Lecture 3

6.8 Choice of Prior and Bayesian Sensitivity Analysis

- Many non-Bayesians admit that the Bayesian development of point estimation offers an attractive theoretical package.

- They remain skeptical, however, of whether \textit{pragmatic considerations} of implementation render such an ideal unattainable, and hence, unsatisfactory as a goal for empirical researchers.
• Bayesian analysis is based on three primary inputs: a *prior*, a *likelihood*, and a *loss function*.

  ◦ Our likelihood-based frequentist analysis last quarter also required the last two inputs.

  ◦ The obvious point for critics to attack is the prior.

  ◦ But because the likelihood and loss function are given more prominent roles in Bayesian analysis than in most frequentist analyses, they also come under attack.
• Once you grant a Bayesian the prior, it is “game over” because the logic of the subsequent analysis (essentially applied probability) is compelling to many researchers.

• The place for the frequentist to circle the wagons is over the choice of prior(s).

• Section 6.7 took the prior as given. We discuss its choice later in this section.
Given a likelihood function, classical point *estimates* are often obtainable in a Bayesian framework by appropriate choice of prior distribution and loss function, possibly as limiting cases. In fact, usually there are many ways of doing it.
Rather than the Bayesian being forced to justify their choice of particular prior and loss function to produce their estimates, the Bayesian can confront the classical advocate with the need to rule out all the other Bayesian estimates that could arise from different implicit choices of prior and/or loss function used by the classical advocate.

The choice of prior and loss function required to rationalize a classical point estimate is often embarrassing to advocates of the classical approach.
In Bayesian eyes classical researchers do not avoid the choice of prior and loss function; rather classical researchers implicitly make unconscious, automatic selections without reference to the problem context.

- Making explicit what is implicit in classical analysis is hardly a shortcoming of the Bayesian approach.

- Formal introduction of subjective information is more intellectually honest than traditional ways of hiding it from the reader.
Choice of prior distribution and loss function are daunting tasks.

- This explains why many argue it is easier to do statistical or econometric theory than to do compelling empirical work.

- The Bayesian approach highlights why the latter is so elusive.
• Choice of Loss Function

  ○ Loss functions are key ingredients to both risk and posterior loss minimization. Choice of a particular loss function is a problem faced by both groups.

  ○ Although some statisticians argue against adopting a decision-making framework, their arguments are unlikely to be compelling to the economist-personality of an econometrician.

  ○ The major value of explicit loss functions is the overall structure they provide to the discussion and their role in defining a “good estimate.”
The disagreement between Bayesians and classicists is whether to minimize $E_{\theta|y}[C(\hat{\theta}, \theta)]$ or $E_{Y|\theta}[C(\overline{\theta}, \theta)]$, i.e.,

$$\theta | Y = y \quad \text{versus} \quad Y | \theta.$$  \hfill (6.8.1)

_A major contention of my text is that the choice posed in (6.8.1) underlies most statistical debates._
There appears to be no *explicit* use of prior information by classical researchers. Many classical proponents argue that nondata-based information has no role in “science.”

- I find it difficult to take these arguments seriously. How else does a researcher choose a window (likelihood function)? What is the vehicle for measuring the *value added* of the researcher?
- Empirical work is a humbling experience. Few engage in it for long before recognizing the many ways nondata-based information enters the analysis.
- Such information is not just the by-product of economic theory; indeed, the latter is often disappointingly quiet on matters of importance to the empirical researcher.

Besides precise elicitation of an individual's beliefs, priors are also chosen with reference to representing the viewpoint of a particular school of thought as to the location of the prior, e.g., sign restrictions on some key parameters.
• An elicitation method forms a bridge between an expert’s opinions and an expression of these opinions in a statistically useful form.

• Two types of prior elicitation: **structural** and **predictive**.

  - In structural elicitation, the analyst is presumed to understand the parameterization of the model and be able to state a prior directly based on subject matter considerations regarding $\theta$.

  - Following Kadane et al. (1980, *JASA*), predictive elicitation uses information (e.g., **quantiles**) for

\[
f(y) = \int_{\Theta} f(y|\theta) f(\theta) \, d\theta
\]

  to “back-out” $f(\theta)$. 
When conjugate priors are employed, prior elicitation amounts to choice of the hyper-parameters that characterize the prior family, e.g., $\alpha$ and $\delta$ (Example 6.7.1)
$\mu$ and $h$ (Example 6.7.2)
$\mu$, $q$, $s^{-2}$, and $v$ (Example 6.7.4)

- These hyperparameters can be chosen by direct introspection of beliefs concerning unknown parameters (Exercise 6.7.3). (structural elicitation)

- Often more appropriately, these hyperparameters can be chosen by indirect introspection in terms of the marginal data p.f. $f(y)$ in (6.7.2). This approach emphasizes observables in the pure subjectivist spirit of Section 5.8. (predictive elicitation)
Conjugate priors are often criticized for being motivated largely by computational considerations. *Why should prior beliefs conform to the conjugate prior form?*

- One reason is that natural conjugate priors have an interpretation in terms of a prior fictitious sample from the same process that gives rise to the likelihood function.

- This corresponds to organizing prior beliefs by viewing the observable world through the same window used for viewing the data.
○ If for the problem at hand the prior information does not conform to a conjugate prior, then by all means dispense with the conjugate prior. Little can be said in general, however, concerning elicitation in such non-conjugate cases which tend to be very problem-specific.

○ In problems with high-dimensional parameter spaces, prior elicitation tends to be both more difficult, and unfortunately, more important. Such considerations are strong motivation for use of parsimonious windows.
Although admitting the need to use non-data-based information in their research, many researchers (both Bayesians and non-Bayesians) argue that its role should in some sense be \textit{minimized}.

This has been particularly the case among \textit{objective Bayesians}, and has led to an immense literature on the development of “noninformative” (a.k.a. \textit{conventional, default, diffuse, flat, indifference, neutral, objective, reference, and vague}) priors.

“Noninformative” priors are probably the most controversial issue in Bayesian statistics with at least as much disagreement among Bayesians as between Bayesian and classical statisticians.
• The issue turns on whether it is ever possible to be in a state of “total ignorance.” As Leamer (1978, p. 61) notes regarding “ignorance”:

“Like the issue of original sin, this is a question that remains unresolved, that attracts a fair amount of theological interest, but that seems rather remote from the concerns of the man on the street.”
The origin of this question goes back to Bayes and Laplace, who suggested the *Principle of Insufficient Reason*, i.e., ignorance should be represented by a probability function that assigns equal probability to all events.

- Unfortunately, there can be no such probability function, because if mutually exclusive events $A$ and $B$ are assigned equal probability, the event $A \cup B$ is implicitly assigned twice the probability.

- Similarly, if a continuous random variable $\theta$ is assigned a uniform distribution as a reflection of ignorance, then $\gamma = \theta^{-1}$ has a nonuniform p.d.f. proportional to $\gamma^{-2}$. 
In other words, the “noninformative” prior for $\theta$ implies an “informative” prior for $\gamma$!

In a situation of “real ignorance” there is insufficient reason to select one event space rather than another, or one parameterization rather than another.

Thus, the Principle of Insufficient Reason is insufficient to determine probabilities.
One way of interpreting a state of prior ignorance is in a *relative* rather than *absolute* sense.

- A “noninformative” prior should yield a posterior distribution that reflects essentially only the sample information embodied in the likelihood function. This can happen when faced with a “large” sample, but what constitutes a “large” sample depends in part on how dogmatic is the prior.

- What is really desired is a prior that is “flat” in the region where the likelihood is non-negligible so that the latter dominates in calculation of the posterior (see Figure 6.8.1).
Figure 6.8.1

Example of a Relatively “Noninformative” Prior
In such cases the prior can be treated as essentially uniform, and the dominance of the data is known as the *Principle of Stable Estimation*.

Ideally, a prior is chosen before observing the data, and hence, the only way to guarantee that the likelihood will dominate is to choose again the prior to be flat everywhere.
• In order to overcome the reparameterization problem, Jeffreys sought a general rule to follow for choosing a prior so that the same posterior inferences were obtained regardless of the parameterization chosen.

  ◦ Jeffreys (1946, 1961) makes a general (but not dogmatic) argument in favor of choosing a “noninformative” prior proportional to the square root of the information matrix, i.e.,

  \[ f(\theta) \propto |J(\theta)|^{\frac{1}{2}}, \quad (6.8.2) \]

  where \( J(\theta) \equiv E_{Y|\theta}[- \frac{\partial^2 L(\theta)}{\partial \theta \partial \theta'}] \) is the information matrix of the sample defined in (6.5.5).
Prior (6.8.2) has the desirable feature that if the model is reparameterized by a one-to-one transformation, say $\alpha = h(\theta)$, then choosing the “noninformative” prior

$$f(\alpha) \propto \left| \mathbb{E}_{Y|\alpha} \left[ -\frac{\partial^2 L(\alpha)}{\partial \alpha \partial \alpha'} \right] \right|^{1/2}$$

(6.8.3)

will lead to the same posterior inferences as (6.8.2).

Priors adopted according to (6.8.3) or (6.8.2) are said to follow **Jeffreys’ Rule**.

Jeffreys’ prior has an established history, but it is a history plagued with ambiguities. To some the associated invariance arguments are compelling, but the motivating arguments for invariance are not as sacrosanct as they appear at first.
Example (Jeffreys’ Prior for Bernoulli Trials): For $T$ Bernoulli trials yielding $m$ success, $\mathcal{L}(\theta) \propto \theta^m (1 - \theta)^{T-m}$. The Hessian of the log-likelihood is

$$\frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial \theta^2} = -\frac{m}{\theta^2} - \frac{T-m}{(1-\theta)^2},$$

and since $E(m|\theta, T) = T\theta$, the information is

$$E \left[ -\frac{\partial^2 \ln \mathcal{L}(\theta)}{\partial \theta^2} \right] = \frac{T}{\theta(1-\theta)}.$$

Therefore, Jeffreys’ prior is $f(\theta) \propto \theta^{\frac{1}{2}} (1 - \theta)^{\frac{1}{2}}$ which is a proper U-shaped $\beta (\frac{1}{2}, \frac{1}{2})$. 
Example: Suppose $Y_t$ ($t = 1, 2, ..., T$) are i.i.d. EXP($\theta$).

(a) Derive Jeffreys’ prior for $\theta$. 
Solution: The sampling density

\[ f_{\text{EXP}}(y_t|\theta) = \theta^{-1} \exp(-\theta^{-1}y_t) \]

for \( 0 < y_t < \infty \), implies the log-likelihood

\[ L(\theta; y) = -T \ln \theta - T \bar{y} \theta^{-1} \]

and the derivatives

\[ \frac{\partial L(\theta; y)}{\partial \theta} = -\frac{T}{\theta} + \frac{T \bar{y}}{\theta^2}, \]

\[ \frac{\partial^2 L(\theta; y)}{\partial \theta^2} = \frac{T}{\theta^2} - \frac{2T \bar{y}}{\theta^3}. \]
Hence, the Fisher-information is

\[
J(\theta) = E \left[ -\frac{\partial^2 L(\theta; y)}{\partial \theta^2} \right]
\]

\[
= -\frac{T}{\theta^2} + \frac{2TE(\bar{y})}{\theta^3}
\]

\[
= \frac{T}{\theta^2}.
\]

Therefore, from (6.8.2), Jeffreys’ prior is

\[
f(\theta) \propto |J(\theta)|^{\frac{1}{2}} \propto \theta^{-1}.
\]
(b) Derive Jeffreys’ prior for $\alpha = \theta^{-1}$. 
Solution: Similarly,

\[ L(\alpha; y) = T \ln \alpha - T \bar{y} \alpha \]

with derivatives

\[ \frac{\partial L(\alpha; y)}{\partial \alpha} = \frac{T}{\alpha} - T \bar{y}, \]

\[ \frac{\partial^2 L(\alpha; y)}{\partial \alpha^2} = -\frac{T}{\alpha^2}. \]

Note that observed information is constant in this case. Therefore, from (6.8.2), Jeffreys’ prior is

\[ f(\alpha) \propto |J(\alpha)|^{\frac{1}{2}} \propto \alpha^{-1}. \]
(c) Find the posterior density of $\theta$ corresponding to the prior density in (a). Be specific in noting the family to which it belongs.
Solution:

$$f(\theta | y) \propto \theta^{-1} \cdot \theta^{-T} \exp\left(-T \bar{y} \theta^{-1}\right)$$

$$\propto \theta^{-(T+1)} \exp\left(-T \bar{y} \theta^{-1}\right),$$

which is recognized from Table 3.3.1 (page 111) to be

$$\text{IG}(T, [T\bar{y}]^{-1}).$$
(d) Find the posterior density of $\alpha$ corresponding to the prior density in (b). Be specific in noting the family to which it belongs.
Solution: The posterior density of $\alpha$ corresponding to the prior in (b) is

$$p(\alpha|y) \propto \alpha^{-1} \cdot \alpha^T \exp(-T \bar{y} \alpha)$$

$$\propto \alpha^{T-1} \exp(-T \bar{y} \alpha),$$

which is recognized as $G(T, [T \bar{y}]^{-1})$.

- The invariance property of Jeffreys’ prior refers to the fact that the densities in (c) and (d) are related by a simple change-of-variable.
The claim to “noninformativeness” for Jeffreys’ prior rests on various arguments using Shannon’s information criterion as a measure of distance between densities.

- There is a fair amount of agreement that Jeffreys’ priors may be reasonable in one-parameter problems, but substantially less agreement (including Jeffreys) in multiple parameter problems.

Jeffreys’ prior is usually (but not always - recall Example 6.8.1) improper, i.e.,

\[ \int_{\mathbb{R}^K} f(\theta) \, d\theta \to \infty. \] (6.8.4)

Such impropriety can endow a posterior computed in the usual fashion using Bayes Theorem with a variety of disturbing properties (e.g., in some cases the posterior may also be improper).
Since Jeffreys’ prior is proportional to the square root of the information matrix, it depends on the expected information in the sample.

- This dependency on a sampling theory expectation makes it sensitive to a host of problems related to the LP.

- If applied to different sampling experiments involving proportional likelihoods (e.g., binomial versus negative binomial as in Example 6.2.4), then different noninformative priors are suggested, and hence, different posteriors arise.
Suppose $y_t$ corresponds to an asset price in a market subject to “circuit breakers” that activate when cumulative price changes over some period exceed predetermined limits.

The *ex ante* sampling distribution will change to take into account the potential censoring, and so (in general) Jeffreys’ prior is different than when no potential censoring is present.
Whether a market breaker is ever activated in the data at hand is *not* relevant.

- Rather the mere *possibility* of censoring, as opposed to its occurrence, is relevant.

- It is not hard to imagine that in almost any asset market, there is some finite change in price over the course of the day that would trigger market closure, if not catastrophe.

- Suppose prior p.d.f. (6.7.5) is uniform, i.e., \( \alpha = \delta = 1 \), or in terms of the equivalent prior sample information interpretation, \( \alpha - 1 = 0 \), successes in a sample of \( T = \alpha + \delta - 2 = 0 \). Combined with a sample of \( m \) successes in \( T \) trials, this prior implies a posterior mean (6.7.52) equaling \( (m + 1)/(T + 2) \) and a posterior mode (6.7.54) equaling the ML estimate \( \bar{y} = m/T \).
Villegas (1977) recommends $\alpha = \delta = 0$ in which case prior p.d.f. (6.7.4) is improper and equals

$$f(\theta) \propto \theta^{-1} (1 - \theta)^{-1}. \quad (6.8.5)$$

(6.8.5) is U-shaped (recall Figure 3.3.4) with a minimum at $\theta = \frac{1}{2}$, and it approaches infinity as $\theta \to 0$ and as $\theta \to 1$.

- The corresponding posterior p.d.f. is a *proper* beta p.d.f. (provided $0 < m < T$) with $\bar{\alpha} = m$ and $\bar{\delta} = T - m$.

- Since the posterior mean (6.7.52) equals the ML estimate $\bar{\theta} = m/T$, some might argue that prior (6.8.5) is “noninformative.”
• Box and Tiao (1973) advocate a *proper* beta prior with \( \alpha = \delta = \frac{1}{2} \).

  - This yields a proper beta posterior distribution for \( \theta \) with \( \bar{\alpha} = m + \frac{1}{2} \) and \( \bar{\delta} = T - m + \frac{1}{2} \).

  - The posterior mean is \( \frac{m + \frac{1}{2}}{T + 1} \) and the posterior mode is \( \frac{m - \frac{1}{2}}{T - 1} \).

  - The rationale for this prior being “noninformative” is that the likelihood is approximately *data translated* in \( \alpha = \sin^{-1}(\theta^{\frac{1}{2}}) \). This prior also satisfies **Jeffreys’ Rule**.
• Zellner (1977) took an information-theoretic approach which led to the proper prior:

\[ f(\theta) = 1.61856 \theta^\theta (1 - \theta)^{1 - \theta}, \quad 0 \leq \theta \leq 1. \]
Example 6.8.2: In the spirit of Example 6.7.6, if we use an “equivalent prior sample size” interpretation of a noninformative prior for the normal sampling problem (with known variance) in Example 6.7.2, then putting $T = 0$ in (6.7.40) implies $h^{-1} = \infty$ and an improper prior distribution $f(\theta) \propto \text{constant}, -\infty < \theta < \infty$. In the limit, however, posterior distribution (6.7.22) is proper:

$$\theta | y \sim \mathcal{N}(\bar{y}, h^{-1}), \quad h = \frac{T}{\sigma^2}. \quad (6.8.6)$$
\[ \theta | y \sim N(\bar{y}, h^{-1}), \quad h = \frac{T}{\sigma^2}. \]  

(6.8.6)

- Although (6.8.6) implies the Bayesian point estimate of $\hat{\theta} = \bar{y}$ under quadratic, symmetric-linear, and all-or-nothing loss functions, the interpretation of (6.8.6) is Bayesian.

- According to (6.8.6), $\theta$ is a normal random variable with fixed mean $\bar{y}$ and variance $h^{-1}$, whereas the standard classical view is that $\bar{y}$ is a normal random variable with fixed mean $\theta$ and variance $h^{-1}$.

- Let $Z \equiv h^{1/2}(\theta - \bar{y})$. According to Bayesian posterior analysis: $Z|y \sim N(0, 1)$. According to standard classical analysis: $Z|\theta \sim N(0, 1)$. While the standard normal distribution $N(0, 1)$ is involved in both cases, $Z$ is random for different reasons under the two paradigms.
Example 6.8.3: Reconsider Example 6.7.4, but with the normal-gamma prior distribution replaced by the “noninformative” prior

\[ f(\theta_1, \theta_2) \propto \theta_2^{-1}. \quad (6.8.7) \]

- Prior density (6.8.7) can be rationalized by:
  - arguing prior beliefs between \( \theta_1 \) and \( \theta_2 \) are independent, and
  - applying Jeffreys’ Rule separately to \( \theta_1 \) and \( \theta_2 \).
• Using (6.7.18) and prior (6.8.7) implies the posterior density

\[ f(\theta_1, \theta_2 | y) \propto [\theta_2]^{(T-3)/2} \exp(-\frac{1}{2} \theta_2 \nu s^2) \varphi(\theta_1 | \bar{y}, [T\theta_2]^{-1}) \]  \hspace{1cm} (6.8.8a)

\[ \propto [\theta_2]^{(T-1)/2} \exp(-\frac{1}{2} \theta_2 \nu s^2 + T(\theta_1 - \bar{y})^2), \] \hspace{1cm} (6.8.8b)

which is recognized [using (6.7.23)] as the kernel of the normal-gamma distribution NG(\(\bar{y}, T^{-1}, s^2, T-1\)).

• The marginal posterior distribution of \(\theta_1\) corresponding to posterior density (6.8.8) is t(\(\bar{y}, s^2/T, T-1\)).
\( \theta_1 \) is a random variable with a t-distribution with fixed mean \( \bar{y} \), variance \( h^{-1} \) and T-1 degrees of freedom, whereas the standard classical view would be that \( \bar{Y} \) is a normal random variable with a t-distribution with fixed mean \( \theta_1 \), variance \( h^{-1} \), and \( \nu = T - 1 \) degrees of freedom.

- Let \( \tau \equiv T^{1/2} (\theta_1 - \bar{y})/s \).
  - According to **Bayesian posterior analysis**
    \[ \tau | y \sim t(0, 1, \nu). \]
  - According to **classical analysis**
    \[ \tau | \theta \sim t(0, 1, \nu). \]
  - Although the Student t-distribution \( t(0, 1, \nu) \) with \( \nu \) degrees of freedom is involved in both cases, \( \tau \) is random for different reasons under the two paradigms.
Given $\mu$ and $\sigma^2$, suppose $Y_t \ (t = 1, 2, ..., T)$ are i.i.d. $N(\mu, \sigma^2)$. Then the information matrix for the sample is

$$J_T(\mu, \sigma^2) = \mathbb{E}_{Y|\mu, \sigma^2} \begin{bmatrix} \frac{T}{\sigma^2} & \frac{T(\bar{Y} - \mu)}{\sigma^4} \\ \frac{T(\bar{Y} - \mu)}{\sigma^4} & -\frac{T}{2\sigma^4} + \frac{1}{\sigma^6} \sum_{t=1}^{T} (Y_t - \theta_t)^2 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{T}{\sigma^2} & 0 \\ 0 & \frac{T}{2\sigma^4} \end{bmatrix},$$

and so Jeffreys’ Rule implies the prior $f(\mu, \sigma^2) \propto \sigma^{-3}$. 
In this case Jeffreys argued against his general rule and advocated treating $\mu$ and $\sigma^2$ independently and using $f(\mu, \sigma^2) \propto \sigma^{-2}$.

- In the case of known $\sigma^2$, the Jeffreys’ prior for $\mu$ is $p(\mu) \propto$ constant.

- In the case of known $\mu$, the Jeffreys’ prior for $\sigma^2$ is $p(\sigma^2) \propto \sigma^{-2}$. 
• Kass and Wasserman (1996, *JASA*) argue:

  ○ Jeffreys evolved toward seeing priors as *chosen by convention* (e.g., like weights and measures), rather than as unique representations of ignorance.
As just noted, not all of Jeffreys’ recommendations always followed Jeffreys’ Rule: \( f(\theta) \propto |J(\theta)|^{\frac{1}{2}} \).

- When \( \Theta \) is finite, Jeffreys assigned equal probabilities to each of the values.

- When \( \Theta \) is a bounded interval, Jeffreys assumed a constant proper prior.

- When \( \Theta = \mathbb{R} \), Jeffreys assumed a constant improper prior.
When $\Theta = [0, \infty)$, Jeffreys chose $f(\theta) = 1/\theta$ because it is invariant under power transformations.

When $\theta = [\theta_1, \theta_2]'$ where $\theta_1$ is a location parameter and $\theta_2$ is a non-location parameter, Jeffreys chose $f(\theta) \propto |J(\theta)|^{1/2}$, where $J(\theta)$ is calculated holding $\theta_1$ fixed (recall Example 6.8.3).

In the case of mixture models, Jeffreys argued that the mixing parameters should be treated independently from the other parameters.
Some researchers view a prior as being “noninformative” if it yields a proper posterior which for a “reasonable” loss function yields a point estimate equal to the MLE (e.g., see Examples 6.8.1-6.8.3).

- A flat prior and an “all-or-nothing” loss function imply that the Bayesian point estimate, i.e., posterior mode, is also the MLE.

- Remember how all of you declined to defend the all-or-nothing loss function?
The motivation for such an interpretation of “noninformative” is that the classical ML point estimate is usually viewed as reflecting only “objective” sample information.

A tongue-in-cheek Bayesian who held this view might further argue that a noninformative prior should reflect “ignorance,” and what could be more ignorant than a prior that produces the point estimate of a maximum likelihood?
• Table 6.8.1 (like Table 6.7.1) summarizes the “noninformative” results for sampling from a univariate normal distribution with a “noninformative” prior.

  ○ Case 1: the mean is unknown and the precision is known (Example 6.8.2).

  ○ Case 2: the mean is known and the precision is unknown (not studied).

  ○ Case 3 both the mean and precision are unknown (Example 6.8.3).
Table 6.8.1: “Noninformative” Analysis:

\[ Y_t (t = 1, 2, ..., T) | \mu, \sigma^{-2} \sim \text{i.i.d. } N(\mu, \sigma^2) \]

<table>
<thead>
<tr>
<th>Prior</th>
<th>Hyperparameter Updating</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f(\mu) \propto c )</td>
<td>( f(\mu</td>
<td>y) = \phi(\mu</td>
</tr>
<tr>
<td>( f(\sigma^{-2}) \propto \sigma^2 )</td>
<td>( \bar{v} \bar{s}^{-2} = vs^2 + T(\bar{y} - \mu)^2 )</td>
<td>( f(\sigma^{-2}</td>
</tr>
<tr>
<td>( f(\mu, \sigma^{-2}) \propto \sigma^2 )</td>
<td>( \bar{v} * \bar{s}^2 = \bar{v} s^2 + \bar{1} (\mu - \bar{\mu})^2 )</td>
<td>Case 2: ( \mu ) known, ( \sigma^{-2} ) unknown</td>
</tr>
<tr>
<td>( \bar{v} * \bar{s}^2 = \bar{v} \bar{s}^{-2} + \bar{1} (\mu - \bar{\mu})^2 )</td>
<td>( \bar{v} * = \bar{v} + 1 )</td>
<td>( f(\mu, \sigma^{-2}</td>
</tr>
<tr>
<td>( \bar{v} * = \bar{v} + 1 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Case 3: \( \mu \) unknown, \( \sigma^{-2} \) unknown
Table 6.8.2 contains the analogous cases for a multivariate normal distribution in which the “noninformative” Jeffreys’ prior

\[ f(\Sigma^{-1}) \propto |\Sigma|^{(M+1)/2} \]

is used.
### Table 6.8.2: “Noninformative” Analysis:

$$Y_t \ (t = 1, 2, ..., T) \mid \mu, \Sigma^{-1} \sim \text{i.i.d. } N_M(\mu, \Sigma)$$

<table>
<thead>
<tr>
<th>Prior</th>
<th>Hyperparameter Updating</th>
<th>Posterior</th>
</tr>
</thead>
<tbody>
<tr>
<td>Case 1: $\mu$ unknown, $\Sigma^{-1}$ known</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f(\mu) \propto \text{constant}$</td>
<td>$f(\mu \mid y) = \phi_M(\mu \mid \bar{y}, T^{-1} \Sigma)$</td>
<td></td>
</tr>
<tr>
<td>$f(\Sigma^{-1}) \propto</td>
<td>\Sigma^{-1}</td>
<td>^{-(M+1)/2}$</td>
</tr>
<tr>
<td>Case 2: $\mu$ known, $\Sigma^{-1}$ unknown</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f(\Sigma^{-1}) \propto</td>
<td>\Sigma^{-1}</td>
<td>^{-(M+1)/2}$</td>
</tr>
<tr>
<td>Case 3: $\mu$ unknown, $\Sigma^{-1}$ unknown</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$f(\mu, \Sigma^{-1}) \propto</td>
<td>\Sigma^{-1}</td>
<td>^{-(M+1)/2}$</td>
</tr>
<tr>
<td>$\bar{S}_* = \nu S + \bar{T}(\mu - \bar{\mu})(\mu - \bar{\mu})'$</td>
<td></td>
<td></td>
</tr>
<tr>
<td>$\nu_* = \nu + 1$</td>
<td>$\phi(\mu \mid \bar{y}, \bar{T}^{-1} \Sigma)W_M(\Sigma^{-1} \mid \bar{S}^{-1}, \omega) =$</td>
<td></td>
</tr>
<tr>
<td></td>
<td>$f(\mu \mid y, \bar{S}/\bar{T}, \omega)W_M(\Sigma^{-1} \mid \bar{S}<em>*^{-1}, \omega</em>*)$</td>
<td></td>
</tr>
</tbody>
</table>
• Kass and Wasserman (1996, *JASA*) survey formal rules for choosing priors that let the *data speak for themselves*.

  ◦ Alfred Marshall once remarked: “the most reckless and dangerous theorist is the man who claims to let the facts speak for themselves.”

  ◦ These priors are intended to lead to proper posteriors dominated by the data, and to serve as *benchmarks* for posteriors derived from subjective considerations.

  ◦ Many of these priors were initially motivated on *simplicity* grounds. But in the process of refinement such priors became more complicated, even more so than eliciting a subjective prior.
• Bernardo (1979, JRSSB) suggested a method for constructing *reference priors* offering two innovations.

  ○ He defined missing information in terms of the *Kullbach-Leibler discrepancy* between the posterior and the prior density.

  ○ He developed a stepwise procedure for handling *nuisance parameters*. If there are no nuisance parameters, then Bernardo’s method usually leads to Jeffreys’ Rule.
Subsequently numerous refinements have been made in joint work with James O. Berger.

- A cottage industry sprung up for generating reference priors in a wide variety of situations.


**Comment:** Some are bothered that the *objective* prior depends on the *subjective* choice of which are parameters of interest and which are nuisance parameters.

  - Some researchers interpret improper priors from the standpoint of finitely (as opposed to countably) additive probability functions. The resulting nonconglomerability (recall footnote 5 on pages 13-14) bothers some researchers.

  - In high-dimensional problems, the effects of the prior can be subtle: it may have little posterior influence on some functions of the data and have an overwhelming influence on other functions.
Improper priors are not guaranteed to lead to admissible Bayesian point estimates.

“Marginalization paradoxes” can occur - see Dawid, Stone and Zidek (1973, JRSSB).

Improper posteriors can arise (e.g., when some parameters are unidentified).
• Digression on Finite Additivity

**Example:** Consider a random variable $Y$ whose possible values are the positive integers $y = 1, 2, 3, \ldots$. The assertion that all such integers are “equally likely,” i.e., $f(y_t) \propto 1$, is meaningless under countable additivity. The assertion $f(y_t) \propto 1$ implies there exist a number $p$ such that $P(Y = k) = p$ for all $k = 1, 2, 3, 4, \ldots$.

- **Countable additivity** is not compatible with the first axiom of probability if $p > 0$ or if $p = 0$.
- But $p = 0$ is compatible with **finite additivity**. In this case although the probability of $Y$ taking on values in a finite set is zero, we have non-zero probabilities for $Y$ falling in infinite sets, e.g., $P(Y$ is even) $= .5$. 
- Unfortunately, there are problems with finitely additive priors.

  - If probability is countably infinite, the distribution of \( Y \) is completely specified by \( P(Y = k) \) for all \( k \), i.e., the probability of any proposition concerning \( Y \) can be derived from these basic probabilities.

  - This is not the case for the finitely additive case, because probabilities of infinite sets are not determined by this information alone.
**Example:** Let $Y$ be the number of successes in $T$ independent trials with a probability $\theta$ of success on each trial. Suppose *both* $\theta$ and $T$ are unknown, and that both have independent uniform prior distributions, i.e.,

$$f(\theta, T) \propto c, \quad 0 \leq \theta \leq 1, \text{ and } T = 1, 2, 3, \ldots$$

where $c$ is a constant. Then the posterior density is

$$f(\theta, T|y) \propto \binom{T}{y} \theta^y (1 - \theta)^{T-y},$$

for $0 \leq \theta \leq 1$ and $T = y, y+1, y+2, \ldots$. Both the prior and posterior are *improper*, and the latter does not go away as additional observations are added.
O’Hagan (1995, p. 79) notes that both posterior marginals are also \textit{improper}, but that both posterior conditionals are \textit{proper}.

Specifically:

\[
f(T|y) \propto \frac{1}{T+1}, \quad T = y, y+1, y+2, ...
\]

\[
f(\theta|y) \propto \frac{1}{\theta}, \quad 0 \leq \theta \leq 1,
\]

\[
f(T|y, \theta) \propto \binom{T}{y}(1-\theta)^{T-y}, \quad T = y, y+1, y+2, ...
\]

\[
f(\theta|y, T) \propto \theta^y(1-\theta)^{T-y}, \quad 0 \leq \theta \leq 1.
\]
Example 6.8.4 [Geweke (1986, JAE)]: Consider a random sample $Y_t \ (t = 1, 2, ..., T)$ from a $N(\theta, T)$ population, where it is known with certainty that $\theta \geq 0$. It is straightforward to show that the MLE of $\theta$ is

$$
\hat{\theta}_{ML} = \begin{cases} 
\bar{y}, & \text{if } \bar{y} \geq 0 \\
0, & \text{if } \bar{y} < 0 
\end{cases}.
$$

(6.8.9)

Consider the “noninformative prior” over $\Re^+$:

$$
f(\theta) = \begin{cases} 
1, & \text{if } \theta \geq 0 \\
0, & \text{if } \theta < 0 
\end{cases}.
$$

(6.8.10)
Then the posterior distribution of $\theta$ is a truncated (on the left at zero) normal distribution (recall Exercise 2.2.5):

$$f(\theta|y) = \begin{cases} 
\frac{\phi(\theta|\bar{y}, 1)}{1 - \Phi(\bar{y})}, & \text{if } \theta \geq 0 \\
0, & \text{if } \theta < 0 
\end{cases}.$$  

(6.8.11)
Under quadratic loss the Bayesian point estimate of $\theta$ is the posterior mean given by [recall Exercise 3.3.2(a)]:

$$E(\theta|y) = \bar{y} + \left[ \frac{\phi(\bar{y})}{1 - \Phi(\bar{y})} \right].$$  \hspace{1cm} (6.8.12)

Estimates (6.8.9) and (6.8.12) differ appreciably as shown in Table 6.8.3. It is left to the reader to judge whether (6.8.9) or (6.8.12) is more reasonable.
Table 6.8.3
Maximum Likelihood Estimate and Posterior Mean
Based on a Random Sample from a
Truncated Standard Normal Distribution

| \( \bar{y} \) | \( \hat{\theta}_{ML} \) | \( E(\theta|y) \) |
|------------|-----------------|-----------------|
| -4         | 0               | .22561          |
| -3         | 0               | .28310          |
| -2         | 0               | .37322          |
| -1         | 0               | .52514          |
| 0          | 0               | .79788          |
| 1          | 1               | 1.28760         |
| 2          | 2               | 2.05525         |
| 3          | 3               | 3.00444         |
| 4          | 4               | 4.00013         |
Summary: Attempts to employ “noninformative” priors raise problems in Bayesian point estimation. Caution is in order. For whom such a state of affairs should be embarrassing is another question.

- Just because priors may be difficult to specify does not mean one should pick a “noninformative” prior.

- There are many candidates for “noninformative” priors. One problem is that there are too many candidates! Recall Example 6.8.1.

- Another problem is that “noninformative” priors often have properties that seem rather non-Bayesian (often improper).
Most “noninformative” priors depend on some or all of the following:

- the likelihood,
- the sample size,
- an expectation with respect to the sampling distribution,
- the parameters of interest, and
- whether the researcher is engaging in estimation, testing or predicting.

Dependency on the form of the likelihood is common (e.g., conjugate priors - informative or otherwise), but some of the other dependencies are less compelling and can have disturbing consequences.
To an objectivist who envisions $\theta$ as having meaning on its own, it is disturbing that a “noninformative” prior for $\theta$ requires so much information regarding the experiment or purpose of the research.

Classical statisticians like to question Bayesians about the interpretation of improper priors.

Some Bayesians argue that improper priors are merely a mathematical convenience for producing a proper posterior which reflects only the sample information embodied in the likelihood.
• Other Bayesians refuse to consider noninformative priors arguing that in a theory of learning one cannot learn anything starting from a state of “total ignorance.” Indeed, under total ignorance the sign of a point estimate cannot be “surprising.”

  ◦ Such Bayesians further argue that we are never in a state of total ignorance in the first place.

  ◦ It is ironic to employ so much information in specifying the likelihood, and then go to the other extreme to claim “no information” about the parameters that characterize the likelihood.
• My advice: *use a “noninformative” prior only with great care, and never alone.*

- I include “noninformative” priors in the class of priors over which I perform my *sensitivity analysis*.

- *Public research* involving only a single prior is likely to draw few readers.
I try to do what I say:


• There have been many criticisms of the “over-precision” demanded by probability elicitors such as de Finetti or Savage.

  ◦ Many Bayesians adhere to the Savage/de Finetti idealism, but temper it with pragmatism in implementation that is commonly referred to as Bayesian sensitivity analysis.

  ◦ This course follows this pragmatic approach and the Pragmatic Principles of Model Building offered in Section 10.4.

  ◦ Anybody who takes seriously choosing a loss function or prior in a specific problem finds the matter difficult - evidence again of why compelling empirical work is so daunting a task.
An obvious approach to mitigating this problem is to try out different prior distributions and loss functions and see if they materially matter, i.e., conduct a Bayesian sensitivity analysis.

- Here the focus is on sensitivity analysis with respect to the prior.

- Since the window is defined as being an entity about which there is *intersubjective agreement* (i.e., the window is expanded until intersubjective agreement is reached), sensitivity analysis with respect to the likelihood is postponed until Chapter 10.
Academic research is a public exercise. Even a “Savage-like” researcher with an infinite amount of time who is able to elicit his/her own coherent prior distribution, must realize that readers are likely to want to see analysis incorporating other priors as well.

- Sensitivity analyses may be local or global, depending on how far away the researcher is willing to wander from a maintained prior.

- Sensitivity analyses can also be performed in reference to all statistical activities: design, estimation, testing and prediction. Often it addresses more than one of these activities [e.g., Ramsay and Novick (1980, *JASA*)].
Conceptually, sensitivity analysis of the prior in an estimation context proceeds as follows. Given the likelihood \( \mathcal{L}(\theta; y) \), consider a family of prior p.f. \( \mathcal{F} = \{f(\theta|\xi), \xi \in \Xi\} \) defined parametrically by a set \( \Xi \), and a loss function \( C(\theta, \theta) \).

- Entertaining various professional positions in terms of \( \theta \) can lead to different choices of \( \xi \).
Given a prior $f(\theta | \xi), \xi \in \Xi$, the optimal Bayesian point estimate is

$$
\hat{\theta}(\xi) = \arg\min_{\hat{\theta} \in \Theta} \int_{\Theta} C(\hat{\theta}, \theta) \left[ \frac{f(\theta | \xi) L(\theta; y)}{f(y | \xi)} \right] d\theta,
$$  \hspace{1cm} (6.8.13)

where the marginal density of the data is

$$
f(y | \xi) = \int_{\Theta} f(y | \theta) f(\theta | \xi) d\theta.
$$  \hspace{1cm} (6.8.14)

Bayesian sensitivity analysis is simply a study of the variability of $\hat{\theta}_\Xi = \{ \theta(\xi), \xi \in \Xi \}$. 
Often a natural metric in which to measure such variability is expected posterior loss:

\[
\min_{\xi \in \Xi} c(\xi) \leq c(\xi) \leq \max_{\xi \in \Xi} c(\xi), \tag{6.8.15}
\]

where the loss yielded by \(\hat{\theta}(\xi)\) is

\[
c(\xi) \equiv \int_{\Theta} C[\hat{\theta}(\xi), \theta] \left[ \frac{f(\theta|\xi) L(\theta; y)}{f(y|\xi)} \right] d\theta. \tag{6.8.16}
\]
If the variability in $\hat{\theta}_\Xi$ or the range (6.8.15) is judged acceptable for the purposes at hand, then further interrogation of prior beliefs is not required and a point estimate $\Theta(\xi)$ may be chosen without much concern, say by choosing $\xi \in \Xi$ based on ease of interpretation or computational considerations.

If the variability in $\hat{\theta}_\Xi$ or the range (6.8.15) is judged unacceptable for the purposes at hand, then further interrogation of beliefs is required in the hope of better articulating the prior family $\mathcal{F}$. 
If a prior p.f. $f(\xi)$ can be articulated, then the problem reduces to a standard case in which a single known prior

$$f(\theta) = \int_{\Xi} f(\theta|\xi) f(\xi) \, d\xi$$  \hspace{1cm} (6.8.17)

can be used. Such situations are known as **Bayesian hierarchical analysis** [see the seminal work of Lindley and Smith (1972) and Exercise 6.8.6].
In most practical problems, however, there is no agreed upon \( f(\xi) \), and we are left with investigating the sensitivity of the analyses across elements in \( \mathcal{F} \).

Sometimes a quantity of interest like a posterior mean \( \mathbb{E}(\theta|y) \) can be analytically restricted to a fairly small set of possible values for any given \( \xi \in \Xi \). The extreme bounds analysis (EBA) of Ed Leamer is a leading example and will be discussed when we deal with regression.
Empirical Bayes methods use the data via

\[ f(y|\xi) = \int_{\Theta} f(y|\theta) f(\theta|\xi) \, d\theta \]

to “estimate” \( \xi \). Type II maximum likelihood treats \( f(y|\xi) \) as an ordinary likelihood function.
Example 6.8.5: Consider Example 6.8.2 involving a random sample of size $T$ from a normal distribution with unknown mean $\theta_1$ and \textit{known} precision $\theta_2$.

- Suppose simple quadratic loss is appropriate, and that prior (6.7.10) can be only partially articulated in the sense that prior beliefs are centered over $\mu$, but the strength of these priors beliefs are unclear.

- This suggests a prior family

$$\mathcal{F} = \{\varphi(\theta_1 \mid \mu, \xi^{-1}), \xi \in \Xi\},$$

where $\Xi \equiv \{\xi \mid 0 \leq \xi < \infty\}$. 

- Given $\xi \in \Xi$, Theorem 6.7.1(a) implies that the Bayesian point estimate of $\theta_1$ is posterior mean (6.7.13) with $\xi$ replacing $h$, i.e.,

$$\hat{\theta}_1(\xi) = \frac{\xi \mu + h \bar{y}}{\xi + h},$$

(6.8.18)

and $h = T \theta_2$. The set $\hat{\theta}_{1,\Xi}$ of possible posterior means is

$$\hat{\Theta}_{1,\Xi} = (\mu, \bar{y}), \quad \text{if } \mu \leq \bar{y},$$

$$\hat{\Theta}_{1,\Xi} = [\bar{y}, \mu), \quad \text{if } \mu > \bar{y}.$$
However, if the prior location $\mu$ is also uncertain, then the posterior mean can be any real number.

If the prior elicitation in Example 6.8.5 is not possible or $\mathcal{F}$ is not reduced by some other means, then for the data at hand a point estimate of $\theta$ cannot be agreed upon. Sometimes the conclusion that the data at hand have not “confessed,” and that there remains an unfortunate level of ambiguity, is a valuable message to convey to consumers of research.
- The facility of a sensitivity analysis is problem specific, and the presentation of such results often relies on artistic merits.

- Conveying the posterior robustness of a point estimate is often easy when analytical results are available and when the class of priors $\mathcal{F}$ is easy to envision, and challenging otherwise.

- There remains much to be learned in this art - the obvious avenue for such learning is increased activity in serious Bayesian empirical studies (not just textbook illustrations) as initiated in Poirier (1991b, *JoE*).