Chapter 3. Construction of Priors
(continued)

♦ Using the Marginal Distribution to
  Determine the Prior

• Marginal Distribution

If $X$ has a probability density $f(x|\theta)$, and $\theta$ has
density $\pi(\theta)$, then the joint density of $X$ and $\theta$ is

$$h(x, \theta) = f(x|\theta)\pi(\theta).$$

The marginal density of $X$ is

$$m(x|\pi) = \int_{\Theta} f(x|\theta) dF_{\pi}(\theta)$$

$$= \begin{cases} \int_{\Theta} f(x|\theta)\pi(\theta) d\theta & \text{(continuous case)}, \\ \sum_{\Theta} f(x|\theta)\pi(\theta) & \text{(discrete case)}. \end{cases}$$

• Example 15: If $X$ (given $\theta$) is $N(\theta, \sigma_f^2)$ and $\pi(\theta)$ is
  a $N(\mu_\pi, \sigma_\pi^2)$ density, then a standard probability
  calculation shows that $m(x|\pi)$ is a $N(\mu_\pi, \sigma_\pi^2 + \sigma_f^2)$
density.
• **Information about** $m$

The sources of information about $m$: subjective knowledge and/or data

For example, suppose $\theta = (\theta_1, \theta_2, \ldots, \theta_p)'$ and the $\theta_i$ are i.i.d. from the density $\pi_0$. Suppose also that the data $X = (X_1, \ldots, X_p)'$, where each $X_i$ has density $f(x_i | \theta_i)$. Then the common marginal distribution of each $X_i$ is

$$m_0(x_i) = \int f(x_i | \theta_i) dF^{\pi_0}(\theta_i),$$

and $X_1, \ldots, X_p$ can be considered to be a simple random sample from $m_0$. Note that

$$m(x) = \int f(x | \theta) \pi(\theta) d\theta$$

$$= \int \left[ \prod_{i=1}^{p} f(x_i | \theta_i) \right] \left[ \prod_{i=1}^{p} \pi_0(\theta_i) \right] d\theta$$

$$= \prod_{i=1}^{p} \int f(x_i | \theta_i) \pi_0(\theta_i) d\theta_i = \prod_{i=1}^{p} m_0(x_i).$$

Thus the data $x$ can be used to estimate $m_0$. This type of situation is typically an *empirical Bayes* or *compound decision* problem (names due to Robbins (1951,1955, 1964)).
• Restrictive Classes of Priors

I. Priors of a Given Functional Form

This class of priors is of the form

$$\Gamma = \{\pi : \pi(\theta) = g(\theta|\lambda), \lambda \in \Lambda\}.$$ 

Here $g$ is a prescribed function, so that choice of a prior reduces to the choice of $\lambda \in \Lambda$. The parameter $\lambda$ (often a vector) is called a hyperparameter of the prior, particularly in situations where it is considered unknown and to be determined from information about the marginal distribution.

• Example 16: Suppose $\theta$ is a normal mean. It is felt that the prior distribution $\pi$, for $\theta$, can be adequately described by the class of normal distributions, and, in addition, it is certain that the prior mean is positive. Then

$$\Gamma = \{\pi : \pi \text{ is } N(\mu_\pi, \sigma_\pi^2), \mu_\pi > 0, \sigma_\pi^2 > 0\},$$

so that $\lambda = (\mu_\pi, \sigma_\pi^2)'$ is the hyperparameter.
II. Priors of a Given Structural Form

Consider $\mathbf{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$. The class of priors is of the form

$$\Gamma = \{ \pi : \pi(\mathbf{\theta}) = \prod_{i=1}^{p} \pi_0(\theta_i), \pi_0 \text{ is an arbitrary density}\}.$$

**Example 17:** Suppose the $X_i$ are independently $N(\theta_i, \sigma_f^2)$ ($\sigma_f^2$ known) and that the $\theta_i$ are likewise felt to be independent with a common $N(\mu_\pi, \sigma_\pi^2)$ prior distribution (call it $\pi_0$), the hyperparameters $\mu_\pi$ and $\sigma_\pi^2$ being completely unknown. Then

$$\Gamma = \{ \pi : \pi(\mathbf{\theta}) = \prod_{i=1}^{p} \pi_0(\theta_i), \pi_0 \text{ being } N(\mu_\pi, \sigma_\pi^2),$$

$$-\infty < \mu_\pi < \infty, \sigma_\pi^2 > 0 \}.$$
III. Priors Close to an Elicited Prior

A rich and calculation ally attractive class to work with is the $\epsilon$-contamination class

$$\Gamma = \{\pi : \pi(\theta) = (1 - \epsilon)\pi_0(\theta) + \epsilon q(\theta), \ q \in Q\}$$

where $0 < \epsilon < 1$ reflects how “close” we feel that $\pi$ must be to $\pi_0$, and $Q$ is a class of possible “contaminations.”

- **Example 2 (continued):** The description of Example 2 can be found on page 4-5. The elicitation process yielded, as one reasonable possibility for $\pi_0$, the $N(0, 2.19)$. Suppose that distributions which have probabilities differing from $\pi_0$ by as much as (say) 0.2 would be plausible priors. Then we could choose $\epsilon = 0.2$. We will defer discussion of the choice of $Q$ in Chapter 4.
The ML-II Approach to Prior Selection

Definition

Suppose $\Gamma$ is a class of priors under consideration, and that $\hat{\pi} \in \Gamma$ satisfies (for the observed data $\boldsymbol{x}$)

$$m(\boldsymbol{x}|\hat{\pi}) = \sup_{\pi \in \Gamma} m(\boldsymbol{x}|\pi).$$

Then $\hat{\pi}$ will be called the Type II maximum likelihood prior, or ML-II prior for short.

For instance, when $\Gamma$ is the class

$$\Gamma = \{\pi : \pi(\theta) = g(\theta|\lambda), \lambda \in \Lambda\},$$

then

$$\sup_{\pi \in \Gamma} m(\boldsymbol{x}|\pi) = \sup_{\lambda \in \Lambda} m(\boldsymbol{x}|g(\theta|\lambda)),$$

so that one simply has to perform a maximization over the hyperparameter $\lambda$. We will call the maximizing hyperparameters the ML-II hyperparameters.
• Example 17 (continued): From Example 15, we have

\[
m(\mathbf{x}|\pi) = \prod_{i=1}^{p} m_0(x_i|\pi_0),
\]

where \(m_0\) is \(N(\mu_\pi, \sigma^2_\pi + \sigma^2_f)\). Thus, we can write

\[
m(\mathbf{x}|\pi) = \prod_{i=1}^{p} \frac{1}{[2\pi(\sigma^2_\pi + \sigma^2_f)]^{1/2}} \exp \left\{ - \frac{(x_i - \mu_\pi)^2}{2(\sigma^2_\pi + \sigma^2_f)} \right\}
\]

\[
= [2\pi(\sigma^2_\pi + \sigma^2_f)]^{-p/2} \exp \left\{ - \frac{\sum_{i=1}^{p} (x_i - \mu_\pi)^2}{2(\sigma^2_\pi + \sigma^2_f)} \right\}
\]

\[
= [2\pi(\sigma^2_\pi + \sigma^2_f)]^{-p/2} \exp \left\{ -\frac{ps^2}{2(\sigma^2_\pi + \sigma^2_f)} \right\}
\]

\[
\times \exp \left\{ \frac{-p(\bar{x} - \mu_\pi)^2}{2(\sigma^2_\pi + \sigma^2_f)} \right\},
\]

where \(\bar{x} = \frac{1}{p} \sum_{i=1}^{p} x_i\) and \(s^2 = \frac{1}{p} \sum_{i=1}^{p} (x_i - \bar{x})^2\).

We seek to maximize \(m(\mathbf{x}|\pi)\) over hyperparameters \(\mu_\pi\) and \(\sigma^2_\pi\). It is easy to see that the maximum over \(\mu_\pi\) is attained at \(\bar{x}\), regardless of the value of \(\sigma^2_\pi\), so that \(\hat{\mu}_\pi = \bar{x}\) is the ML-II choice of \(\mu_\pi\).
Inserting this value into the expression for \( m(x|\pi) \), it remains only to maximize

\[
\psi(\sigma^2_\pi) = [2\pi(\sigma^2_\pi + \sigma^2_f)]^{-p/2} \exp \left\{ \frac{-ps^2}{2(\sigma^2_\pi + \sigma^2_f)} \right\}
\]

over \( \sigma^2_\pi \). Now

\[
\frac{d}{d\sigma^2_\pi} \log \psi(\sigma^2_\pi) = \frac{-p/2}{\sigma^2_\pi + \sigma^2_f} + \frac{ps^2}{2(\sigma^2_\pi + \sigma^2_f)^2}.
\]

This equals to zero at \( \sigma^2_\pi = s^2 - \sigma^2_f \), provided that \( s^2 \geq \sigma^2_f \). If \( s^2 < \sigma^2_f \), the derivative is always negative, so that the maximum is achieved at \( \sigma^2_\pi = 0 \). Thus, we have that the ML-II estimate of \( \sigma^2_\pi \) is

\[
(s^2 - \sigma^2_f)^+ = \max\{0, s^2 - \sigma^2_f\}.
\]

In conclusion, the ML-II prior, \( \hat{\pi}_0 \), is

\[
N(\hat{\mu}_\pi, \hat{\sigma}^2_\pi) = N(\bar{x}, \max\{0, s^2 - \sigma^2_f\}).
\]
**Example 18:** For any $\pi$ in the $\epsilon$-contamination class

$$
\Gamma = \{ \pi : \pi(\theta) = (1 - \epsilon)\pi_0(\theta) + \epsilon q(\theta), \, q \in \mathcal{Q} \},
$$

it is clear that

$$
m(x|\pi) = \int_\Theta f(x|\theta)[(1 - \epsilon)\pi_0(\theta) + \epsilon q(\theta)]d\theta
= (1 - \epsilon)m(x|\pi_0) + \epsilon m(x|q).
$$

Thus, the ML-II prior can be found by maximizing $m(x|q)$ over $q \in \mathcal{Q}$, and thus using the maximizing $\hat{q}$ in the expression for $\pi$.

If $\mathcal{Q}$ is the class of all possible distributions, then

$$
m(x|q) = \int_\Theta f(x|\theta)q(\theta)d\theta \leq f(x|\hat{\theta}),
$$

where $\hat{\theta}$ maximizes $f(x|\theta)$ (i.e., $\hat{\theta}$ is a maximum likelihood estimate (MLE) of $\theta$). It is easy to see that the maximum value for $m(x|q)$ is achieved by taking $q$ to be concentrated at $\hat{\theta}$. Thus, we have that the ML-II prior $\hat{\pi}$ is

$$
\hat{\pi} = (1 - \epsilon)\pi_0(\theta) + \epsilon<\hat{\theta}>.
$$

Note that if $\pi_0$ is a continuous density, $\hat{\pi}$ is a mixture of a continuous and a discrete probability distribution.
• The Moment Approach to Prior Selection

The moment approach applies when $\Gamma$ is of the “given functional form” type and it is possible to relate prior moments to moments of the marginal distribution, the latter being supposedly either estimated from data or determined subjectively.

• Lemma 1: Let $\mu_f(\theta)$ and $\sigma_f^2(\theta)$ denote the conditional mean and variance of $X$ (i.e., the mean and variance with respect to the density $f(x|\theta)$). Let $\mu_m$ and $\sigma_m^2$ denote the marginal mean and variance of $X$ (with respect to $m(x)$). Assuming these quantities exist, then

$$
\mu_m = E^\pi[\mu_f(\theta)],
$$

$$
\sigma_m^2 = E^\pi[\sigma_f^2(\theta)] + E^\pi[(\mu_f(\theta) - \mu_m)^2].
$$
Proof:

\[ \mu_m = E^m[X] = \int_X x m(x) dx = \int_X x \int_{\Theta} f(x|\theta) \pi(\theta) d\theta dx \]

\[ = \int_{\Theta} \pi(\theta) \int_X x f(x|\theta) dx d\theta \]

\[ = \int_{\Theta} \pi(\theta) \mu_f(\theta) d\theta = E^\pi[\mu_f(\theta)]. \]

Similarly,

\[ \sigma^2_m = E^m[(X - \mu_m)^2] = E^\pi \left\{ E^\theta_f[(X - \mu_m)^2|\theta] \right\} \]

\[ = E^\pi \left\{ E^\theta_f[(X - \mu_f(\theta) + \mu_f(\theta) - \mu_m)^2|\theta] \right\} \]

\[ = E^\pi \left\{ E^\theta_f[(X - \mu_f(\theta))^2] + (\mu_f(\theta) - \mu_m)^2 \right\} \]

\[ = E^\pi \left[ \sigma^2_f(\theta) \right] + E^\pi \left[ (\mu_f(\theta) - \mu_m)^2 \right]. \]
• Corollary 1:

(i) If $\mu_f(\theta) = \theta$, then $\mu_m = \mu_\pi$, where $\mu_\pi = E^\pi[\theta]$ is
the prior mean.

(ii) If, in addition, $\sigma^2_f(\theta) = \sigma^2_f$, then $\sigma^2_m = \sigma^2_f + \sigma^2_\pi$,
where $\sigma^2_\pi$ is the prior variance.

• Example 19: Suppose $X \sim N(\theta, 1)$, and that the
class, $\gamma$, of all $N(\mu_\pi, \sigma^2_\pi)$ priors for $\theta$ is considered
reasonable. Subjective experience yields a
“prediction” that $X$ will be about 1, with associated
“prediction variance” of 3. Thus we estimate that
$\mu_m = 1$ and $\sigma^2_m = 3$. Using Corollary 1, noting that
$\sigma^2_f = 1$, we have that $1 = \mu_m = \mu_f$ and
$3 = \sigma^2_m = 1 + \sigma^2_\pi$. Solving for $\mu_\pi$ and $\sigma^2_\pi$, we conclude
that the $N(1, 2)$ prior should be used.
\textbf{Example 17 (continued):} We again seek to determine $\mu_\pi$ and $\sigma^2_\pi$. Treating $X_1, X_2, \ldots, X_p$ as a sample from $m_0$, the standard method of moments estimates for $\mu_{m_0}$ and $\sigma^2_{m_0}$ are $\bar{x}$ and $s^2 = \frac{1}{p} \sum_{i=1}^{p} (x_i - \bar{x})^2$. Note that the moment estimate for the second marginal moment $\mu_{2,m_0}$ is $\frac{1}{p} \sum_{i=1}^{p} x_i^2$. Thus, the moment estimate for $\sigma^2_{m_0}$ is

$$\frac{1}{p} \sum_{i=1}^{p} x_i^2 - \bar{x}^2 = s^2.$$

It follows that the moment estimates of $\mu_\pi$ and $\sigma^2_\pi$ are $\hat{\mu}_\pi = \bar{x}$ and $\hat{\sigma}^2_\pi = s^2 - \sigma^2_f$. Note that $\hat{\sigma}^2_\pi$ could be negative, a recurring problem with moment estimates.
• The Distance Approach to Prior Selection

We directly estimate $m$ and then use

$$m(x) = \int_{\Theta} f(x|\theta) dF^\pi (\theta).$$

to determine $\pi$.

If a large amount of data $x_1, x_2, \ldots, x_p$ is available, we use the density estimate method to estimate $m(x)$ by

$$\hat{m}(x) = \frac{1}{p} \text{[the number of } x_i \text{ equal to } x].$$

The difficult encountered in using an estimate, $m$, is that the equation

$$\hat{m}(x) = \int_{\Theta} f(x|\theta) dF^\pi (\theta)$$

need have no solution, $\pi$. Hence all we can seek is an estimate of $\pi$, say, $\hat{\pi}$, for which

$$\hat{m}_{\hat{\pi}}(x) = \int_{\Theta} f(x|\theta) dF^{\hat{\pi}} (\theta)$$

is close (in some sense) to $\hat{m}(x)$.
A reasonable measure of “distance” between two such densities is

\[ d(\hat{m}, m_\hat{\pi}) = E^\hat{m} \left[ \log \frac{\hat{m}(X)}{m_\hat{\pi}(X)} \right] \]

\[ = \left\{ \begin{array}{ll}
\int_{X} \hat{m}(x) \left[ \log \frac{\hat{m}(x)}{m_\hat{\pi}(x)} \right] dx & \text{(continuous case)} \\
\sum_{X} \hat{m}(x) \left[ \log \frac{\hat{m}(x)}{m_\hat{\pi}(x)} \right] & \text{(discrete case)}
\end{array} \right. \]

\[ = E^\hat{m} [\log \hat{m}(X)] - E^\hat{m} [\log m_\hat{\pi}(X)]. \]

Since only the last term of this expression depends on \( \hat{\pi} \), it is clear that minimizing \( d(\hat{m}, m_\hat{\pi}) \) over \( \hat{\pi} \) is equivalent to maximizing

\[ E^\hat{m} [\log m_\hat{\pi}(X)]. \]

Finding the maximizer \( \hat{\pi} \) is difficult. However, when \( \Theta = \{\theta_1, \ldots, \theta_k\} \), letting \( p_i = \hat{\pi}(\theta_i) \), we have

\[ \hat{m}_\hat{\pi}(x) = \sum_{i=1}^{k} f(x|\theta_i)p_i. \]

Hence, finding the optimal \( \hat{\pi} \) reduces to the problem of maximizing

\[ E^\hat{m} \left[ \log \left( \sum_{i=1}^{k} f(x|\theta_i)p_i \right) \right] \]

over all \( p_i \) such that \( 0 \leq p_i \leq 1 \) and \( \sum_{i=1}^{k} p_i = 1 \). For
the density estimate of $m(x)$, the above expression becomes

$$E^n \left[ \log \left( \sum_{i=1}^{k} f(x|\theta_i)p_i \right) \right]$$

$$= \sum_{j=1}^{p} \frac{1}{p} \log \left( \sum_{i=1}^{k} f(x|\theta_i)p_i \right).$$

The maximization of this last quantity over the $p_i$ is a linear programming problem.
Hierarchical Priors

There are often two more stages. The hierarchical approach is most commonly used when the first stage, $\Gamma$, consists of priors of a certain functional form. Thus, if

$$\Gamma = \{\pi_1(\theta|\lambda) : \quad \pi_1 \text{ is of a given functional form & } \lambda \in \Lambda\},$$

then the second stage would consist of putting a prior distribution, $\pi_2(\lambda)$, on the hyperparameter $\lambda$. Such a second stage prior is sometimes called a hyperprior.

- **Example 17 (continued)**: The structural assumption of independence of the $\theta_i$, together with the assumption that they have a common normal distribution, led to (where $\lambda = (\mu_\pi, \sigma^2_\pi)'$)

$$\Gamma = \left\{ \pi_1 : \quad \pi_1(\theta) = \prod_{i=1}^{p} \pi_0(\theta_i), \pi_0 \text{ being } N(\mu_\pi, \sigma^2_\pi), \quad -\infty < \mu_\pi < \infty, \sigma^2_\pi > 0 \right\}.$$  

A second stage prior, $\pi_2(\lambda)$ could be chosen for the hyperparameters according to subjective beliefs.
For instance, in the example where the $X_i$ are test scores measuring the “true abilities” $\theta_i$, one could interpret $\mu_\pi$ and $\sigma_\pi^2$ as the population mean and variance of the $\theta_i$. Suppose that the “mean true ability’ $\mu_\pi$ is near 100, with a “standard error” of $\pm 20$, while the “variance of true abilities”, $\sigma_\pi^2$, is about 200, with a “standard error” of $\pm 100$. A reasonable prior for $\mu_\pi$ would then be $N(100, 400)$, while $\mathcal{IG}(6, 0.001)$ distribution might be a reasonable prior for $\sigma_\pi^2$. Furthermore, it is reasonable to assume the prior independence of $\mu_\pi$ and $\sigma_\pi^2$. Thus, the second stage prior for $\lambda$ is the product of the $N(100, 400)$ density times the $\mathcal{IG}(6, 0.001)$ density.