Tutorials on Objective Bayesian Analysis

12:00 noon-1:30 Tutorial A
- TA1. Introduction and History (Jim Berger)
- TA2. Reference Analysis (Jose Bernardo)

1:45-3:15 Tutorial B
- TB1. Other Proposed Objective Priors (Jim Berger)
- TB2. Eliminating Nuisance Parameters: Comparison with Empirical Bayes and Likelihood Methods (Brunero Liseo)

3.30-5:00 Tutorial C
- TC1. Hypothesis Testing (M.J. Bayarri)
- TC2. Model Selection (M.J. Bayarri)

5.15-6:45pm Tutorial D
- TD1. Linear Mixed Models (Dongchu Sun)
- TD2. Spatial and Temporal Models (Dongchu Sun)
Introduction and History of Objective Bayesian Analysis

James O. Berger
Duke University and the
Statistical and Applied Mathematical Sciences Institute

Branson, Missouri
June 4, 2005
A. Introduction to Objective Bayesian Analysis

Objective Bayesian analysis proceeds by

- modeling the data probabilistically;
- representing initial uncertainty about unknown features of the model for the data using ‘noninformative’ prior probability distributions;
- using probability theory (usually Bayes theorem) to find the posterior probability distribution of quantities of interest, given the data.
A Medical Diagnosis Example (with Mossman, 2001)

The Medical Problem:

- Within a population, $p_0 = Pr(\text{Disease } D)$.
- A diagnostic test results in either a Positive (P) or Negative (N) reading.
- $p_1 = Pr(P | \text{patient has } D)$.
- $p_2 = Pr(P | \text{patient does not have } D)$.

It follows from Bayes theorem that

$$\theta = Pr(D | P) = \frac{p_0 p_1}{p_0 p_1 + (1 - p_0)p_2}.$$
The Statistical Problem: The $p_i$ are unknown. Based on (independent) data $X_i \sim \text{Binomial}(n_i, p_i)$ (arising from medical studies), find a $100(1 - \alpha)\%$ confidence set for $\theta$.

Suggested Solution: Assign $p_i$ the Jeffreys-rule prior

$$\pi(p_i) \propto p_i^{-1/2} (1 - p_i)^{-1/2}$$

(superior to the uniform prior $\pi(p_i) = 1$). By Bayes theorem, the posterior distribution of $p_i$ given the data, $x_i$, is

$$\pi(p_i \mid x_i) = \frac{p_i^{-1/2} (1 - p_i)^{-1/2} \times \binom{n}{x_i} p_i^{x_i} (1 - p_i)^{n_i-x_i}}{\int p_i^{-1/2} (1 - p_i)^{-1/2} \times \binom{n}{x_i} p_i^{x_i} (1 - p_i)^{n_i-x_i} \, dp_i},$$

which is the Beta($x_i + \frac{1}{2}, n_i - x_i + \frac{1}{2}$) distribution.
Finally, compute the desired confidence set (formally, the $100(1 - \alpha)\%$ equal-tailed posterior credible set) by

- drawing random $p_i$ from the Beta($x_i + \frac{1}{2}, n_i - x_i + \frac{1}{2}$) distributions, $i = 0, 1, 2$;
- computing the associated $\theta = p_0p_1/[p_0p_1 + (1 - p_0)p_2]$;
- repeating this process 10,000 times;
- using the $\frac{\alpha}{2}\%$ upper and lower percentiles of these generated $\theta$ to form the desired confidence limits.
\[ n_0 = n_1 = n_2 = (x_0, x_1, x_2) \]

95% confidence interval

<table>
<thead>
<tr>
<th>(n_0 = n_1 = n_2)</th>
<th>((x_0, x_1, x_2))</th>
<th>95% confidence interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>20</td>
<td>(2,18,2)</td>
<td>(0.107, 0.872)</td>
</tr>
<tr>
<td>20</td>
<td>(10,18,0)</td>
<td>(0.857, 1.000)</td>
</tr>
<tr>
<td>80</td>
<td>(20,60,20)</td>
<td>(0.346, 0.658)</td>
</tr>
<tr>
<td>80</td>
<td>(40,72,8)</td>
<td>(0.808, 0.952)</td>
</tr>
</tbody>
</table>

Table 1: The 95% equal-tailed posterior credible interval for \(\theta = p_0 p_1 / [p_0 p_1 + (1 - p_0) p_2]\), for various values of the \(n_i\) and \(x_i\).
B. A Brief History of Objective Bayesian Analysis
The Reverend Thomas Bayes, began the objective Bayesian theory, by solving a particular problem

- Suppose $X$ is Binomial $(n,p)$; an ‘objective’ belief would be that each value of $X$ occurs equally often.
- The only prior distribution on $p$ consistent with this is the uniform distribution.
- Along the way, he codified Bayes theorem.
- Alas, he died before the work was finally published in 1763.
The real inventor of Objective Bayes was Simon Laplace (also a great mathematician, astronomer and civil servant) who wrote *Théorie Analytique des Probabilités* in 1812

- He virtually always utilized a ‘constant’ prior density (and clearly said why he did so).
- He established the ‘central limit theorem’ showing that, for large amounts of data, the posterior distribution is asymptotically normal (and the prior does not matter).
- He solved very many applications, especially in physical sciences.
- He had numerous methodological developments, e.g., a version of the Fisher exact test.
What’s in a name, part I

- It was called *probability theory* until 1838.
- From 1838-1950, it was called *inverse probability*, apparently so named by Augustus de Morgan.
- From 1950 on it was called *Bayesian analysis* (as well as the other names).

Augustus de Morgan
The importance of inverse probability b.f. (before Fisher): as an example, Egon Pearson in 1925 finding the ‘right’ objective prior for a binomial proportion

- Gathered a large number of estimates of proportions $p_i$ from different binomial experiments
- Treated these as arising from the predictive distribution corresponding to a fixed prior.
- Estimated the underlying prior distribution (an early empirical Bayes analysis).
- Recommended something close to the currently recommended ‘Jeffreys prior’ $p^{-1/2}(1-p)^{-1/2}$. 
Fig. 3. Distribution of Frequencies of $\frac{\chi+\tau}{\alpha+\beta}$ in 100 samples (made symmetrical).
1930’s: ‘inverse probability’ gets ‘replaced’ in mainstream statistics by two alternatives

- For 50 years, Boole, Venn and others had been calling use of a constant prior logically unsound (since the answer depended on the choice of the parameter), so alternatives were desired.
- R.A. Fisher’s developments of ‘likelihood methods,’ ‘fiducial inference,’ … appealed to many.
- Jerzy Neyman’s development of the frequentist philosophy appealed to many others.
Harold Jeffreys (also a leading geophysicist) revived the Objective Bayesian viewpoint through his work, especially the *Theory of Probability* (1937, 1949, 1963)

- The now famous *Jeffreys prior* yielded the same answer no matter what parameterization was used.
- His priors yielded the ‘accepted’ procedures in all of the standard statistical situations.
- He began to subject Fisherian and frequentist philosophies to critical examination, including his famous critique of p-values: “An hypothesis, that may be true, may be rejected because it has not predicted observable results that have not occurred.”
What’s in a name, part II

• In the 50’s and 60’s the subjective Bayesian approach was popularized (de Finetti, Rubin, Savage, Lindley, …)

• At the same time, the objective Bayesian approach was being revived by Jeffreys, but Bayesianism became incorrectly associated with the subjective viewpoint. Indeed,
  – only a small fraction of Bayesian analyses done today heavily utilize subjective priors;
  – objective Bayesian methodology dominates entire fields of application today.
What’s in a name, part III

- Some contenders for the name (other than Objective Bayes):
  - Probability
  - Inverse Probability
  - Noninformative Bayes
  - Default Bayes
  - Vague Bayes
  - Matching Bayes
  - Non-subjective Bayes

- But ‘objective Bayes’ has a website and soon will have *Objective Bayesian Inference* (coming soon to a bookstore near you)
C. Current State of Objective Bayesian Analysis

It is used all the time in Bayesian practice and has become central to numerous application areas and a number of other disciplines. Much of it is *ad hoc objective Bayesian analysis*, characterized by use of one of the following:

- A constant prior
- Vague proper priors
- Proper priors chosen over ‘reasonable ranges’ or in a data-dependent fashion.
Ad hoc objective Bayesian analysis can be successful, especially if validated by experience or extensive sensitivity studies, but there are potential problems:

- A constant prior can yield improper posteriors (see, e.g., Berger, De Oliveira and Sanso, 2001) or can swamp the data in high dimensions (e.g., large sparse contingency tables).

- Vague proper priors are often even worse (e.g., the usual prior used with BUGS for a hyper-variance results in a posterior very concentrated at zero, when the data suggests quite the opposite).

- Ad hoc proper priors can strongly affect the answer in unintended ways (producing too-tight posteriors).
In contrast, *Formal objective Bayesian analysis* seeks methodology that comes with a ‘guarantee’ of success.

**First effort:** Egon Pearson’s effort to empirically determine the objective prior for a binomial parameter.

**Second effort:** The *Jeffreys-rule prior* was developed by Harold Jeffreys to produce objective Bayesian answers that were invariant to the parameterization used for the problem. If the data model density is \( p(x | \theta) \) the Jeffreys-rule prior for the unknown \( \theta = \{\theta_1, \ldots, \theta_k\} \) has the form

\[
|I(\theta)|^{1/2}d\theta_1 \ldots d\theta_k
\]

where \( I(\theta) \) is the \( k \times k \) matrix Fisher’s information matrix with \((i, j)\) element

\[
I(\theta)_{ij} = E_{x | \theta} \left[ -\frac{\partial^2}{\partial \theta_i \partial \theta_j} \log p(x | \theta) \right].
\]
D. Potential of Objective Bayesian Analysis

- Unification of statistics
  - Objective Bayesian methods typically have very good frequentist properties.
  - Objective Bayesian answers solve the two major problems facing frequentists:
    * How to properly condition.
    * What to do for small sample sizes.

- Objective Bayesian methodology can be used in the (very many) situations where serious elicitation of prior distributions is not feasible, either because it is too difficult (and note that casual elicitation of proper priors can be very dangerous), or because an appearance of objectivity is required.

- Teaching of statistics is greatly simplified.
Other Proposed Objective Priors

James O. Berger
Duke University and the
Statistical and Applied Mathematical Sciences Institute

Branson, Missouri
June 4, 2005
A. Formal Approaches to Objective Bayes

- Objective priors for estimation
  - *Already discussed*: constant priors, Jeffreys-rule priors, reference priors, maximum entropy priors
  - *Not to be discussed*: data translated priors (Box and Tiao), maximal data information priors (Zellner), minimum description length priors (Rissanen, ...), data-dependent priors (e.g., Fraser)
  - Probability matching priors
  - Invariance priors
    * Right-Haar priors and left-Haar (Structural) priors (Fraser)
    * Fiducial distributions (Fisher) and specific invariance (Jaynes)
  - Priors on the boundary of admissibility

- Objective testing and model selection – Susie’s tutorial.

- Nonparametric priors – several talks.

- Robust Bayes analysis
Matching Priors (Peers; Datta and Mukerjee, 2004)

An objective prior is often evaluated by the frequentist coverage of its credible sets (when interpreted as confidence intervals). If $\xi$ is the parameter of interest (with $\theta$ the entire parameter), it suffices to study one-sided intervals $(-\infty, q_{1-\alpha}(x))$, where $q_{1-\alpha}(x)$ is the posterior quantile of $\xi$, defined by

$$P(\xi < q_{1-\alpha}(x) \mid x) = \int_{-\infty}^{q_{1-\alpha}(x)} \pi(\xi \mid x) d\xi = 1 - \alpha.$$  

Of interest is the frequentist coverage of the one-sided intervals

$$C(\theta) = P(q_{1-\alpha}(X) > \xi \mid \theta).$$

**Definition 1** An objective prior is exact matching for a parameter $\xi$, if it's 100(1 − $\alpha$)% one-sided posterior credible sets for $\xi$ have frequentist coverage equal to $1 - \alpha$. An objective prior is matching if this is true asymptotically up to a term of order $1/n$. 
Medical Diagnosis Example: Recall that the goal was to find confidence sets for

\[ \theta = Pr(D \mid P) = \frac{p_0p_1}{p_0p_1 + (1 - p_0)p_2}. \]

Consider the frequentist percentage of the time that the 95% Bayesian credible sets (found earlier) miss on the left and on the right (ideal would be 2.5% each) for the indicated parameter values when \( n_0 = n_1 = n_2 = 20. \)

<table>
<thead>
<tr>
<th>((p_0, p_1, p_2))</th>
<th>O-Bayes</th>
<th>Log Odds</th>
<th>Gart-Nam</th>
<th>Delta</th>
</tr>
</thead>
<tbody>
<tr>
<td>((\frac{1}{4}, \frac{3}{4}, \frac{1}{4}))</td>
<td>2.86, 2.71</td>
<td>1.53, 1.55</td>
<td>2.77, 2.57</td>
<td>2.68, 2.45</td>
</tr>
<tr>
<td>((\frac{1}{10}, \frac{9}{10}, \frac{1}{10}))</td>
<td>2.23, 2.47</td>
<td>0.17, 0.03</td>
<td>1.58, 2.14</td>
<td>0.83, 0.41</td>
</tr>
<tr>
<td>((\frac{1}{2}, \frac{9}{10}, \frac{1}{10}))</td>
<td>2.81, 2.40</td>
<td>0.04, 4.40</td>
<td>2.40, 2.12</td>
<td>1.25, 1.91</td>
</tr>
</tbody>
</table>
Invariance Priors

Generalizes ‘invariance to parameterization’ to other transformations that seem to leave a problem unchanged. There are many illustrations of this in the literature, but the most systematically studied (and most reliable) invariance theory is ‘invariance to a group operation.’

An example: location-scale group operation on a normal distribution:

- Suppose $X \sim N(\mu, \sigma)$.
- Then $X^* = aX + b \sim N(\mu^*, \sigma^*)$, where $\mu^* = a\mu + b$ and $\sigma^* = a\sigma$.

Desiderata:

- Final answers should be the same for the two problems.
- Since the $X$ and $X^*$ problems have identical ‘structure,’ $\pi(\mu, \sigma)$ should have the same form as $\pi(\mu^*, \sigma^*)$.

Mathematical consequence: use an invariant measure corresponding to the ‘group action’ of the problem, the Haar measure if unique, and the right-Haar measure otherwise (optimal from a frequentist perspective). For the example, $\pi^{RH}(\mu, \sigma) = \sigma^{-1}d\mu d\sigma$ (independence-Jefferys prior).
Example: The Bivariate Normal Model (with Dongchu Sun)

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 \(\overline{x} = (\overline{x}_1, \overline{x}_2)'\), where \(\overline{x}_i = n^{-1} \sum_{j=1}^{n} x_{ij}\), and

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

where \(s_{ij} = \sum_{k=1}^{n} (x_{ik} - \overline{x}_i)(x_{jk} - \overline{x}_j)\), \(r = s_{12}/\sqrt{s_{11}s_{22}}\).
An Extended Conjugate Class of Priors: For \(a, b > 0\), the priors for \((\mu_1, \mu_2, \sigma_1, \sigma_2, \rho)\) of the form

\[
\pi_{\text{ab}}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{\sigma_1^{3-a} \sigma_2^{2-b} (1 - \rho^2)^{2-b/2}}
\]

have essentially a closed form posterior, and include the following important special cases:

- \(\pi_{11}\), which is a one-at-a-time reference prior.
- \(\pi_{\text{RH}} = \pi_{12}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \sigma_1^{-2} (1 - \rho^2)^{-1}\), which is the right-Haar prior corresponding to invariance w.r.t. the group action \(\mu^* = \mu + a, \Sigma^* = T' \Sigma T\), with \(T\) a triangular matrix. It is not known if this is derivable as a reference prior.
- \(\pi_J = \pi_{10}\), the Jeffreys-rule prior \((|\Sigma|^{-2} d\Sigma)\).
- \(\pi_{IJ} = \pi_{21}\), the independence-Jeffreys prior \((|\Sigma|^{-3/2} d\Sigma)\).
Credible Intervals for \( \rho \), under the right-Haar prior

\[
\pi^{RH}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{\sigma_1^2(1 - \rho^2)},
\]

can be found 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 \( Y = -\frac{Z}{\sqrt{\chi^2_{n-1}}} + \frac{\sqrt{\chi^2_{n-2}} r}{\sqrt{\chi^2_{n-1}} \sqrt{1-r^2}} \);
- 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.
Lemma 1

1. This Bayesian credible set, $C(r)$, when considered as a frequentist confidence interval, has exact coverage $1 - \alpha$.

2. This credible set can be shown to be the same as the fiducial confidence interval obtained by Fisher in 1930.

Two Historical Curiosities:
1. Was it known that Fisher’s fiducial interval has exact frequentist coverage of $1 - \alpha$?
2. In the early 60’s, results of Lindley and Brillinger showed that, if one starts with the density $p(r \mid \rho)$ 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 $\pi^{RH}(\rho \mid x)$ can be shown only to depend on the data through $r$, we have a marginalization paradox (Dawid, Stone and Zidek, 1973): $\pi^{RH}(\rho \mid x) = g(\rho, r) \neq p(r \mid \rho)\pi(\rho)$ for any $\pi(\cdot)$. 
<table>
<thead>
<tr>
<th>parameter</th>
<th>exact matching constructive posteriors</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mu_1$</td>
<td>$\overline{x}_1 + \frac{Z^<em>}{\sqrt{\chi^</em><em>n}} \sqrt{\frac{s</em>{11}}{n}}$</td>
</tr>
<tr>
<td>$\mu_2$</td>
<td>$\overline{x}_2 + \frac{Z^<em>}{\sqrt{\chi^</em><em>n}} \sqrt{\frac{s</em>{22}}{n}}$</td>
</tr>
<tr>
<td>$d'(\mu_1, \mu_2), \ d \in \mathbb{R}^2$</td>
<td>$d'(\overline{x}_1, \overline{x}_2) + \frac{Z^<em>}{\sqrt{\chi^</em>_n}} \sqrt{\frac{d'Sd}{n}}$</td>
</tr>
<tr>
<td>$\sigma_1$</td>
<td>$\sqrt{s_{11}/\chi^*_n}$</td>
</tr>
<tr>
<td>$\frac{\mu_1}{\sigma_1}$</td>
<td>$\frac{Z^<em>}{\sqrt{n}} + \frac{\overline{x}<em>1}{\sqrt{s</em>{11}}} \sqrt{\chi^</em>_n}$</td>
</tr>
<tr>
<td>$\rho$</td>
<td>$\psi \left( -\frac{Z^<em>}{\sqrt{\chi^</em>_n}} + \sqrt{\frac{\chi^<em>_n}{\chi^</em>_n}} \frac{r}{\sqrt{1-r^2}} \right), \psi(x) = \frac{x}{\sqrt{1+x^2}}$</td>
</tr>
<tr>
<td>$\frac{\rho \sigma_2}{\sigma_1}$</td>
<td>$\frac{r \sqrt{s_{22}}}{\sqrt{s_{11}}} - \frac{Z^<em>}{\sqrt{\chi^</em><em>n}} \sqrt{\frac{1-r^2}{1-r^2}} \sqrt{s</em>{22}}$</td>
</tr>
<tr>
<td>$</td>
<td>\Sigma</td>
</tr>
<tr>
<td>$\sigma_2^2 (1 - \rho^2)$</td>
<td>$s_{22} (1 - r^2)/\sqrt{\chi^*_n}$</td>
</tr>
<tr>
<td>$- \frac{\rho}{\sigma_1 \sqrt{1 - \rho^2}}$</td>
<td>$\frac{Z^<em>}{\sqrt{s_{11}}} - \frac{\sqrt{\chi^</em><em>n}}{\sqrt{s</em>{11}}} \frac{r}{\sqrt{1-r^2}}$</td>
</tr>
</tbody>
</table>
The Lindley/Bayarri Reference Prior for $\rho$:

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

This does not have a closed form posterior like the $\pi_{ab}$ priors, but here is a rejection method to exactly generate from the posterior $\pi^{LB}(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho \mid X)$.

*Step 1.* Simulate

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \rho \sigma_1 \sigma_2 \\ \rho \sigma_1 \sigma_2 & \sigma_2^2 \end{pmatrix} \sim \text{Inverse Wishart}(S^{-1}, n - 1).$$

*Step 2.* Simulate $u \sim \text{unif}(0, 1)$. If $u \leq \sqrt{1 - \rho^2}$, report $\Sigma$.

Otherwise go back to *Step 1*.

*Step 3.* Simulate $\mathbf{u} = (\mu_1, \mu_2)' \sim N_2(\bar{x}, n^{-1} \Sigma)$. 

The ‘Eigenvalue’ Reference Prior (Yang and Berger, 1994)

\[
\pi^E(\mu_1, \mu_2, \sigma_1, \sigma_2, \rho) = \frac{1}{|\Sigma|(d_1 - d_2)},
\]

where \(d_1 > d_2\) are the ordered eigenvalues of \(\Sigma\). Here is a Metropolis method to exactly generate from its posterior.

**Step 1.** At each iteration, generate

\[
\Sigma^* \sim \text{Inverse Wishart}(S^{-1}, n - 1).
\]

**Step 2.** Set

\[
\Sigma' = \left\{ \begin{array}{ll}
\Sigma^* & \text{with probability } \alpha, \\
\Sigma & \text{otherwise},
\end{array} \right.
\]

where

\[
\alpha = \min \left\{ 1, \frac{(d_1^* - d_2^*)}{(d_1 - d_2)} \cdot \frac{|\Sigma|^{(k-1)/2}}{|\Sigma^*|^{(k-1)/2}} \right\}.
\]
What is a Good Overall Prior?

- The right-Haar prior appears to be great (e.g., exactly frequentist matching) for many parameters, including $\rho$, but it can be bad for inference about general features such as $\sigma_1/\sigma_2$, and it has a marginalization paradox.

- The Lindley-Bayarri prior is not exact matching for anything, but it is first order matching for many parameters (e.g. $\rho$), has reasonable numerical coverage for all parameters, and no marginalization paradox.

- The eigenvalue reference prior also seems like a good general prior, but is much harder to compute with.
Figure 2. Frequentist coverage of one-sided Bayesian credible sets for $\sigma_1/\sigma_2$. 
Multivariate Versions of $\pi^{LB}$ and $\pi^E$

A sample of $k \times 1$ vectors $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ is observed from the multivariate normal model, $N_k(\mu, \Sigma)$. Inference about the unknown $\Sigma$ can then be based on $\overline{\mathbf{x}} = n^{-1} \sum_{i=1}^n \mathbf{x}_i$ and $\mathbf{S} = \sum_{i=1}^n (\mathbf{x}_i - \overline{\mathbf{x}})'(\mathbf{x}_i - \overline{\mathbf{x}}) \sim \text{Wishart}(n, \Sigma)$. Then

$$\pi^{LB}(\mu, \Sigma) \, d\mu \, d\Sigma = \frac{1}{|\Sigma|^{k/2}} \cdot \frac{1}{|I + \Sigma \Sigma^{-1}|^{1/2}} \, d\mu \, d\Sigma,$$

where $*$ denotes the Hadamard product of matrices (this was derived as a reference prior by Chang and Eaves, 1990) and

$$\pi^E(\mu, \Sigma) \, d\mu \, d\Sigma = \frac{I_{[d_1 > \ldots > d_k]}}{|\Sigma| \prod_{i < j} (d_i - d_j)} \, d\mu \, d\Sigma,$$

where $d_1 > d_2 > \ldots > d_k$ are the eigenvalues of $\Sigma$. 
To generate from the posterior $\pi^{LB}(\mu, \Sigma \mid X)$:

**Step 1.** Generate $\Sigma \sim \text{Inverse Wishart}(S^{-1}, n - 1)$.

**Step 2.** Simulate $u \sim \text{unif}(0, 1)$. If $u \leq 2^k |I + \Sigma^* \Sigma^{-1}|^{-1/2}$, report $\Sigma$. Otherwise go back to **Step 1**.

**Step 3.** Simulate $\mu \sim N_k(\overline{x}, n^{-1} \Sigma)$.

To generate from the posterior $\pi^{E}(\mu, \Sigma \mid X)$:

**Step 1.** At each iteration, generate

$$\Sigma^* \sim \text{Inverse Wishart}(S^{-1}, n - 1).$$

**Step 2.** Set $\Sigma' = \begin{cases} 
\Sigma^* & \text{with probability } \alpha, \\
\Sigma & \text{otherwise,}
\end{cases}$

where $\alpha = \min \left\{1, \frac{\prod_{i < j} (d^*_i - d^*_j)}{\prod_{i < j} (d_i - d_j)} \cdot \frac{|\Sigma|^{(k-1)/2}}{|\Sigma^*|^{(k-1)/2}} \right\}$. 

Priors on the Boundary of Admissibility

Objective priors can be too diffuse, or can be too concentrated. Some problems this can cause:

1. **Posterior Impropriety:** If an objective prior does not yield a proper posterior for reasonable sample sizes, it is grossly defective. (One of the very considerable strengths of reference priors – and to an extent Jeffreys priors – is that they almost never result in this problem.)

2. **Inconsistency:** Inconsistent behavior can result as the sample size $n \to \infty$. For instance, in the Neyman-Scott problem of observing $X_{ij} \sim N(\mu_i, \sigma^2), i = 1, \ldots, k; j = 1, 2$, the Jeffrey-rule prior leads to an inconsistent estimator of $\sigma^2$ as $n = 2k \to \infty$; the reference prior is fine.

3. **Priors Overwhelming the Data:** As an example, in large sparse contingency tables, priors will often be much more influential than the data, if great care is not taken.
4. **Admissibility:** A more refined question, from the decision-theoretic point of view, is that of choosing priors so that, in say estimating $\theta$ by its posterior mean $\delta^\pi(x)$, under mean squared error risk

$$R(\theta, \delta^\pi) = E_X \left[ (\theta - \delta^\pi(X))^t(\theta - \delta^\pi(X)) \right],$$

$\delta^\pi$ is admissible. (The estimator $\delta$ is *inadmissible* if there exists another estimator with risk function nowhere bigger and somewhere smaller. If no such better estimator exists, $\delta$ is *admissible.*) Priors ‘on the boundary of admissibility’ are typically very appropriately balanced between being too vague and too concentrated.
Example: Good hyperpriors for normal hierarchical models (hyperpriors yielding admissible mean estimates) are given in Berger, Strawderman and Tang (2005) for the following block multivariate normal situation:

For $i = 1, 2, \ldots, m$,

- $X_i = \theta_i + \epsilon_i, \quad \epsilon_i \sim N_k(0, \Sigma_i)$, are $k \times 1$ observation vectors, $k \geq 2$, with the $\Sigma_i$ known.

- $\theta_i = z_i \beta + \xi_i, \quad \xi_i \sim N_k(0, V)$, with the $z_i$ being specified standardized $k \times l$ covariate matrices,
- $\beta$ is an $l \times 1$ unknown ‘hyper-mean’ vector,
- $V$ is an unknown $k \times k$ ‘hyper-covariance matrix’.

Goal: Find good hyperpriors $\pi(\beta, V) = \pi(\beta)\pi(V)$. 
For $\pi(\beta)$:

- Use of the constant prior $\pi(\beta) = 1$ results in inadmissibility, except when $l = 2$.

- The prior $\pi(\beta) \propto [1 + \|\beta\|^2]^{-(l-1)/2}$ is excellent from the perspective of admissibility for all $l$ (even $l = 2$).

To compute with this prior, use the equivalent representation

$$
\beta \mid \lambda \sim N_l(0, \lambda I), \quad \lambda \sim \lambda^{-1/2} e^{-1/2\lambda},
$$

- sample $\lambda$ from its full conditional, the
  Inverse Gamma($(l - 1)/2, 2/[1 + \|\beta\|^2]$) density;
- given $\lambda$ (and $V$ and the $\theta_i$), Gibbs sampling of $\beta$ can be done from its full conditional, which is

$$
N_l \left( \left( \frac{1}{\lambda} I + \sum_{i=1}^{m} z_i' V^{-1} z_i \right)^{-1} \sum_{i=1}^{m} z_i' V^{-1} \theta_i, \left( \frac{1}{\lambda} I + \sum_{i=1}^{m} z_i' V^{-1} z_i \right)^{-1} \right).
$$
For the hyper-covariance matrix $V$,

- The standard Jeffreys prior for a covariance matrix, 
  \[ \pi(V) = |V|^{-(k+1)/2}, \]
  yields an improper posterior.

- The commonly used constant prior, \( \pi(V) = 1 \), yields an extremely inadmissible estimator. Furthermore, roughly \( 2k \) observations are needed for a proper posterior, while \( k + 1 \) observations are enough for identifiability.

- Excellent admissible priors are modified reference priors; or modified Jeffreys priors of the form
  \[ \pi(V) = |\Sigma + V|^{-(k+1)/2}. \]

- Reference priors are similar but considerably more complicated.

*Note:* The recommended priors automatically result in the right type of frequentist ‘shrinkage.’
MCMC Sampling of $V$ in the Posterior for $\pi(V) = |\Sigma + V|^{-(k+1)/2}$:

Defining $W(\theta, \beta) = \sum_{i=1}^{m} (\theta_i - z_i' \beta)(\theta_i - z_i' \beta)^t$, the resulting full conditional for $V$ can be written

$$\pi(V | \theta, \beta) \propto \frac{1}{|\Sigma + V|^{(k+1)/2}|V|^{m/2}} \exp \left( -\frac{1}{2} \text{tr}(V^{-1}W(\theta, \beta)) \right),$$

One can sample from this full conditional using the following accept-reject sampling algorithm:

**Propose** a candidate $V^*$ from the InverseWishart ($W(\theta, \beta), m)$ density

**Accept** the candidate with probability $P = (|V|/|\Sigma + V|)^{(k+1)/2}$, returning to the proposal step if the candidate is rejected, and moving on to another full conditional if it is accepted.