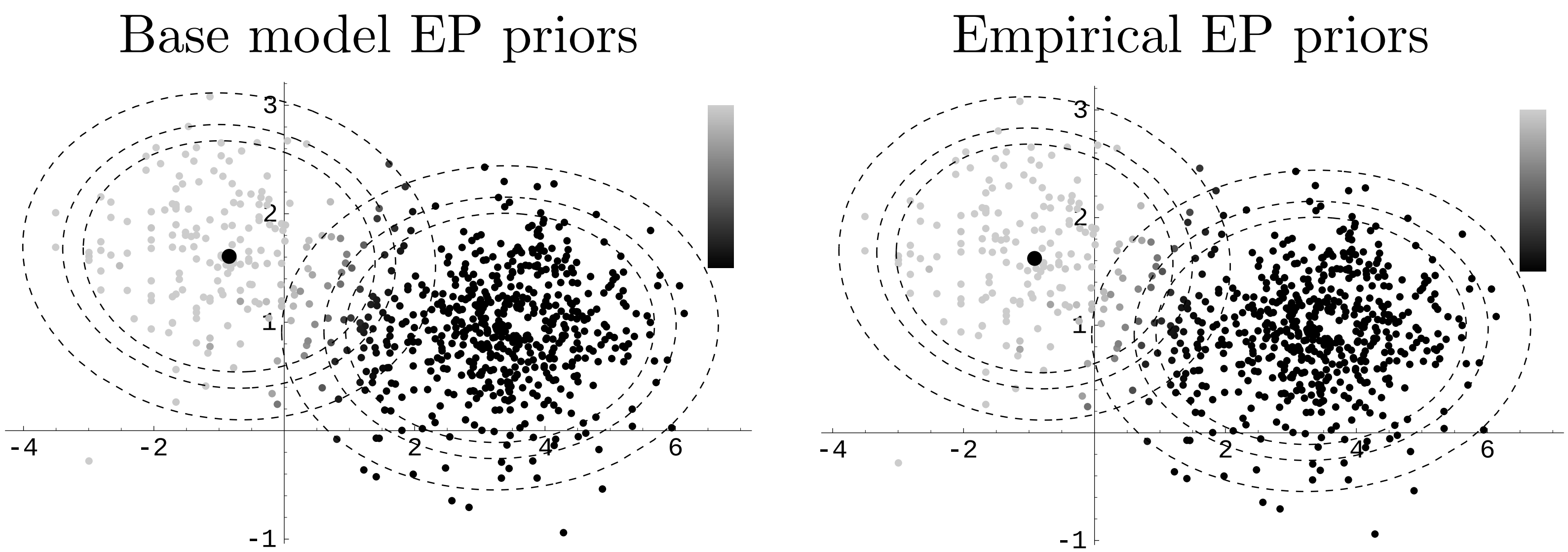


Figure 1: BATSE classification probabilities. Color bar indicates value of $p(z_i = 2 \mid \boldsymbol{y})$.