Chapter 3. Construction of Priors (continued)

♠ Informative Priors

• Rationale for Informative Priors

Prior elicitation perhaps plays the most crucial role in Bayesian inference. Although, noninformative and improper priors may be useful and easier to specify for certain problems, they cannot be used in all applications, such as model selection or model comparison, as it is well known that proper priors are required to compute Bayes factors and posterior model probabilities. It is well known that Bayes factors are generally quite sensitive to the choices of hyperparameters of vague proper priors, and thus one cannot simply specify vague proper priors in model selection contexts to avoid informative prior elicitation. In addition, noninformative priors can cause instability in the posterior estimates and convergence problems for the Gibbs sampler. This can occur if the posterior surface is flat when using

noninformative or improper priors. Moreover, noninformative priors do not make use of real prior information that one may have for a specific application. Thus, informative priors are essential in these situations, and in general, they are useful in applied research settings where the investigator has access to previous studies measuring the same response and covariates as the current study. For example, in many cancer and AIDS clinical trials, current studies often use treatments that are very similar or slight modifications of treatments used in previous studies. We refer to data arising from previous similar studies as *historical data*. In carcinogenicity studies, for example, large historical databases exist for the control animals from previous experiments. In all of these situations, it is natural to incorporate the historical data into the current study by quantifying it with a suitable prior distribution on the model parameters. The informative prior elicitation discussed here can be applied to each of these situations as well as in other applications that involve historical data.

- **Power Prior Distributions**
  - One particular class of informative prior distributions is the class of *Power Priors*, which are constructed from historical data.
  - The power prior construction is based on the notion of the existence of a previous similar study that measures the same response variable and covariates as the current study.
  - Let D_0 denote the historical data.
  - Let $L(\boldsymbol{\theta}|D_0)$ denote the likelihood function of $\boldsymbol{\theta}$ based on the historical data. Here, $\boldsymbol{\theta}$ is a generic label for the vector of parameters of the model.

– The power prior is defined as

$$\pi(\boldsymbol{\theta}|D_0, a_0) \propto L(\boldsymbol{\theta}|D_0)^{a_0}\pi_0(\boldsymbol{\theta}).$$

– $\pi_0(\boldsymbol{\theta})$ is the *initial prior* for $\boldsymbol{\theta}$. That is, $\pi_0(\boldsymbol{\theta})$ is the prior for $\boldsymbol{\theta}$ before the historical data D_0 is observed.

– $0 \leq a_0 \leq 1$ is a scalar precision parameter.

- a_0 controls the heaviness of the tails of the prior. The smaller the a_0, the heavier the tails.
- $a_0 = 0$ corresponds to no incorporation of historical data. $a_0 = 1$ corresponds to the Bayesian update of $\pi_0(\boldsymbol{\theta})$.

• **Example 1: Normal Linear Regression Model**

Consider the normal linear regression model

$$\boldsymbol{y}_0|X_0, \boldsymbol{\beta} \sim N_{n_0}(X_0\boldsymbol{\beta}, I),$$

where X_0 is an $n \times p$ design matrix and $\boldsymbol{\beta}$ is a p-dimensional vector of regression coefficients. Suppose that we take a noninformative initial prior $\pi_0(\boldsymbol{\beta}) = 1$. Let $D_0 = (\boldsymbol{y}_0, X_0)$ denote the historical data. Then, the likelihood function based on the data D_0 is

$$L(\boldsymbol{\beta}|D_0) \propto \exp\left\{y_0' X_0\boldsymbol{\beta} - \frac{1}{2}\,\boldsymbol{\beta}' X_0' X_0\boldsymbol{\beta}\right\},$$

and the power prior is given by

$$\begin{aligned}
&\pi(\boldsymbol{\beta}|D_0, a_0) \propto [L(\boldsymbol{\beta}|D_0)]^{a_0}\,\pi_0(\boldsymbol{\beta}) \\
&\propto \exp\left\{a_0\left[y_0' X_0\boldsymbol{\beta} - \frac{1}{2}\,\boldsymbol{\beta}' X_0' X_0\boldsymbol{\beta}\right]\right\} \\
&\propto \exp\left\{-\frac{a_0}{2}\,(\boldsymbol{\beta} - \boldsymbol{\mu}_0)'\,(X_0' X_0)\,(\boldsymbol{\beta} - \boldsymbol{\mu}_0)\right\},
\end{aligned}$$

where $\boldsymbol{\mu}_0 = (X_0' X_0)^{-1} X_0' \boldsymbol{y}_0$.

Thus

$$\boldsymbol{\beta} \mid D_0, a_0 \sim N_p(\boldsymbol{\mu}_0, a_0^{-1}\ (X_0'X_0)^{-1}).$$

In this example, we can see the precise role of a_0. That is, a_0 is a precision parameter that quantifies the degree of prior belief in $\boldsymbol{\mu}_0$, and hence D_0.

• **Example 2: Logistic Regression Model**

To further illustrate the role of a_0 in the power priors, we consider the following logistic regression model. We simulated a data set consisting $n_0 = 200$ independent Bernoulli observations with success probability

$$p_{0i} = \frac{\exp\{-0.5 + 0.5x_{0i}\}}{1 + \exp\{-0.5 + 0.5x_{0i}\}}\ , \qquad i = 1, 2, \ldots, n_0\ ,$$

where the x_{0i} are *i.i.d.* normal random variables with mean 0 and standard deviation 0.5.

Let y_{0i} denote the binary response for the i^{th} observation and $\boldsymbol{x}_{0i} = (1, x_{0i})'$ for $i = 1, 2, \ldots, n_0$. Again, the historical data $D_0 = (\boldsymbol{y}_0, X_0)$, where $\boldsymbol{y}_0 = (y_{01}, y_{02}, \ldots, y_{0n_0})'$ and X_0 is a $n_0 \times 2$ matrix with the i^{th} row $\boldsymbol{x}_{0i}'$. Also Let $\boldsymbol{\beta} = (\beta_0, \beta_1)'$. Then, the likelihood function is given by

$$L(\boldsymbol{\beta}|D_0) = \prod_{i=1}^{n_0} \frac{\exp\{y_{0i}\boldsymbol{x}_{0i}'\boldsymbol{\beta}\}}{1 + \exp\{\boldsymbol{x}_{0i}'\boldsymbol{\beta}\}},$$

and the power prior with an improper uniform initial prior is thus given by

$$\pi(\boldsymbol{\beta}|D_0, a_0) \propto \prod_{i=1}^{n_0} \frac{\exp\{a_0 y_{0i}\boldsymbol{x}_{0i}'\boldsymbol{\beta}\}}{(1 + \exp\{\boldsymbol{x}_{0i}'\boldsymbol{\beta}\})^{a_0}}.$$

Figure 1 shows the contours of the power prior for various a_0 values.

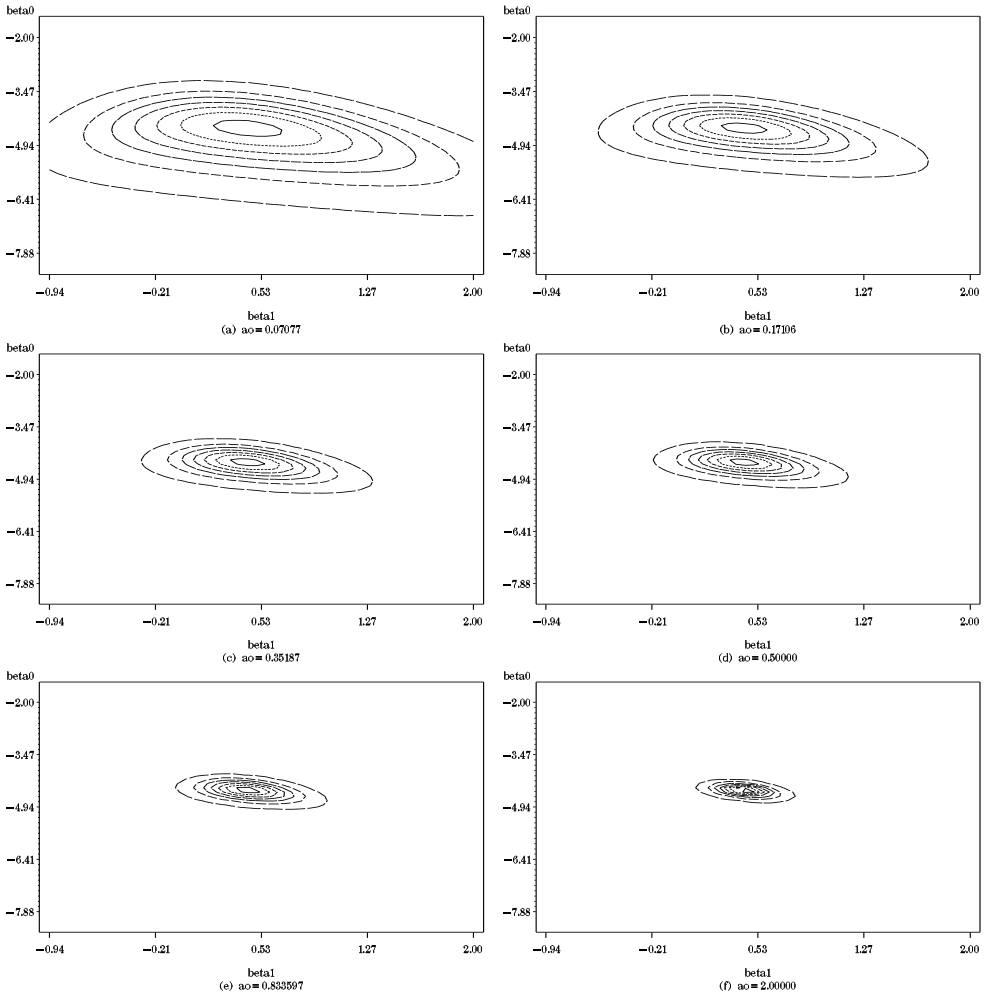

FIGURE 1. Contours of the Power Prior for $a_0 = 0.07$, 0.17, 0.35, 0.5, 0.83

From Figure 1, we can see that

(i) the centers of the power priors remain the same for different a_0 values; and

(ii) the tails of the power priors become heavier and the prior surfaces are getting flatter, as a_0 becomes smaller.

Thus, it becomes even clearer that a_0 serves as a dispersion parameter, which controls the heaviness of the tails of the prior.

From the practical perspective, a_0 can be viewed as a weight parameter. Small values of a_0 give little prior weight to the historical data. The special case $a_0 = 0$ gives a zero weight to the historical data, and the power prior with $a_0 = 0$ hence reduces to the initial prior $\pi_0(\boldsymbol{\theta})$.

• **The Bayesian Paradigm**

The Bayesian paradigm is based on specifying a probability model for the observed data D, given a vector of unknown parameters $\boldsymbol{\theta}$, leading to the likelihood function $L(\boldsymbol{\theta}|D)$.

Then we assume that $\boldsymbol{\theta}$ is random and has a *prior* distribution denoted by $\pi(\boldsymbol{\theta})$.

Inference concerning $\boldsymbol{\theta}$ is then based on the *posterior* distribution, which is obtained by Bayes' theorem. The posterior distribution of $\boldsymbol{\theta}$ is given by

$$\pi(\boldsymbol{\theta}|D) = \frac{L(\boldsymbol{\theta}|D)\pi(\boldsymbol{\theta})}{\int_{\Theta} L(\boldsymbol{\theta}|D)\pi(\boldsymbol{\theta})\, d\boldsymbol{\theta}},$$

where Θ denotes the parameter space of $\boldsymbol{\theta}$.

• **Posterior Under Power Prior**

Under the power prior, the posterior distribution of $\boldsymbol{\theta}$ can be written as

$$\pi(\boldsymbol{\theta}|D, D_0, a_0) \propto L(\boldsymbol{\theta}|D)L(\boldsymbol{\theta}|D_0)^{a_0}\pi_0(\boldsymbol{\theta}).$$

There are two interesting special cases.

- First, when $a_0 = 0$, this leads to the posterior

$$\pi(\boldsymbol{\theta}|D, D_0, a_0 = 0) \propto L(\boldsymbol{\theta}|D)\pi_0(\boldsymbol{\theta}).$$

- The other special case of interest is $a_0 = 1$, which leads to

$$\pi(\boldsymbol{\theta}|D, D_0, a_0 = 1) \propto L(\boldsymbol{\theta}|D)L(\boldsymbol{\theta}|D_0)\pi_0(\boldsymbol{\theta}).$$

Thus $\pi(\boldsymbol{\theta}|D, D_0, a_0 = 0)$ and $\pi(\boldsymbol{\theta}|D, D_0, a_0 = 1)$ represent the two extremes. In one case, no historical data is used and in the other case, the historical and current data are equally weighted, and thus $\pi(\boldsymbol{\theta}|D, D_0, a_0 = 1)$ corresponds to pooling the historical and current data.

• **Optimality Result**

The power prior can be justified as the minimizer of the convex sum of the Kullback-Leibler (KL) divergences between the posterior densities given in $\pi(\boldsymbol{\theta}|D, D_0, a_0 = 0)$ and $\pi(\boldsymbol{\theta}|D, D_0, a_0 = 1)$. Towards this goal, recall the definition of the KL divergence. Suppose f_1 and f_2 are two densities with respect to Lebesgue measure. Then the KL directed divergence between f_1 and f_2 is defined as

$$K(f_1, f_2) = \int \log\left(\frac{f_1(\boldsymbol{x})}{f_2(\boldsymbol{x})}\right) \; f_1(\boldsymbol{x}) d\boldsymbol{x}.$$

Now let $g(\boldsymbol{\theta})$ denote an arbitrary density function of $\boldsymbol{\theta}$. For convenience, denote $f_0 = \pi(\boldsymbol{\theta}|D, D_0, a_0 = 0)$ and $f_1 = \pi(\boldsymbol{\theta}|D, D_0, a_0 = 1)$. Now we consider the problem of finding the density g that minimizes the convex sum

$$K_g = (1 - a_0)K(g, f_0) + a_0 K(g, f_1),$$

where $0 \leq a_0 \leq 1$. It turns out that the density $g \equiv g(\boldsymbol{\theta})$ that minimizes K_g, denoted by g_{opt}, is

$$g_{opt} = \pi(\boldsymbol{\theta}|D, D_0, a_0) \propto L(\boldsymbol{\theta}|D)L(\boldsymbol{\theta}|D_0)^{a_0}\pi_0(\boldsymbol{\theta}).$$

This tells us that the power prior is the unique prior that minimizes K_g. We now state this as a formal theorem.

Theorem 1: *Let* $f_0 = \pi(\boldsymbol{\theta}|D, D_0, a_0 = 0)$ *and* $f_1 = \pi(\boldsymbol{\theta}|D, D_0, a_0 = 1)$. *The density* $g \equiv g(\boldsymbol{\theta})$ *that minimizes*

$$K_g = (1 - a_0)K(g, f_0) + a_0 K(g, f_1)$$

is

$$g_{opt} = \pi(\boldsymbol{\theta}|D, D_0, a_0) \propto L(\boldsymbol{\theta}|D)L(\boldsymbol{\theta}|D_0)^{a_0}\pi_0(\boldsymbol{\theta}).$$

Proof: We have

$$\begin{aligned}
&(1 - a_0)K(g, f_0) + a_0 K(g, f_1) \\
=&(1 - a_0)\int \log\Big(\frac{g(\boldsymbol{\theta})}{f_0(\boldsymbol{\theta})}\Big)g(\boldsymbol{\theta})d\boldsymbol{\theta} + a_0 \int \log\Big(\frac{g(\boldsymbol{\theta})}{f_1(\boldsymbol{\theta})}\Big)g(\boldsymbol{\theta})d\boldsymbol{\theta} \\
=&\int \log\left[\frac{g(\boldsymbol{\theta})}{f_0(\boldsymbol{\theta})}\right]^{1-a_0} g(\boldsymbol{\theta})d\boldsymbol{\theta} + \int \log\left[\frac{g(\boldsymbol{\theta})}{f_1(\boldsymbol{\theta})}\right]^{a_0} g(\boldsymbol{\theta})d\boldsymbol{\theta} \\
=&\int \log\left[\frac{g(\boldsymbol{\theta})}{f_0(\boldsymbol{\theta})^{1-a_0} f_1(\boldsymbol{\theta})^{a_0}}\right] g(\boldsymbol{\theta})d\boldsymbol{\theta} \\
=&K\left(g, \frac{f_0^{1-a_0} f_1^{a_0}}{h(a_0)}\right) - \log(h(a_0)),
\end{aligned}$$

where

$$h(a_0) = \int f_0(\boldsymbol{\theta})^{1-a_0} f_1(\boldsymbol{\theta})^{a_0} d\boldsymbol{\theta}$$

is the normalizing constant of $f_0^{1-a_0} f_1^{a_0}$.

Now clearly

$$K\left(g, \frac{f_0^{1-a_0} f_2^{1-a_0}}{h(a_0)}\right)$$

is minimized and equal to 0 when

$$g = g_{opt} = \frac{f_0^{1-a_0} f_2^{a_0}}{h(a_0)} \propto f_0^{1-a_0} f_2^{a_0}.$$

Since $f_0 = \pi(\boldsymbol{\theta}|D, D_0, a_0 = 0)$ and
$f_1 = \pi(\boldsymbol{\theta}|D, D_0, a_0 = 1)$, we have

$$\begin{aligned} g_{opt} &\propto f_0^{1-a_0} f_1^{a_0} \\ &= (\pi(\boldsymbol{\theta}|D, D_0, a_0 = 0))^{1-a_0} \, (\pi(\boldsymbol{\theta}|D, D_0, a_0 = 1))^{a_0} \\ &\propto L(\boldsymbol{\theta}|D) L(\boldsymbol{\theta}|D_0)^{a_0} \pi_0(\boldsymbol{\theta}). \end{aligned}$$

Thus, the posterior density g that achieves the desired minimum is precisely the one based on the power prior. □

The theorem thus tells us that the power prior is in this sense an optimal prior to use and in fact minimizes the convex combination of KL divergences between two extremes: one in which no historical data is used, and the other in which the historical data and current data are given equal weight (i.e., pooled). As a corollary to the theorem, we can see that $K = K(g, f_0) + K(g, f_1)$ is minimized when

$$g \propto (f_0 f_1)^{1/2} \propto L(\boldsymbol{\theta}|D) L(\boldsymbol{\theta}|D_0)^{1/2} \pi_0(\boldsymbol{\theta}).$$

This implies that if we directly minimize the sum of KL divergences between g and f_0 and g and f_1, then that minimizer is the posterior distribution based on a power prior using $a_0 = 0.5$. This result tells us that a choice of $a_0 = 0.5$ is a reasonable starting value to use in an analysis, and from which sensitivity analyses can be based.

• **Example 3: Normal Linear Model**

Consider historical data

$$\boldsymbol{y}_0 = X_0\boldsymbol{\beta} + \boldsymbol{\epsilon}_0,$$

where $\boldsymbol{\epsilon}_0 \sim N_n(0, \sigma^2 I)$, X_0 is $n_0 \times p$ of rank p, and $\boldsymbol{\beta}$ is $p \times 1$. Assume that the initial prior is $\pi_0(\boldsymbol{\beta}) \propto 1$. Similar to Example 1, the power prior is given by

$$\begin{aligned}\pi(\boldsymbol{\beta}|D_0, a_0) &\propto \exp\left\{-\frac{a_0}{2\sigma^2}(\boldsymbol{y}_0 - X_0\boldsymbol{\beta})'(\boldsymbol{y}_0 - X_0\boldsymbol{\beta})\right\} \\ &\propto \exp\left\{-\frac{a_0}{2\sigma^2}(\boldsymbol{\beta} - \boldsymbol{\mu}_0)'(X_0'X_0)(\boldsymbol{\beta} - \boldsymbol{\mu}_0)\right\},\end{aligned}$$

where $\boldsymbol{\mu}_0 = (X_0'X_0)^{-1}X_0'\boldsymbol{y}_0$. Thus, we see in this case that

$$\pi(\boldsymbol{\beta}|D_0, a_0) = N_p\left(\boldsymbol{\mu}_0, a_0^{-1}\sigma^2(X_0'X_0)\right).$$

Also, consider the current data which follows the linear model

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} \sim N_n(0, \sigma^2 I)$.

Without loss of generality, and for ease of exposition, suppose that $\sigma^2 = 1$. Let $D = (n, \boldsymbol{y}, X)$ and $D_0 = (n_0, \boldsymbol{y}_0, X_0)$. It is easily shown that the posterior

$$\pi(\boldsymbol{\beta}|D, D_0, a_0) = N_p(\boldsymbol{\mu}, \Sigma),$$

where

$$\boldsymbol{\mu} = \Lambda\boldsymbol{\mu}_0 + (I - \Lambda)\boldsymbol{\mu}_1,$$

$$\boldsymbol{\mu}_1 = (X'X)^{-1}X'\boldsymbol{y},$$

$$\Lambda = (X'X + a_0X_0'X_0)^{-1}(a_0X_0'X_0),$$

and

$$\Sigma = (X'X + a_0X_0'X_0)^{-1}.$$

• **Example 4: Exponential Model**

Suppose the current data y_i has an exponential distribution with mean $1/\theta$, $i = 1, 2, \ldots, n$, and the y_i's are i.i.d. Let the historical data y_{0i} have the same distribution as y_i, and let $\boldsymbol{y}_0 = (y_{01}, y_{02}, \ldots, y_{0n_0})'$, and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$. Further let $D = (n, \boldsymbol{y})$ and $D_0 = (n_0, \boldsymbol{y}_0)$. Take

$$\pi_0(\theta) \propto \theta^{-1}.$$

Then

$$\pi(\theta|D, D_0, a_0 = 1) = \mathcal{G}\Big(n + n_0, (n\bar{y} + n_0\bar{y}_0)^{-1}\Big),$$

where $\bar{y}$ and $\bar{y}_0$) denote the sample means corresponding to the current and historical data, respectively. Also

$$\pi(\theta|D, D_0, a_0 = 0) = \mathcal{G}\Big(n, (n\bar{y})^{-1}\Big).$$

The derivation of $\pi(\theta|D, D_0, a_0)$ is left as an exercise.

• **Example 5: Poisson Model**

Suppose the current data y_i has a Poisson distribution with mean θ, $i = 1, 2, \ldots, n$, and the y_i's are i.i.d. Let the historical data y_{0i} have the same distribution as y_i, and let $\boldsymbol{y}_0 = (y_{01}, y_{02}, \ldots, y_{0n_0})'$, and $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)'$. Further let $D = (n, \boldsymbol{y})$ and $D_0 = (n_0, \boldsymbol{y}_0)$. Take

$$\pi_0(\theta) \propto \theta^{-1}.$$

Then

$$\pi(\theta|D, D_0, a_0 = 1) = \mathcal{G}\Big(n\bar{y} + n_0\bar{y}_0, (n + n_0)^{-1}\Big).$$

Also,

$$\pi(\theta|D, D_0, a_0 = 0) = \mathcal{G}\Big(n\bar{y}, n^{-1}\Big).$$

The derivation of $\pi(\theta|D, D_0, a_0)$ is left as an exercise.

• **Extension to Multiple Historical Data Sets**

If there are K_0 historical studies, we define $D_{0k} = (n_{0k}, X_{0k}, \boldsymbol{y}_{0k})$ to be the historical data based on the k^{th} study, $k = 1, 2, \ldots, K_0$, and $D_0 = (D_{01}, \ldots, D_{0K_0})$. Letting $\boldsymbol{a}_0 = (a_{01}, \ldots, a_{0K_0})'$, the prior can be generalized as

$$\pi(\boldsymbol{\theta}|D_0, \boldsymbol{a}_0) \propto \left(\prod_{k=1}^{K_0} [L(\boldsymbol{\theta}|D_{0k})]^{a_{0k}} \right) \pi_0(\boldsymbol{\theta}).$$

Under the power prior, the posterior distribution of $\boldsymbol{\theta}$ can be written as

$$\pi(\boldsymbol{\theta}|D, D_0, \boldsymbol{a}_0) \propto L(\boldsymbol{\theta}|D) \left(\prod_{k=1}^{K_0} [L(\boldsymbol{\theta}|D_{0k})]^{a_{0k}} \right) \pi_0(\boldsymbol{\theta}),$$

where $L(\boldsymbol{\theta}|D)$ denotes the likelihood function of $\boldsymbol{\theta}$ given the current data D.

There are $(K_0 + 1)$ interesting special cases of $\pi(\boldsymbol{\theta}|D_0, \boldsymbol{a}_0)$ These special cases are at the extremes $\boldsymbol{a}_0 = (0, 0, \ldots, 0)$ and

$$\begin{aligned} &a_0 = e_k \\ \equiv &(a_{01} = 0, \ldots, a_{0,k-1} = 0, a_{0k} = 1, a_{0,k+1} = 0, \ldots, a_{0K_0} = 0), \end{aligned}$$

where e_k is a vector with a 1 in the k^{th} position and zero's elsewhere, for $k = 1, 2, \ldots, K_0$. First, when $a_0 = 0$, this leads to the posterior

$$\pi(\boldsymbol{\theta}|D, D_0, \boldsymbol{a}_0 = 0) \propto L(\boldsymbol{\theta}|D)\pi_0(\boldsymbol{\theta}).$$

The other special cases are $a_0 = e_k$, leading to

$$\pi(\boldsymbol{\theta}|D, D_0, \boldsymbol{a}_0 = e_k) \propto L(\boldsymbol{\theta}|D)L(\boldsymbol{\theta}|D_{0k})\pi_0(\boldsymbol{\theta}),$$

for $k = 1, 2, \ldots, K_0$. In $\pi(\boldsymbol{\theta}|D, D_0, \boldsymbol{a}_0 = 0)$, no historical data is used, and in $\pi(\boldsymbol{\theta}|D, D_0, \boldsymbol{a}_0 = e_k)$, only the k^{th} historical data set is used for $k = 1, 2, \ldots, K_0$.

Let $g(\boldsymbol{\theta})$ denote an arbitrary density function of $\boldsymbol{\theta}$. Also let

$$f_0(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|D, D_0, a_0 = 0)$$

and

$$f_k(\boldsymbol{\theta}) = \pi(\boldsymbol{\theta}|D, D_0, a_0 = e_k)$$

for $k = 1, 2, \ldots, K_0$. Now, we define

$$K_g^* = \left(1 - \sum_{k=1}^{K_0} a_{0k}\right) K(g, f_0) + \sum_{k=1}^{K_0} a_{0k} K(g, f_k).$$

Theorem 2: *Assume* $a_{0k} \geq 0$ *for* $k = 1, 2, \ldots, K_0$ *and* $\sum_{k=1}^{K_0} a_{0k} \leq 1$. *The density* g *that minimizes* K_g^* *is*

$$g = \pi(\boldsymbol{\theta}|D, D_0, a_0)$$

$$\propto L(\boldsymbol{\theta}|D) \left(\prod_{k=1}^{K_0} [L(\boldsymbol{\theta}|D_{0k})]^{a_{0k}}\right) \pi_0(\boldsymbol{\theta}).$$

Proof: We can write

$$
\begin{aligned}
K_g^* &= \left(1 - \sum_{k=1}^{K_0} a_{0k}\right) K(g, f_0) + \sum_{k=1}^{K_0} a_{0k} K(g, f_k) \\
&= \int \log \left(\frac{g^{1-\sum_{k=1}^{K_0} a_{0k}}}{f_0^{1-\sum_{k=1}^{K_0} a_{0k}}} \cdot \prod_{k=1}^{K_0} \frac{g^{a_{0k}}}{f_k^{a_{0k}}} \right) g d\theta \\
&= \int \log \left(\frac{g}{f_0^{1-\sum_{k=1}^{K_0} a_{0k}} \prod_{k=1}^{K_0} f_k^{a_{0k}}} \right) g d\boldsymbol{\theta} \\
&= K \left(g, \frac{f_0^{1-\sum_{k=1}^{K_0} a_{0k}} \prod_{k=1}^{K_0} f_k^{a_{0l}}}{h^*(a_0)} \right) - \log(h^*(a_0)),
\end{aligned}
$$

where

$$
\begin{aligned}
h_0^*(a_0) &= \int f_0^{1-\sum_{k=1}^{K_0} a_{0k}} \prod_{k=1}^{K_0} f_k^{a_{0k}} d\boldsymbol{\theta} \\
&= \int L(\boldsymbol{\theta}|D) \left(\prod_{k=1}^{K_0} [L(\boldsymbol{\theta}|D_{0k})]^{a_{0k}} \right) \pi_0(\boldsymbol{\theta}) d\boldsymbol{\theta}.
\end{aligned}
$$

Similar to the proof of Theorem 1, K_g^* is minimized and equal to $-\log(h^*(a_0))$ when

$$
g = g_{opt}^* \propto L(\boldsymbol{\theta}|D) \left(\prod_{k=1}^{K_0} [L(\boldsymbol{\theta}|D_{0k})]^{a_{0k}} \right) \pi_0(\boldsymbol{\theta})
$$

as desired. □

• **Random** a_0

To express certain uncertainty on a_0, we can use the hierarchical power prior specification, which is obtained by specifying a proper prior distribution for a_0.

Thus the joint power prior distribution for (θ, a_0) takes the form

$$\pi(\boldsymbol{\theta}, a_0|D_0) \ \propto \ L(\boldsymbol{\theta}|D_0)^{a_0}\pi_0(\boldsymbol{\theta})\ \pi(a_0|\boldsymbol{\gamma}_0)\ ,$$

where $\boldsymbol{\gamma}_0$ is a specified hyperparameter vector. A natural choice for $\pi(a_0|\boldsymbol{\gamma}_0)$ is a beta prior. However, other choices, including a truncated gamma prior or a truncated normal prior can be used. These three priors for a_0 have similar theoretical properties, and our experience shows that they have similar computational properties. In practice, they yield similar results when the hyperparameters are appropriately chosen. Thus, for a clear focus and exposition, we will use a beta distribution for $\pi(a_0|\boldsymbol{\gamma}_0)$.

The beta prior for a_0 appears to be the most natural prior to use and leads to the most natural elicitation scheme. Furthermore, in the presence of multiple historical datasets, the Dirichlet prior for a_0 appears the most natural. One attractive feature of $\pi(\boldsymbol{\theta}, a_0|D_0)$ is that it creates heavier tails for the marginal prior of $\boldsymbol{\theta}$ than the prior $\pi(\boldsymbol{\theta}|D_0, a_0)$, which assumes that a_0 is a fixed value. This is a desirable feature since it gives the investigator more flexibility in weighting the historical data.

Suppose that we write

$$\pi(a_0|\boldsymbol{\gamma}_0) \propto a_0^{\delta_0 - 1}(1 - a_0)^{\lambda_0 - 1}.$$

For elicitation purposes, it is easier to work with the prior mean and variance of a_0, that is,

$$\mu_{a_0} = \frac{\delta_0}{\delta_0 + \lambda_0}$$

and

$$\sigma^2_{a_0} = \mu_{a_0}(1 - \mu_{a_0})(\delta_0 + \lambda_0 + 1)^{-1}.$$

For elicitation purposes, it is typically easier to specify $(\mu_{a_0}, \sigma^2_{a_0})$ and then solve for (δ_0, λ_0) from the implied equations.

• **Example 6: Logistic Regression Model**

This is a continuation of Example 2. However, we recode β_1 and β_2 for intercept and slope. When a_0 is random, the joint prior for $(\boldsymbol{\beta}, a_0)$ is given by

$$\pi(\boldsymbol{\beta}, a_0|D_0)$$

$$\propto \prod_{i=1}^{n_0} \frac{\exp\{a_0 y_{0i} \boldsymbol{x}_{0i}' \boldsymbol{\beta}\}}{(1+\exp\{\boldsymbol{x}_{0i}' \boldsymbol{\beta}\})^{a_0}} \times a_0^{\delta_0 - 1}(1-a_0)^{\lambda_0 - 1}.$$

Figure 2 shows the marginal prior densities of β_1 and β_2 for three choices of $(\mu_{a_0}, \sigma_{a_0})$. From Figure 2, it is easy to see that as μ_{a_0} gets smaller, both marginal prior density curves get flatter; but the prior modes of β_1 and β_2 for all three choices of $(\mu_{a_0}, \sigma_{a_0})$ are almost the same. Although it is not shown in Figure 2, we also obtained the marginal prior densities for β_1 and β_2 for $(\delta_0, \lambda_0) = (3, 3)$, which are nearly uniform over the real line. For Figure 2, the corresponding (δ_0, λ_0) values are $(50, 3)$, $(20, 20)$, and $(5, 5)$, respectively.

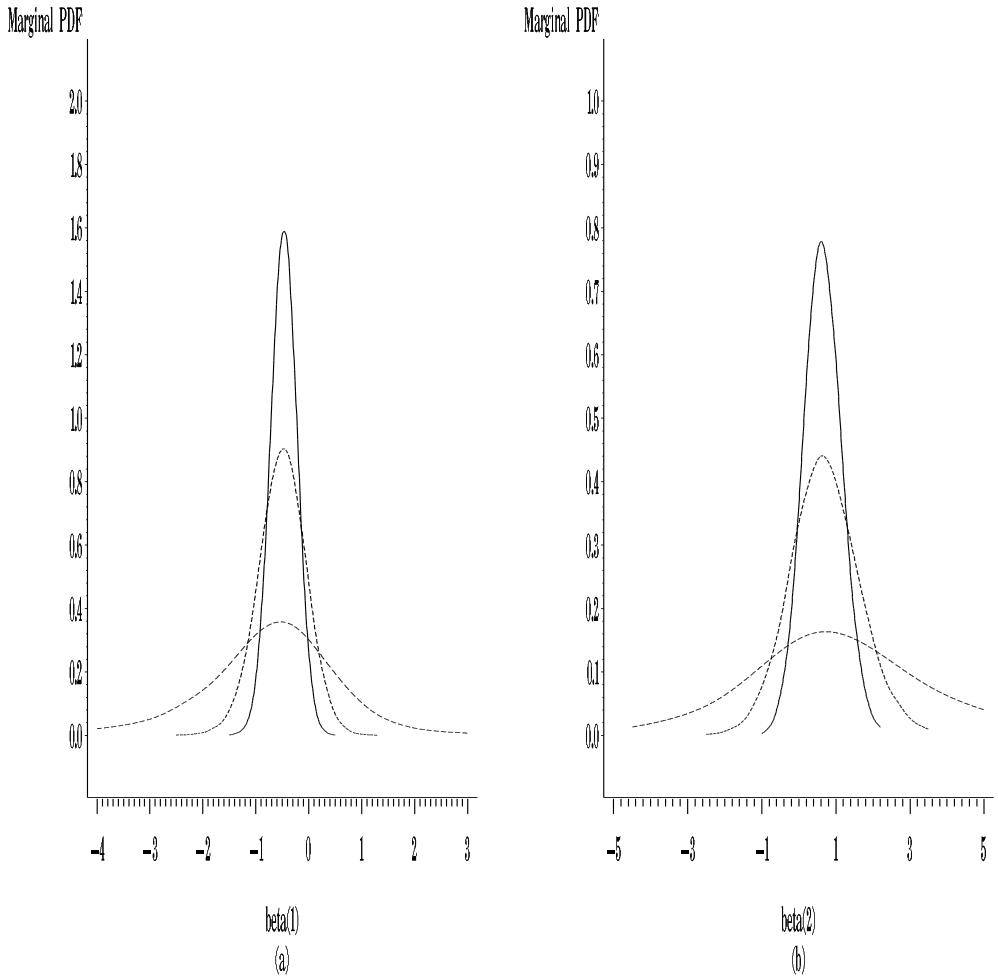

FIGURE 2. Plots of Marginal Posterior Densities for β_1 and β_2 where the solid curve is for $(\mu_{a_0}, \sigma_{a_0}) = (0.94, 0.031)$, the doted curve is for $(\mu_{a_0}, \sigma_{a_0}) = (0.5, 0.078)$, and the dashed curve is for $(\mu_{a_0}, \sigma_{a_0}) = (0.5, 0.151)$.

- **Comments**:
  - When a_0 is random, the joint power prior can be justified in a similar fashion as the fixed a_0 case.
  - For multiple historical data sets, we can choose $\pi(a_0|\boldsymbol{\gamma}_0)$ to be a K_0-dimensional Dirichlet distribution.
  - The power prior is different, but related to a prior based on a "meta-analysis" formulation, in which the parameters for the historical data and current data are different but come from a common distribution.
  - The power prior does not have a closed form in general, but it has several attractive theoretical and computational properties. This is discussed extensively in Ibrahim and Chen (2000, Statistical Science).
  - Under certain mild regularity conditions, the power prior is proper. The detailed discussions and proofs can be found in Chen, Ibrahim, and Shao (2000, JSPI).