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 θ has density $\pi(\theta)$, then the joint density of X and θ is

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

The *marginal density* of X is

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

• **Example 15**: If X (given θ) 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 $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)'$ and the θ_i are i.i.d. from the density π_0. Suppose also that the data $\boldsymbol{X} = (X_1, \ldots, X_p)'$, 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

$$\begin{aligned} m(\boldsymbol{x}) &= \int f(\boldsymbol{x}|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}) d\boldsymbol{\theta} \\ &= \int \left[\prod_{i=1}^{p} f(x_i|\theta_i) \right] \left[\prod_{i=1}^{p} \pi_0(\theta_i) \right] d\boldsymbol{\theta} \\ &= \prod_{i=1}^{p} \int f(x_i|\theta_i) \pi_0(\theta_i) d\theta_i = \prod_{i=1}^{p} m_0(x_i). \end{aligned}$$

Thus the data $\boldsymbol{x}$ 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**

<u>I. Priors of a Given Functional Form</u>

This class of priors is of the form

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

Here g is a prescribed function, so that choice of a prior reduces to the choice of $\boldsymbol{\lambda} \in \Lambda$. The parameter $\boldsymbol{\lambda}$ (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 θ is a normal mean. It is felt that the prior distribution π, for θ, 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 $\boldsymbol{\lambda} = (\mu_\pi, \sigma_\pi^2)'$ is the hyperparameter.

II. Priors of a Given Structural Form

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

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

• **Example 17**: Suppose the X_i are independently $N(\theta_i, \sigma_f^2)$ (σ_f^2 known) and that the θ_i are likewise felt to be independent with a common $N(\mu_\pi, \sigma_\pi^2)$ prior distribution (call it π_0), the hyperparameters μ_π and σ_π^2 being completely unknown. Then

$$\Gamma = \Big\{\pi : \ \ \pi(\boldsymbol{\theta}) = \prod_{i=1}^{p} \pi_0(\theta_i), \pi_0 \text{ being } N(\mu_\pi, \sigma_\pi^2),$$
$$- \infty < \mu_\pi < \infty, \sigma_\pi^2 > 0\Big\}.$$

III. Priors Close to an Elicited Prior

A rich and calculation ally attractive class to work with is the ϵ-*contamination class*

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

where $0 < \epsilon < 1$ reflects how "close" we feel that π must be to π_0, and $\mathcal{Q}$ 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 π_0, the $N(0, 2.19)$. Suppose that distributions which have probabilities differing from π_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 $\mathcal{Q}$ in Chapter 4.

- **The ML-II Approach to Prior Selection**

<u>**Definition**</u>

Suppose Γ 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 Γ is the class

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

then

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

so that one simply has to perform a maximization over the hyperparameter $\boldsymbol{\lambda}$. We will call the maximizing hyperparameters the ML-II *hyperparameters*.

• **Example 17 (continued)**: From Example 15, we have

$$m(\boldsymbol{x}|\pi) = \prod_{i=1}^{p} m_0(x_i|\pi_0),$$

where m_0 is $N(\mu_\pi, \sigma_\pi^2 + \sigma_f^2)$. Thus, we can write

$$\begin{aligned} m(\boldsymbol{x}|\pi) = & \prod_{i=1}^{p} \frac{1}{[2\pi(\sigma_\pi^2 + \sigma_f^2)]^{1/2}} \exp\left\{ - \frac{(x_i - \mu_\pi)^2}{2(\sigma_\pi^2 + \sigma_f^2)} \right\} \\ = & [2\pi(\sigma_\pi^2 + \sigma_f^2)]^{-p/2} \exp\left\{ - \frac{\sum_{i=1}^{p}(x_i - \mu_\pi)^2}{2(\sigma_\pi^2 + \sigma_f^2)} \right\} \\ = & [2\pi(\sigma_\pi^2 + \sigma_f^2)]^{-p/2} \exp\left\{ \frac{-ps^2}{2(\sigma_\pi^2 + \sigma_f^2)} \right\} \\ & \times \exp\left\{ \frac{-p(\bar{x} - \mu_\pi)^2}{2(\sigma_\pi^2 + \sigma_f^2)} \right\}, \end{aligned}$$

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 $m(\boldsymbol{x}|\pi)$ over hyperparameters μ_π and σ_π^2. It is easy to see that the maximum over μ_π is attained at $\bar{x}$, regardless of the value of σ_π^2, so that $\hat{\mu}_\pi = \bar{x}$ is the ML-II choice of μ_π.

Inserting this value into the expression for $m(\boldsymbol{x}|\pi)$, it remains only to maximize

$$\psi(\sigma_\pi^2) = [2\pi(\sigma_\pi^2 + \sigma_f^2)]^{-p/2} \exp\left\{\frac{-ps^2}{2(\sigma_\pi^2 + \sigma_f^2)}\right\}$$

over σ_π^2. Now

$$\frac{d}{d\sigma_\pi^2} \log \psi(\sigma_\pi^2) = \frac{-p/2}{(\sigma_\pi^2 + \sigma_f^2)} + \frac{ps^2}{2(\sigma_\pi^2 + \sigma_f^2)^2}.$$

This equals to zero at $\sigma_\pi^2 = s^2 - \sigma_f^2$, provided that $s^2 \geq \sigma_f^2$. If $s^2 < \sigma_f^2$, the derivative is always negative, so that the maximum is achieved at $\sigma_\pi^2 = 0$. Thus, we have that the ML-II estimate of σ_π^2 is

$$(s^2 - \sigma_f^2)^+ = \max\{0, s^2 - \sigma_f^2\}.$$

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

$$N(\hat{\mu}_\pi, \hat{\sigma}_\pi^2) = N(\bar{x}, \max\{0, s^2 - \sigma_f^2\}).$$

• **Example 18**: For any π in the ϵ-contamination class

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

it is clear that

$$\begin{aligned} m(\boldsymbol{x}|\pi) = & \int_{\Theta} f(\boldsymbol{x}|\theta)[(1-\epsilon)\pi_0(\theta) + \epsilon q(\theta)]d\theta \\ = & (1-\epsilon)m(\boldsymbol{x}|\pi_0) + \epsilon m(\boldsymbol{x}|q). \end{aligned}$$

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

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

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

where $\hat{\theta}$ maximizes $f(\boldsymbol{x}|\theta)$ (i.e., $\hat{\theta}$ is a maximum likelihood estimate (MLE) of θ). It is easy to see that the maximum value for $m(\boldsymbol{x}|q)$ 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 π_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 Γ 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 μ_m and σ_m^2 denote the marginal mean and variance of X (with respect to $m(x)$). Assuming these quantities exist, then*

$$\begin{aligned}\mu_m =& E^\pi[\mu_f(\theta)],\\ \sigma_m^2 =& E^\pi[\sigma_f^2(\theta)] + E^\pi[(\mu_f(\theta) - \mu_m)^2].\end{aligned}$$

Proof:

$$
\begin{aligned}
\mu_m =& E^m[X] = \int_{\mathcal{X}} x m(x) dx = \int_{\mathcal{X}} x \int_{\Theta} f(x|\theta)\pi(\theta) d\theta dx \\
=& \int_{\Theta} \pi(\theta) \int_{\mathcal{X}} x f(x|\theta) dx d\theta \\
=& \int_{\Theta} \pi(\theta)\mu_f(\theta) d\theta = E^\pi[\mu_f(\theta)].
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\sigma_m^2 =& E^m[(X-\mu_m)^2] = E^\pi\{E_\theta^f[(X-\mu_m)^2|\theta]\} \\
=& E^\pi\{E_\theta^f[(X-\mu_f(\theta)+\mu_f(\theta)-\mu_m)^2|\theta]\} \\
=& E^\pi\{E_\theta^f[(X-\mu_f(\theta))^2] + (\mu_f(\theta)-\mu_m)^2\} \\
=& E^\pi[\sigma_f^2(\theta)] + E^\pi[(\mu_f(\theta)-\mu_m)^2].
\end{aligned}
$$

• **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, $\sigma_f^2(\theta) = \sigma_f^2$, then $\sigma_m^2 = \sigma_f^2 + \sigma_\pi^2$, where σ_π^2 is the prior variance.

• **Example 19**: Suppose $X \sim N(\theta, 1)$, and that the class, γ, of all $N(\mu_\pi, \sigma_\pi^2)$ priors for θ 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 $\sigma_m^2 = 3$. Using Corollary 1, noting that $\sigma_f^2 = 1$, we have that $1 = \mu_m = \mu_f$ and $3 = \sigma_m^2 = 1 + \sigma_\pi^2$. Solving for μ_π and σ_π^2, we conclude that the $N(1, 2)$ prior should be used.

• **Example 17 (continued)**: We again seek to determine μ_π and σ_π^2. Treating X_1, X_2, ..., X_p as a sample from m_0, the standard method of moments estimates for μ_{m_0} and $\sigma_{m_0}^2$ 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 μ_{2,m_0} is $\frac{1}{p}\sum_{i=1}^p x_i^2$. Thus, the moment estimate for $\sigma_{m_0}^2$ is

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

It follows that the moment estimates of μ_π and σ_π^2 are $\hat{\mu}_\pi = \bar{x}$ and $\hat{\sigma}_\pi^2 = s^2 - \sigma_f^2$. Note that $\hat{\sigma}_\pi^2$ 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 π.

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, $\hat{m}$, is that the equation

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

need have no solution, π. Hence all we can seek is an estimate of π, 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

$$
\begin{aligned}
d(\hat{m}, m_{\hat{\pi}}) &= E^{\hat{m}}\left[\log \frac{\hat{m}(X)}{\hat{m}_{\hat{\pi}}(X)}\right] \\
&= \begin{cases} \int_{\mathcal{X}} \hat{m}(x)\left[\log \frac{\hat{m}(x)}{\hat{m}_{\hat{\pi}}(x)}\right] dx & \text{(continuous case)} \\ \sum_{\mathcal{X}} \hat{m}(x)\left[\log \frac{\hat{m}(x)}{\hat{m}_{\hat{\pi}}(x)}\right] & \text{(discrete case)} \end{cases} \\
&= E^{\hat{m}}[\log \hat{m}(X)] - E^{\hat{m}}[\log \hat{m}_{\hat{\pi}}(X)].
\end{aligned}
$$

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

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

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

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

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, Γ, consists of priors of a certain functional form. Thus, if

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

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

• **Example 17 (continued)**: The structural assumption of independence of the θ_i, together with the assumption that they have a common normal distribution, led to (where $\boldsymbol{\lambda} = (\mu_\pi, \sigma_\pi^2)'$)

$$\Gamma = \Big\{\pi_1 : \ \ \pi_1(\boldsymbol{\theta}) = \prod_{i=1}^{p} \pi_0(\theta_i), \pi_0 \text{ being } N(\mu_\pi, \sigma_\pi^2),$$
$$-\infty < \mu_\pi < \infty, \sigma_\pi^2 > 0\Big\}.$$

A second stage prior, $\pi_2(\boldsymbol{\lambda})$ 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" θ_i, one could interpret μ_π and σ_π^2 as the population mean and variance of the θ_i. Suppose that the "mean true ability' μ_π is near 100, with a "standard error" of ± 20, while the "variance of true abilities", σ_π^2, is about 200, with a "standard error" of ± 100. A reasonable prior for μ_π would then be $N(100, 400)$, while $\mathcal{IG}(6, 0.001)$ distribution might be a reasonable prior for σ_π^2. Furthermore, it is reasonable to assume the prior independence of μ_π and σ_π^2. Thus, the second stage prior for $\boldsymbol{\lambda}$ is the product of the $N(100, 400)$ density times the $\mathcal{IG}(6, 0.001)$ density.