THE UNIVERSITY OF TEXAS AT SAN ANTONIO, COLLEGE OF BUSINESS

Working Paper SERIES

December 2nd, 2008

Wp# 0058MSS-432-2008

Normalized Power Prior Bayesian Analysis

Yuyan Duan
Bristol-Myers Squibb, USA

Keying Ye
Department of Management Science and Statistics
The University of Texas at San Antonio
One UTSA Circle
San Antonio, TX 78249, USA
Email: keying@utsa.edu

Department of Management Science & Statistics,
University of Texas at San Antonio,
San Antonio, TX 78249, U.S.A

UTSA COLLEGE OF BUSINESS KNOWLEDGE FOR A NEW WORLD

ONE UTSA CIRCLE
SAN ANTONIO, TEXAS 78249-0631
210 458-4317 | BUSINESS.UTSA.EDU

Normalized Power Prior Bayesian Analysis

Yuyan Duan,
Bristol-Myers Squibb, USA
and Keying Ye*,
University of Texas at San Antonio, USA

Summary. The elicitation of power prior distributions is based on the availability of historical data, and is realized by raising the likelihood function of the historical data to a fractional power. However, an arbitrary positive constant before the likelihood function of the historical data could change the inferential results when one uses the original power prior. This raises a question that which likelihood function should be used, one from raw data, or one from a sufficient-statistics. We propose a normalized power prior that can better utilize the power parameter in quantifying the heterogeneity between current and historical data. Furthermore, when the power parameter is random, the optimality of the normalized power priors is shown in the sense of maximizing Shannon's mutual information. Some comparisons between the original and the normalized power prior approaches are made and a water-quality monitoring data is used to show that the normalized power prior is more sensible. .

Keywords: Bayesian analysis, historical data, normalized power prior, power prior, prior elicitation, Shannon's mutual information.

*Corresponding author: Keying Ye, Department of Management Science and Statistics, University of Texas at San Antonio, One UTSA Circle, San Antonio, TX 78249, USA; E-mail: keying@utsa.edu

1 Introduction

In applying statistics to real experiments, it is common that the sample size in the current study is often inadequate to provide necessary precision for parameter estimation, while plenty of historical data or data from similar studies or research settings are available. For example, to assess violations of water quality standards, measurements of chemical constituents are typically collected on a monthly or quarterly basis at each monitoring station, and then analyzed to evaluate the percentage of samples exceeding the standard. Under the Clean Water Act, only observations over a two year period are allowed to be counted as current data in the assessment. The lack of sufficient data often leads to unacceptable levels of uncertainty. In a situation like this, "historical" data, a data set from previous time periods or from adjacent stations, can be very useful in interpreting the current status of water quality, if it can be combined with current data in some way.

Due to the nature of sequential information updating, it is natural to use a Bayesian approach with an informative prior on the model parameters to incorporate the historical data into the current study. A traditional approach to incorporating historical data is to construct an informative prior using the historical data and such a prior is combined with the likelihood to yield the posterior distribution in statistical inference. This implies a simple pooling of current data and historical data together, since the two data sets are equally weighted. This approach can be well justified by assuming that the current and historical data come from exactly the same population. However, although the current and historical data are usually assumed to follow distributions in the same family, the population parameters may change over time, or over different settings. If the sample size of the historical data is much larger than that of the current data and heterogeneity exists between these data sets, historical information could dominate the analysis and the data pooling may result in misleading conclusions.

To address this issue, Ibrahim and Chen ([10], and thereafter [3], [4], [11], [12], and

others) proposed the concept of *power priors*, based on the notion of the availability of historical data. The basic idea is to let a power parameter δ ($0 \leq \delta \leq 1$) tell us how much historical data is to be used in the current study. However, in their approach, the ways in determining the historical likelihood, e.g., using the joint density of all the data, the joint densities of various sufficient statistic settings and so on, would change inferential conclusions due to the fact that the posterior distributions vary when the constants before the likelihood functions vary. Also, the power parameter has a tendency to be close to zero, which suggests that much of a historical data set may not be used in decision making. In this article, we propose a normalized power prior Bayesian approach, in which the power parameter quantifies the heterogeneity between current and historical data automatically, and hence controls the influence of historical data on the current study in a sensible way.

The article is organized as follows. In Section 2, the general development of the normalized power prior approach is given and certain properties of the approach for the Bernoulli and normal families are discussed. In Section 3, optimality of the normalized power prior approach in the sense of maximizing Shannon's mutual information will be investigated. Section 4 contains brief comparisons of the power parameters between the original and the normalized power prior methods. More of such comparisons can be found in [7]. In Section 5, as an illustration, we apply the normalized power prior to water quality data where there are clear distinction between historical and current data sets. Finally in Section 6, we summarize the properties of the normalized power prior, close the article with a brief discussion.

2 A Normalized Power Prior Approach

2.1 The Normalized Power Prior

Suppose that θ is the parameter of interest, for instance, concentration of a chemical level in a water quality measurement. Assume that such a measurement follows a distribution and $L(\theta|D_0)$ is the likelihood function of θ based on the historical data,

denoted by D_0. In this article, we assume that, given θ, historical data D_0 and current data, denoted by D, are independent random samples from an exponential family. Furthermore, denote by $\pi(\theta)$ the initial prior, which can be a noninformative prior. Given δ, the power parameter, Ibrahim and Chen ([11]) defined the power prior of θ for the current study as

$$\pi(\theta|D_0, \delta) \propto L(\theta|D_0)^{\delta}\pi(\theta). \tag{1}$$

The power parameter δ measures the portion of historical information needed in the current study.

The power prior $\pi(\theta|D_0, \delta)$ in (1) was initially elicited for fixed δ. However, since δ is not necessarily pre-determined and also because it is often difficult to specify it in practice, we may extend the case further to a random δ. A random variable δ provides the researcher with more flexibility in weighting the historical data. A natural prior for δ would be a $Beta(\alpha, \beta)$ distribution, or simply a uniform distribution, since $0 \leq \delta \leq 1$. The elicitation of the power prior on (θ, δ) is then completed by specifying a prior distribution for δ. Ibrahim and Chen ([11]) constructed the joint power prior of (θ, δ) as

$$\pi(\theta, \delta|D_0) \propto L(\theta|D_0)^{\delta}\pi(\theta)\pi(\delta), \tag{2}$$

with the posterior, given the current data D,

$$\pi(\theta, \delta|D_0, D) = \frac{L(\theta|D)L(\theta|D_0)^{\delta}\pi(\theta)\pi(\delta)}{\int_{\Theta\times\Delta} L(\theta|D)L(\theta|D_0)^{\delta}\pi(\theta)\pi(\delta)d\theta d\delta}. \tag{3}$$

In (3), any constant before $L(\theta|D_0)$ cannot be canceled out on both numerator and denominator. This could yield different posteriors if different forms of the likelihood functions are used. For instance, one can use the joint density of the whole data, or one can use the joint densities of the different forms of sufficient statistics. On the other hand, another problem of this power prior approach arises as we investigate the application of power priors on Bernoulli and normal mean models. The influence of historical data is generally small, i.e., δ is close to 0, no matter how compatible the current and historical data are. In such a case, the inference on θ is not much different from the inference when the historical data is ignored (more discussion is

referred to Section 4). Finally, this prior could also be improper. We feel that once the historical information is available, a prior elicited from such information would better be proper.

Therefore, we propose a normalized joint power prior distribution for (θ, δ) as

$$\pi(\theta, \delta|D_0) \propto \frac{L(\theta|D_0)^\delta \pi(\theta)\pi(\delta)}{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta}, \tag{4}$$

in the region of δ such that the denominator in (4) is finite.

The difference in the forms between (2) and (4) is that the prior distribution of (θ, δ) expressed in (4) is always proper given that $\pi(\delta)$ is proper, whereas it is not necessarily the case for that in (2). More importantly, multiplying the likelihood function in (2) by an arbitrary positive number may change the prior, whereas the constant is canceled out in (4). More discussion will be given in Sections 4 and 5.

Using current data to update the prior distribution $\pi(\theta, \delta|D_0)$ in (4), we derive the joint posterior distribution for (θ, δ) as

$$\pi(\theta, \delta|D_0, D) \propto L(\theta|D)\pi(\theta, \delta|D_0) \propto \frac{L(\theta|D)L(\theta|D_0)^\delta \pi(\theta)\pi(\delta)}{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta}.$$

Integrating θ out of the expression above, the marginal posterior distribution of δ can be expressed as

$$\pi(\delta|D_0, D) \propto \pi(\delta)\frac{\int_\Theta L(\theta|D)L(\theta|D_0)^\delta \pi(\theta)\ d\theta}{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta}. \tag{5}$$

Similarly, the marginal posterior distribution of θ, $\pi(\theta|D_0, D)$, is obtained by integrating δ out. If our interest is only in θ, δ may be integrated out at an earlier stage. Then $\pi(\theta|D_0, D)$ may also be developed in the way described below.

If we integrate δ out in $\pi(\theta, \delta|D_0)$ we obtain a new prior for θ, a prior that is updated by the historical information,

$$\pi(\theta|D_0) = \int \pi(\theta, \delta|D_0)d\delta \propto \pi(\theta) \int \frac{L(\theta|D_0)^\delta \pi(\delta)}{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta}\ d\delta. \tag{6}$$

With historical data appropriately incorporated, $\pi(\theta|D_0)$ can be viewed as an informative prior for the Bayesian analysis to the current data. Consequently, the

posterior distribution of θ can be written as

$$\pi(\theta|D_0, D) \propto \pi(\theta|D_0)L(\theta|D_0, D) \propto \pi(\theta)L(\theta|D) \int \frac{L(\theta|D_0)^{\delta}\pi(\delta)}{\int_{\Theta} L(\theta|D_0)^{\delta}\pi(\theta)\ d\theta}\ d\delta. \qquad (7)$$

Furthermore, similar to the extension given by Ibrahim and Chen ([11]), the priors defined in (4) can easily be generalized to multiple historical data sets. Suppose there are m historical studies. Denote by D_{0j} the historical data for the jth study, $j = 1, ..., m$ and $D_0 = (D_{01}, ..., D_{0m})$. Different weight parameter δ_j for each historical study can be used. Assume that δ_j's are i.i.d. *Beta* random variables with parameters (α, β). Denote $\underline{\delta} = (\delta_1, ..., \delta_m)$. The normalized power prior in (4) can be generalized as

$$\pi(\theta, \underline{\delta}|D_0) \propto \frac{\left\{\prod_{j=1}^{m} L(\theta|D_{0j})^{\delta_j}\pi(\delta_j|\alpha, \beta)\right\}\pi(\theta)}{\int \left\{\prod_{j=1}^{m} L(\theta|D_{0j})^{\delta_j}\right\}\pi(\theta)\ d\theta}. \qquad (8)$$

There are actually different ways this prior can be defined, depending on the way of normalization. Duan and Ye ([8]) find out that (8) is the most reasonable one.

Heterogeneity often exists among different studies but data collected at one study are relatively homogeneous. The framework introduced above would accommodate potential heterogeneity among data sets from different sources or collected at different times. For example, in water quality assessment, we could take data observed at neighboring sites as different "historical" data sets. Moreover, data collected over a long period may be divided into several historical data sets to ensure the homogeneity within each data set. In such a way, the role of historical data can be more accurately evaluated ([9]). Examples of implementing the normalized power prior approach using multiple sites information can be found therein.

2.2 Normalized Power Prior Approach for Exponential Family

In this section we are interested in making inference on the parameter θ (possibly vector-valued) of an exponential family, by incorporating both current and historical data. Denote by $D = (x_1, ..., x_n)$ the current data and $D_0 = (x_{01}, ..., x_{0n_0})$ the

historical data. Suppose that current data come from an exponential family with probability density function or probability mass function of the form (see, e.g., [5])

$$f(x|\theta) = h(x) \exp\left\{ \sum_{i=1}^{k} w_i(\theta) t_i(x) + \tau(\theta) \right\}, \tag{9}$$

where the dimension of θ is no larger than k. Here $h(x) \geq 0$ and $t_1(x), ..., t_k(x)$ are real-valued functions of the observation x, and $w_1(\theta), ..., w_k(\theta)$ are real-valued functions of the parameter θ. Define $\underline{w}(\theta) = (w_1(\theta), \ldots, w_k(\theta))'$. Furthermore, define

$$\underline{C}(\underline{x}) = \left(\frac{1}{n} \sum_{j=1}^{n} t_1(x_j), ..., \frac{1}{n} \sum_{j=1}^{n} t_k(x_j) \right)' \tag{10}$$

as the *compatibility statistic* to measure how compatible a sample $\underline{x} = (x_1, ..., x_n)$ is with other samples in providing information about θ. The density function of the current data may be expressed as

$$f(D|\theta) = h(D) \exp\left[n\{\underline{C}(D)'\underline{w}(\theta) + \tau(\theta)\} \right], \tag{11}$$

where $h(D) = \prod_{j=1}^{n} h(x_j)$ and $\underline{C}(D)$ stands for the compatibility statistic related to the current data D. Accordingly, the compatibility statistic and the density function similar to (10) and (11) respectively for the historical data D_0 can be defined as well.

Denote by $\pi(\theta)$ the initial prior distribution of θ and $\pi(\delta)$ denote the prior distribution of the power parameter. We write the joint posterior distribution of (θ, δ) as

$$\pi(\theta, \delta|D_0, D) \propto \frac{\exp\left[\{\delta n_0 \underline{C}(D_0)' + n\underline{C}(D)'\}\underline{w}(\theta) + (\delta n_0 + n)\tau(\theta) \right] \pi(\theta)\pi(\delta)}{\int_\Theta \exp\left[\delta n_0 \{\underline{C}(D_0)'\underline{w}(\theta) + \tau(\theta)\} \right] \pi(\theta)\ d\theta} \tag{12}$$

Integrating θ out in (12), the marginal posterior distribution of δ is given by

$$\pi(\delta|D_0, D) \propto \pi(\delta) \frac{\int_\Theta \exp\left[(\delta n_0 \underline{C}(D_0)' + n\underline{C}(D)')\underline{w}(\theta) + (\delta n_0 + n)\tau(\theta) \right] \pi(\theta)\ d\theta}{\int_\Theta \exp\left\{ \delta n_0 [\underline{C}(D_0)'\underline{w}(\theta) + \tau(\theta)] \right\} \pi(\theta)\ d\theta}.$$

The behavior of the power parameter δ can be examined from this marginal posterior distribution. Similarly, the marginal posterior distribution of θ can be derived by integrating δ out in $\pi(\theta, \delta|D_0, D)$, but it often does not have a closed form. Instead the posterior distribution of θ given D_0, D and δ is often in a more familiar form. Therefore we may learn the characteristic of the marginal posterior of θ by studying the conditional posterior distribution $\pi(\theta|D_0, D, \delta)$, combined with $\pi(\delta|D_0, D)$.

2.2.1 Bernoulli Population

Suppose we are interested in making inference on the probability of success p from a Bernoulli population. Define $y_0 = \sum_{i=1}^{n_0} x_{0i}$, and $y = \sum_{j=1}^{n} x_j$. The joint posterior distribution of p and δ can be easily derived as the result below and the proof is omitted.

Result 1. Assume that the initial prior distribution of p follows a $Beta(\alpha_p, \beta_p)$, and the prior distribution of δ follows a $Beta(\alpha_\delta, \beta_\delta)$ distribution, where the hyper-parameters $\alpha_p, \beta_p, \alpha_\delta$ and β_δ are all known. The joint posterior distribution of (p, δ) is

$$\pi(p, \delta | D_0, D) \propto \frac{p^{\delta y_0 + y}(1-p)^{\delta(n_0 - y_0)+(n-y)} \delta^{\alpha_\delta - 1}(1-\delta)^{\beta_\delta - 1}}{B(\delta y_0 + \alpha_p, \delta(n_0 - y0) + \beta_p)},$$

where $B(a, b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)}$ stands for the *beta* function.

Integrating p out in $\pi(p, \delta | D_0, D)$, the marginal posterior distribution of δ is given by

$$\pi(\delta | D_0, D) \propto \frac{B(\delta y_0 + y + \alpha_p, \delta(n_0 - y_0) + n - y + \beta_p)}{B(\delta y_0 + \alpha_p, \delta(n_0 - y0) + \beta_p)} \delta^{\alpha_\delta - 1}(1-\delta)^{\beta_\delta - 1}.$$

The conditional posterior distribution of p given δ follows a $Beta(\delta y_0 + y + 1, \delta(n_0 - y_0) + (n - y) + 1)$. However, the marginal posterior distribution of p does not have a close form . An application of the normalized power prior for Bernoulli data can be found in [6].

2.2.2 Normal Population

Suppose we are interested in making inference on the normal mean from a normal $N(\mu, \sigma^2)$ population with unknown mean μ and variance σ^2. Define

$$\bar{x}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} x_{0i}, \quad \bar{x} = \frac{1}{n}\sum_{j=1}^{n} x_j, \quad \hat{\sigma}_0^2 = \frac{1}{n_0}\sum_{i=1}^{n_0}(x_{0i} - \bar{x}_0)^2, \text{ and } \hat{\sigma}^2 = \frac{1}{n}\sum_{j=1}^{n}(x_i - \bar{x})^2.$$

Following (4), the normalized power prior for the normal population with unknown variance is given in the following result.

Result 2. Denote by $\pi(\mu, \sigma^2)$ the initial prior distribution for (μ, σ^2). Assume that the prior distribution of δ follows a $beta(\alpha, \beta)$, where parameters α and β are known. The normalized power prior distribution of (μ, σ^2, δ) is

$$\pi(\mu, \sigma^2, \delta|D_0) \propto \frac{(\sigma^2)^{-\frac{\delta n_0}{2}} \exp\left[-\frac{\delta n_0}{2\sigma^2}\{\hat{\sigma}_0^2 + (\mu - \bar{x}_0)^2\}\right] \pi(\mu, \sigma^2)\delta^{\alpha-1}(1-\delta)^{\beta-1}}{\int_0^\infty \int_{-\infty}^{+\infty} (\sigma^2)^{-\frac{\delta n_0}{2}} \exp\left[-\frac{\delta n_0}{2\sigma^2}\{\hat{\sigma}_0^2 + (\mu - \bar{x}_0)^2\}\right] \pi(\mu, \sigma^2) d\mu d\sigma^2}.$$

When considering a special form of $\pi(\mu, \sigma^2)$, we are led to Corollaries 2.1, 2.2, and 2.3 whose proofs are simple and thus omitted.

Corollary 2.1. Suppose that we use the prior $\pi(\mu, \sigma^2) \propto (\frac{1}{\sigma^2})^a$ as the initial prior of (μ, σ^2), where $a > 0$ is a pre-determined constant. The joint power prior distribution of (μ, σ^2, δ) can be expressed as

$$\pi(\mu, \sigma^2, \delta|D_0) \propto \frac{\delta^{\frac{\delta n_0}{2}+a+\alpha_\delta-2}(1-\delta)^{\beta_\delta-1}}{\left(\frac{2\sigma^2}{n_0\hat{\sigma}_0^2}\right)^{\frac{\delta n_0}{2}+a}\Gamma\left(\frac{\delta n_0-3}{2}+a\right)} \exp\left[-\frac{\delta n_0}{2\sigma^2}\{\hat{\sigma}_0^2 + (\mu - \bar{x}_0)^2\}\right].$$

Note that $a = 1$ corresponds to the reference prior ([2]), while $a = \frac{3}{2}$ corresponds to the Jeffreys prior ([13]).

Corollary 2.2. Assume $\pi(\mu, \sigma^2) \propto (\frac{1}{\sigma^2})^a$. The marginal posterior distribution of δ is

$$\pi(\delta|D_0, D) \propto \frac{\delta^{\frac{\delta n_0}{2}+a+\alpha_\delta-2}(1-\delta)^{\beta_\delta-1}\Gamma(\frac{\delta n_0+n-3}{2}+a)}{\left\{\frac{\delta n}{\delta n_0+n}\frac{(\bar{x}_0-\bar{x})^2}{\hat{\sigma}_0^2} + \delta + \frac{n}{n_0}\frac{\hat{\sigma}^2}{\hat{\sigma}_0^2}\right\}^{\frac{\delta n_0+n-3}{2}+a}\Gamma(\frac{\delta n_0-3}{2}+a)}.$$

Corollary 2.3. Assume $\pi(\mu, \sigma^2) \propto (\frac{1}{\sigma^2})^a$. The conditional posterior distribution of μ, given δ and data (D_0, D), follows a Student t-distribution with, respectively, the location parameter and the scale parameter

$$\left(\frac{\delta n_0\bar{x}_0 + n\bar{x}}{\delta n_0 + n}, \qquad \sqrt{\frac{2}{G(\delta)}\frac{1}{(\delta n_0 + n + 2a - 3)(\delta n_0 + n)}}\right),$$

and degrees of freedom $\delta n_0 + n + 2a - 3$, where

$$G(\delta) = \frac{2}{\frac{\delta n_0 n(\bar{x}_0-\bar{x})^2}{\delta n_0+n} + \delta n_0\hat{\sigma}_0^2 + n\hat{\sigma}^2}.$$

Furthermore, the conditional posterior distribution of σ^2, given δ and the data, follows an inverse-gamma distribution with parameters $\frac{\delta n_0+n+2a-3}{2}$ and $G(\delta)^{-1}$.

Duan, *et al.* ([9]) provides an example of implementing the normalized power prior for a normal population with unknown variance.

3 Optimality Properties of the Normalized Power Prior

The optimality properties of the normalized power prior will be investigated in two steps. Section 3.1 shows that, given a fixed δ, the derived posterior $\pi(\theta|D_0, D, \delta)$ minimizes the expected loss from the true posterior distribution of θ. In Section 3.2, with δ being random, the normalized power prior yields a posterior $\pi(\delta|D_0, D)$ that maximizes the observed mutual information between historical and current data.

3.1 Optimality of Power Priors Conditional on δ

Assuming that the power parameter δ is fixed, the normalized power prior can be justified as a minimizer of the expected loss. Since the Kullback-Leibler (KL) divergence ([14]) is commonly used to measure the distance between two densities, here we use the KL divergence as the loss function between the true posterior density of θ and its estimated density. Recall the definition of the KL divergence,

$$K(g, f) = \int \log\left(\frac{g(\theta)}{f(\theta)}\right) g(\theta)\ d(\theta),$$

where g and f are two densities with respect to Lebesgue measure.

If the historical data truly come from the population underlying the current data, two samples should be pooled and hence the true posterior density of θ is $C_1 L(\theta|D_0)L(\theta|D)\pi(\theta)$, denoted by f_1. Otherwise, if the historical data and current data come from different populations so that they should not be pooled together for inference, no historical data should be incorporated and hence the true posterior density of θ is $C_0 L(\theta|D)\pi(\theta)$, denoted by f_0. Both C_1 and C_0 are normalization constants. Now let $g(\theta)$ denote an arbitrary density function of θ and $f(\theta)$ denote the true posterior distribution of θ. Then the expected loss of using the density g to

estimate the true posterior distribution of θ can be written as

$$L_g \equiv E(K(g,f)) = Pr(f = f_0)K(g,f_0) + Pr(f = f_1)K(g,f_1).$$

Furthermore, δ can be interpreted as the probability that D_0 follow the same distribution as D, since δ is initiated to measure how much of historical data should be used in analyzing current data's distribution. It follows that

$$L_g = (1-\delta)K(g,f_0) + \delta K(g,f_1).$$

It has been shown by Ibrahim *et al.* ([12]) that the unique minimizer for L_g is the posterior distribution derived using the power prior.

$$\pi(\theta|D_0, D, \delta) \propto L(\theta|D_0)^{\delta} L(\theta|D)\pi(\theta). \tag{13}$$

The $\pi(\theta|D_0, D, \delta)$ based on the normalized power prior is the same as that based on the original approach proposed by Ibrahim and Chen ([11]). Therefore the normalized power prior is optimal in a sense that its conditional posterior distribution of θ is expected to be closest to the true posterior when using KL divergence as the loss function.

In addition, $\pi(\theta|D_0, D, \delta)$ in (13) is a 100% efficient information processing rules (IPR) in the sense that the ratio of the output to input information is equal to 1, as showed by Ibrahim *et al.* ([12]).

Based on Zellner's theory of IPR ([16] and [17]), a weighted version of the information criterion function $\Delta[g(\theta)]$ is considered in our scenario.

$$\begin{aligned}
&\Delta[g(\theta)] = \text{Output information} - \text{Input information} \\
&= \int g(\theta) \ln g(\theta)\ d\theta + \int g(\theta) \ln m(D, D_0)\ d\theta \\
&\quad - \left\{ \int g(\theta) \ln \pi(\theta)\ d\theta + \int g(\theta) \ln L(\theta|D)\ d\theta + \delta \int g(\theta) \ln L(\theta|D_0)\ d\theta \right\},
\end{aligned} \tag{14}$$

where $g(\theta)$ denotes a proper posterior density $\pi(\theta|D, D_0)$ in our setting.

Zellner defined a rule to be 100% efficient whenever $\Delta[g(\theta)] = 0$; that is, output information equals input information. It turns out that the $g^*(\theta) = \pi(\theta|D_0, D, \delta)$

obtained using power prior yields $\Delta[g^*(\theta)] = 0$. To satisfy Zellner's optimal IPR, further results are derived below and those will contribute to next section's discussion on optimality of $\pi(\delta|D_0, D)$.

Meanwhile, to achieve $\Delta[g^*(\theta)] = 0$, the $m(D, D_0)$ in (14) has to be in the form of

$$m^*(D, D_0) = \int_\Theta L(\theta|D)L(\theta|D_0)^\delta \pi(\theta)\ d\theta.$$

This can be easily verified by substituting $g(\theta)$ with $\pi(\theta|D_0, D, \delta)$ in (14). Notice that $m^*(D, D_0)$ depends on δ. However, it is not necessarily a proper probability density function with respect to D and D_0. The marginal density of (D, D_0) given δ can be derived by normalizing $m^*(D, D_0)$.

$$\begin{aligned} m(D, D_0|\delta) &= \frac{\int_\Theta L(\theta|D)L(\theta|D_0)^\delta \pi(\theta)\ d\theta}{\int\int \left\{\int_\Theta L(\theta|D)L(\theta|D_0)^\delta \pi(\theta) d\theta\right\}\ dDdD_0} \\ &= \frac{\int_\Theta L(\theta|D)L(\theta|D_0)^\delta \pi(\theta)\ d\theta}{\int\int_\Theta L(\theta|D_0)^\delta \pi(\theta) d\theta\ dD_0}. \end{aligned} \tag{15}$$

If current data have not come into play in Zellner's IPR, i.e., no $\int g(\theta) \ln L(\theta|D) d\theta$ in (14), we have

$$m^*(D_0) = \int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta.$$

Consequently, we obtain the marginal density of D_0 given δ by normalizing $m^*(D_0)$.

$$m(D_0|\delta) = \frac{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta}{\int\int_\Theta L(\theta|D_0)^\delta \pi(\theta) d\theta\ dD_0}. \tag{16}$$

Following (15) and (16), $m(D|D_0, \delta)$ can be written as

$$m(D|D_0, \delta) = \frac{m(D, D_0|\delta)}{m(D_0|\delta)} = \frac{\int_\Theta L(\theta|D)L(\theta|D_0)^\delta \pi(\theta)\ d\theta}{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta}. \tag{17}$$

We will use (17) for our further investigation on optimality when δ is random.

3.2 Optimality of the Normalized Power Prior When δ Is Random

Define $\ln[m(D|D_0)/m(D)]$ as the *observed* mutual information between two arbitrary samples D_0 and D, where $m(D) = \int_\Theta L(\theta|D)\pi(\theta)\ d\theta$ is the marginal density of D, and

$m(D|D_0)$ is the density of D given that D_0 is observed. This concept was first used by Shannon ([15]) in his theory of mutual information to measure the dependency between two variables X and Y. Shannon's mutual information is defined by the expected entropy difference,

$$\vartheta(Y \wedge X) \equiv H(Y) - E_x\{H(Y|x)\} = E_{(x,y)}\left\{ \ln \frac{f(x|y)}{f(x)} \right\},$$

where $H(Y)$ is the entropy of $f(y)$ and $H(Y|x)$ is the entropy of the conditional distribution $f(y|x)$. Shannon's mutual information is a measure of the expected information about Y transmitted through a "noisy" channel, which is represented by X. In our case, the observed mutual information $\ln \frac{m(D|D_0)}{m(D)}$ measures the amount of information in historical data that is useful in interpreting the current data.

The true $m(D|D_0)$ and $m(D)$ are learned through the sampling distribution of current data as well as priors on model parameters. In addition, $m(D|D_0)$ also depends on how historical data are incorporated, which can be recognized in the following breakdown

$$\ln m(D|D_0) = \ln \frac{m(D|D_0, \delta)\pi(\delta|D_0)}{\pi(\delta|D_0, D)}. \tag{18}$$

Note that our discussion in this section is within the power prior framework defined by (1). The power prior method with a fixed δ has been well justified as an optimal method in Section 3.1. So here it is sufficient to show, among extensions to the case in which δ is random, our proposed normalized power prior provides an optimal way to handle the random δ.

As discussed in Section 2, we believe that historical data alone does not provide additional information about δ, because δ is introduced to measure the compatibility between the historical and current data. This implies that the information of δ in $\pi(\delta|D_0)$ should be the same as that in $\pi(\delta)$. Using Zellner's definition ([16]),

$$\text{the information of } \delta \text{ in } \pi(\delta|D_0) = E_{\pi(\delta|D,D_0)} \ln \pi(\delta|D_0), \text{ and}$$
$$\text{the information of } \delta \text{ in } \pi(\delta) = E_{\pi(\delta|D,D_0)} \ln \pi(\delta)$$

are hence interchangeable during the derivation.

Considering the above characteristics of the framework of power prior Bayesian analysis, the observed mutual information between D and D_0, which measures the information in historical data transmitted through a power prior model, can be written as

$$\begin{aligned}\varpi(D \wedge D_0) &= \ln \frac{m(D|D_0)}{m(D)} = E_{\pi(\delta|D,D_0)}\left\{ \ln \frac{m(D|D_0)}{m(D)} \right\} \\ &= E_{\pi(\delta|D,D_0)}\left\{ \ln \frac{m(D|D_0,\delta)\pi(\delta)}{\pi(\delta|D_0,D)} - \ln m(D) \right\},\end{aligned}$$

where $m(D|D_0,\delta)$ is defined in (17). We have the following result whose proof is given in the Appendix.

Theorem 1: *The density $\pi(\delta|D_0,D)$ that maximizes $\varpi(D_0 \wedge D)$ is*

$$\pi^*(\delta|D_0,D) \propto \pi(\delta)\frac{\int_\Theta L(\theta|D)L(\theta|D_0)^\delta \pi(\theta)\ d\theta}{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\ d\theta}.$$

Note that $\pi^*(\delta|D_0,D)$ is precisely the marginal posterior of δ based on the normalized power prior (5). Theorem 1 states that the maximum expected information of the current data through the "noise-channel" of historical data is achieved by using the normalized power prior. Hence when the power parameter δ is random, the normalized power prior reaches optimum when Shannon's mutual information criterion is of interest.

4 Behavioral Comparisons Between Two Power-Prior Approaches

As mentioned in Section 2.1, for the original power prior, multiplying the likelihood function $L(\theta|D_0)$ by a positive constant k could change inferential results. However, the results would not change for the normalized power prior approach.

Although the joint power priors of (θ,δ) are different, the conditional power prior $\pi(\theta|D_0,\delta)$ in (1) and the conditional posterior $\pi(\theta|D_0,D,\delta)$ in (13) are the same for both approaches. This feature indicates that the two approaches are equivalent for a fixed δ, which is expected because both approaches are rooted in the same idea

presented by the definition of $\pi(\theta|D_0, \delta)$. This also implies that the differences in results between two approaches come from their difference in the posterior marginal distributions of δ. Therefore we may examine their differences in $\pi(\theta, \delta|D_0, D)$ by comparing $\pi(\delta|D_0, D)$ between two approaches.

The marginal posterior mode of δ represents the most likely value of δ given by the historical and current data and it will be used to compare the posterior distributions of the two approaches. Since $\pi(\delta|D_0, D)$ is often asymmetric, the marginal posterior mode of δ is an important statistic for studying the marginal posterior distribution of δ.

To discuss how well the marginal posterior mode of δ responses to the compatibility between the current and historical data, the notion of "compatibility statistic" is defined in (10) for the exponential family with density (9).

Clearly, $T(\underline{x}) = (\sum_{j=1}^{n} t_1(x_j), ..., \sum_{j=1}^{n} t_k(x_j))$ is a sufficient statistic for θ ([5]). One underlying assumption of this sufficiency is that the sample size n is fixed when the experiment is performed repeatedly. However, the current and historical data often have different sample sizes. This then raises the question of how to measure the difference between two samples with unequal sizes in terms of their information about θ.

Using (10) as the compatibility statistic of a sample $\underline{x} = (x_1, ..., x_n)$ for θ, we note that $C(\underline{x}) = \frac{y}{n} = \bar{x}$ for the Bernoulli case, and $\underline{C}(\underline{x}) = (\bar{x}, \hat{\sigma}^2)$ for the normal case, where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$ is the maximum likelihood estimator of σ^2.

Applying the concept of the compatibility statistic on our investigation of power priors, we have the following result whose proof is in the Appendix.

Theorem 2: *Suppose that historical data D_0 and current data D are two independent random samples from an exponential family given in (9). Define the compatibility statistic for the historical data and current data are $\underline{C}(D_0)$ and $\underline{C}(D)$ respectively.Then the marginal posterior mode of δ is always* 1 *under the normalized power prior approach, if*

$$\frac{d}{d\delta}\ln\pi(\delta) + h_1(D_0, D, \delta) + n_0\{\underline{C}(D_0) - \underline{C}(D)\}'\underline{h}_2(D_0, D, \delta) \geq 0, \qquad (19)$$

for all $0 \leq \delta \leq 1$, *where*

$$h_1(D_0, D, \delta) = \frac{n_0}{n} \int_\Theta \ln L(\theta|D)\{\pi(\theta|D_0, D, \delta) - \pi(\theta|D_0, \delta)\}d\theta,$$

and

$$\underline{h}_2(D_0, D, \delta) = \int_\Theta \underline{w}(\theta)\{\pi(\theta|D_0, D, \delta) - \pi(\theta|D_0, \delta)\}d\theta.$$

The first term in (19) is always non-negative if the prior density of δ is a non-decreasing function. The second term, $h_1(D_0, D, \delta)$, is always non-negative by using the property of Kullback-Liebler divergence (see proof in the Appendix), and it is 0 if and only if $\pi(\theta|D_0, D, \delta) = \pi(\theta|D_0, \delta)$ of which the current data, D, does not contribute to any information about θ, given δ. This could be a rare case. The values in third term depends on how closely the compatibility statistics $\underline{C}(D_0)$ and $\underline{C}(D)$ are to each other. In a special case that when $\underline{C}(D_0)$=$\underline{C}(D)$ (historical and current data are fully compatible), the posterior mode of δ is always 1. This is rational since when the historical data contribute necessary information into the current study, it should be used as much as possible to achieve higher precision.

Although the probability of being fully compatible between D_0 and D is theoretically impossible in continuous distribution cases, as long as the difference between $\underline{C}(D_0)$ and $\underline{C}(D)$ is negligible from a practical point of view, it is appropriate to view the historical and current samples as fully compatible, and hence the marginal posterior mode of δ would be 1 or very close to 1 under the normalized power prior approach.

On the other hand, in the original power prior approach, the posterior mode of δ changes if we multiply the likelihood function by a constant. We have the following result.

Theorem 3: *Suppose that current data D are from a population with a density function $f(x|\theta)$, and D_0 is a related historical data set. Furthermore, suppose that the prior $\pi(\delta)$ is a non-increasing function and the conditional posterior distribution of θ on δ is proper for any δ. Then for any D_0 and D, if*

$$\max_{0\leq\delta\leq1} \frac{\int \pi(\theta) f(D|\theta) f(D_0|\theta)^\delta \ln f(D_0|\theta) d\theta}{\int \pi(\theta) f(D|\theta) f(D_0|\theta)^\delta d\theta} < \infty, \tag{20}$$

then there exists at least one positive constant k_0 *such that* $\pi(\delta|D_0, D)$ *has mode at* $\delta = 0$ *under the original power prior approach, where* $L(\theta|x) = k_0 f(x|\theta)$.

The assumption in (20) is valid in the case that all the integrals in the numerator as well as denominator are finite positive values when δ is either 0 or 1. Usually this condition satisfies when $\pi(\theta)$ is smooth. The proof of this result is also given in the Appendix. For a normal or a Bernoulli population, our research reveals that $\pi(\delta|D_0, D)$ has mode at $\delta = 0$ in many scenarios. This strong tendency of δ towards 0 in the original approach compromises the flexibility of using a random δ. Also, the role of historical data is underestimated. In Section 5, we illustrate this in an example.

5 Applying Normalized Power Prior to a Water-Quality Data

When applying Bayesian analysis with power priors to water quality data, past information could be utilized. In this example, we use measurements of pH to evaluate impairment of four sites in Virginia individually. Of interest in these data sets is the determination of whether the pH values at a site indicate that the site violates a (lower) standard of 6.0 more than 10% of the time. For each site, larger sample size is associated with the historical and smaller with the current data. We compare the normalized power prior approach with a traditional Bayesian approach using the reference prior, and the original power prior approach. Suppose that the measurements of water quality follow a normal distribution, and for ease of comparison, the normal model with a simple mean is considered. Note that there are many other things, such as spatial and temporal features of the data and so on, may be considered in this data, we only use it as an illustration to implement our normalized power prior method.

In this example, pH data collected over a two-year or three-year period are treated as the current data, while pH data collected over the previous nine years represents

one single historical data set. The current data and historical data are plotted side by side for each site in Figure 1. In the power prior approach, a violation is evaluated using a Bayesian test of

$$H_0 : L \geq 6.0 \text{ (no impairment, don't list)},$$
$$H_1 : L < 6.0 \text{ (impairment, list)},$$

where L is the lower 10th percentile of the distribution for pH. Comparison of results from different methods is presented in Table 1.

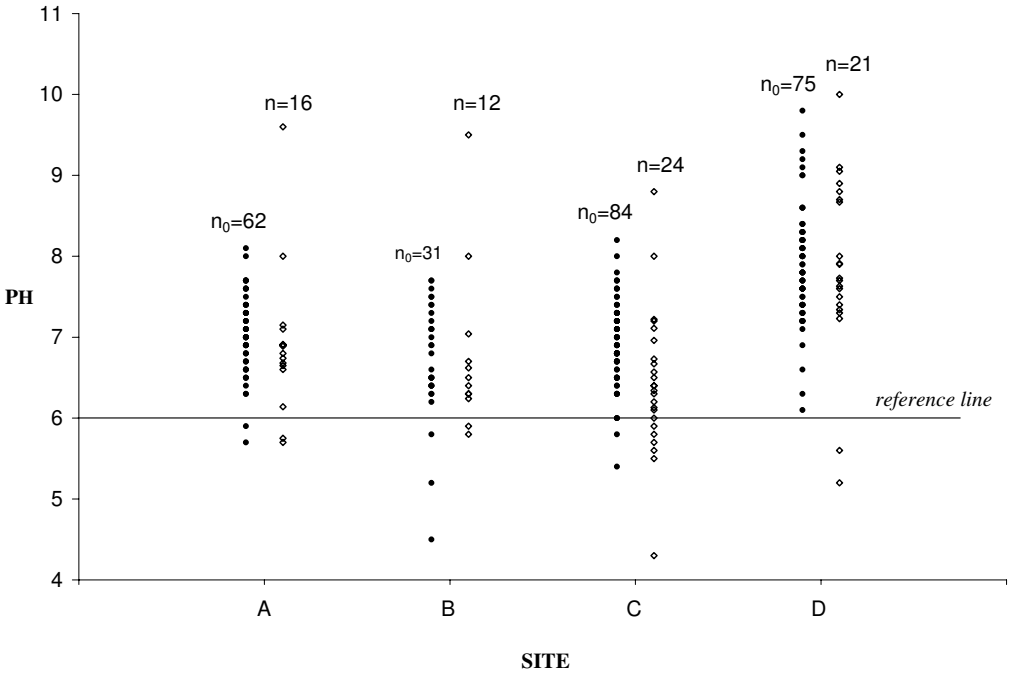

Figure 1: pH data collected at four stations. For each site, historical data are on the left (circle) and current data on the right (diamond).

In Table 1, the summarization of the current and historical data are given. The test results using the reference prior analysis (without incorporating historical data) and both normalized and original power prior analyses (with reference prior as the initial prior for (μ, σ^2), i.e, $a = 1$ in Section 2.2.2) are presented. As shown in Theorem 3, the posterior mode of δ changes in the original power prior approach if we multiply the likelihood function by a constant. Therefore, results from the original power prior are calculated using three different likelihood functions: (1) use the joint density of

Table 1: *Comparison of the power prior method with alternative methods in evaluating site impairment when one historical data set is available. In the table, n and n_0 are sample sizes, mean (s.d.) refers to sample mean (sample standard deviation), and s.d. of L is the posterior standard deviation of L.*

Site	Current data		Historical data		Posterior probability of H_0 (s.d. of L)				
	n	mean (s.d.)	n_0	mean (s.d.)	Reference prior	Normalized power prior	Original power prior (1)	(2)	(3)
A	16	6.91 (0.90)	62	7.05 (0.47)	0.2074 (0.27)	0.6027 (0.21)	0.996 (0.01)	0.9982 (0.01)	0.2362 (0.26)
B	12	6.78 (1.03)	31	6.73 (0.71)	0.0627 (0.34)	0.0294 (0.19)	0.0252 (0.03)	0.024 (0.01)	0.0609 (0.33)
C	24	6.43 (0.88)	84	6.95 (0.49)	0.0003 (0.26)	0.0017 (0.24)	0.0003 (0.26)	0.4601 (0.18)	0.0002 (0.26)
D	21	7.87 (1.11)	75	7.88 (0.67)	0.8673 (0.36)	0.9831 (0.25)	0.8879 (0.34)	0.9199 (0.32)	0.8759 (0.35)

sufficient statistics, i.e. $L(\mu, \sigma^2|D_0) = f(\bar{x}_0, S_0^2|\mu, \sigma^2)$, where $\bar{x}_0$ and S_0^2 are the sample mean and variance of historical data, respectively; (2) use the likelihood function without constant, i.e., $L(\mu, \sigma^2|D_0) = \frac{1}{(\sigma^2)^{n_0/2}} \exp\left[-\{n_0(\bar{x}_0 - \mu)^2 + (n-1)s_0^2\}/2\sigma^2\right]$; (3) use an arbitrary constant, $L(\mu, \sigma^2|D_0) = e^{-200}(2\pi)^{n_0/2} f(\underline{x_0}|\mu, \sigma^2)$.

If the 0.05 significance level is used, the reference prior Bayesian test using the reference prior would only indicate site C as impaired. Here we use the posterior probability of H_0 as equivalent to the p-value (see [1]) for testing a one-sided hypothesis. Using historical data does lead to different conclusions for site B. The test using either normalized or original power prior with density of sufficient statistics as likelihood results in significance for sites B & C. In the case of site B, there are around 10% of historical observations below 6.0. Hence our prior opinion of the site is suggestive of impairment. Less information is therefore required to declare impairment relative to a reference prior and the result is a smaller p-value. However, if one use the likelihood function in case (2) of the original power prior method, the test result is ambiguous. Furthermore, if we use an arbitrary constant as in case (3) of the original power prior situations, the marginal posterior modes of δ are always 0 and the results can be different from the others. Hence, this example shows that inference results are

sensitive to which likelihood form one would like to use in employing original power prior approach.

Another notable advantage of the power prior method is that it improves the estimation of L by using past information. This can be shown by the consistently smaller posterior standard deviation of L with the power prior than with the reference prior for all four sites.

6 Discussion

The power prior method provides a framework to incorporate data from alternative sources, whose influence on inference is automatically adjusted according to its availability and discrepancy from current data. As consequence of using more data, the power prior method has advantages in terms of power and estimation precision for decisions with small sample sizes (see [9] for more discussion).

On one hand, the power prior method can be used to solve the problems with small sample size. On the other hand, the power prior may be viewed as a general class of informative priors in Bayesian inference. The power prior is elicited to take into account the heterogeneity between historical and current data when we are not able to describe or adequately model the heterogeneity explicitly. The power priors are semi-automatic, in the sense that they take the form of raising the likelihood function based on the historical data to a fractional power regardless of the specific form of heterogeneity. The fact that we often do not have enough knowledge to model such heterogeneity or to specify a fixed power makes this power prior with a random power parameter δ especially attractive in practice.

The normalized power prior with a random power parameter is very flexible in determining the role of historical data. The subjective information about the difference in two populations is incorporated by adjusting the hyperparameters in the prior for δ; and the discrepancy between two samples is automatically taken into account through a random δ.

The normalized and original power prior approaches are essentially the same when the power parameter is fixed. Therefore the normalized power prior shares all the nice properties of the original one discussed in a series of papers by Ibrahim and Chen ([11], [12]), such as the generality of this methodology, the optimality from the aspect of information processing, the flexibility in expressing the uncertainty about the power parameter, and the wide applications. In addition, the controlling role of the power parameter in the normalized power prior approach is adjusted automatically based on the compatibility between the historical and current samples, and also based on their sample sizes. With the normalized power prior, the power parameter behaves in a sensible and desirable way. However, the original power prior approach underestimates the influence of historical data on the current study in general and therefore little benefits are gained from incorporation of historical data. Furthermore, empirical evidences show that the normalized power prior leads to smaller MSE for estimated θ than the original one, when the divergency between historical and current populations is small to moderate (see [7]).

Acknowledgement

This research was supported in part by a grant from the College of Business at University of Texas at San Antonio.

References

[1] Berger, J.O. (1985). *Statistical Decision Theory and Bayesian Analysis*, 2nd edition, Springer Verlag, New York.

[2] Berger, J.O. and Bernardo, J.M. (1992). On the development of reference priors, in *Bayesian Statistics 4: Proceedings of the Fourth Valencia International Meeting*, Bernardo, J.M, Berger, J.O., Dawid, A.P. and Smith, A.F.M. eds., 35-60, Clarendon Press: Oxford.

[3] Chen, M.-H. and Ibrahim, J.G. (2006). The relationship between the power prior and hierarchical models. *Bayesian Analysis* **1**: 551-574.

[4] Chen, M.-H., Ibrahim, J.G., and Shao, Q.-M. (2000). Power prior distributions for generalized linear models. *Journal of Statistical Planning and Inference* **84**: 121-137.

[5] Casella, G. and Berger, J.O. (2001). *Statistical Inference*, 2nd edition, Duxbury.

[6] Duan, Y., Smith E.P., and Ye, K. (2006) Using Power Priors to Improve the Binomial Test of Water Quality. *Journal of Agricultural, Biological, and Environmental Statistics* **11**: 1-18.

[7] Duan, Y. and Ye, K. (2008), Comparisons between two power prior approaches. Manuscript under preparation.

[8] Duan, Y. and Ye, K. (2008), Normalized power prior Bayesian analysis with multiple sites. Manuscript under preparation.

[9] Duan, Y., Ye, K., and Smith E.P. (2006), Evaluating water quality: using power priors to incorporate historical information. *Environmetrics* **17**: 95-106.

[10] Ibrahim, J.G., and Chen, M.-H. (1998). Prior distributions and Bayesian computation for proportional hazard models. *Sankhya, Series B* **60**: 48-64.

[11] Ibrahim, J.G., and Chen, M.-H. (2000). Power prior distributions for regression models. *Statstical Science* **15**: 46-60.

[12] Ibrahim, J. G., Chen, M.-H., and Sinha D. (2003). On optimality properties of the power prior. *Journal of the American Statistical Association* **98**: 204-213.

[13] Jeffreys, H. (1946). An invariant form for the prior probability in estimation problems, em Proceedings of the Royal Statistical Society of London, Series A, **186**, 453-461.

[14] Kullback, S. and Leiber, R.A. (1951). On information and sufficiency, *Annals of Mathematical Statistics*, **22**, 79-86.

[15] Shannon, C.E. (1948). A mathematical theory of communication. *Bell Systme Technical Journal* **27**: 379-423.

[16] Zellner, A. (1988). Optimal information processing and Bayes' theorem. *The American Statistician* **42**: 278-284.

[17] Zellner, A., and Min, C. (1993). Bayesian analysis, model selection and prediction. In *Physics and Probability: Essays in Honor of Edwin T. Jaynes*, W.J. Grandy and P. Milonmi, Eds. U.K.: Cambridge University Press, pp. 195-206.

Appendix: Proofs of Theorems

Proof of Theorem 1:

It is known that

$$\begin{aligned}\varpi(D_0 \wedge D) &\equiv E_{\pi(\delta|D,D_0)}\left\{ \ln \frac{m(D|D_0,\delta)\pi(\delta)}{\pi(\delta|D_0,D)} - \ln m(D) \right\} \\ &= \int \pi(\delta|D_0,D) \ln \frac{m(D|D_0,\delta)\pi(\delta)}{\pi(\delta|D_0,D)} d\delta - \ln m(D) \\ &= -K\left\{ \pi(\delta|D_0,D), \frac{m(D|D_0,\delta)\pi(\delta)}{M} \right\} + \ln M - \ln m(D),\end{aligned}$$

where $M = \int m(D|D_0,\delta)\pi(\delta)\, d\delta$ is the normalizing constant of $m(D|D_0,\delta)\pi(\delta)$. Now clearly $-K\left\{ \pi^*(\delta|D_0,D), \frac{m(D|D_0,\delta)\pi(\delta)}{M} \right\}$ is maximized and equal to 0 when

$$\pi^*(\delta|D_0,D) = \frac{m(D|D_0,\delta)\pi(\delta)}{M} \propto m(D|D_0,\delta)\pi(\delta).$$

Combined with (17), it leads to

$$\pi^*(\delta|D_0,D) \propto \pi(\delta) \frac{\int_\Theta L(\theta|D) L(\theta|D_0)^\delta \pi(\theta)\, d\theta}{\int_\Theta L(\theta|D_0)^\delta \pi(\theta)\, d\theta}.$$

Proof of Theorem 2:

Applying the property of the *Kullback-Leibler* divergence between two distributions,

$$K(f_1, f_2) = \int f_1(x) \ln \frac{f_1(x)}{f_2(x)} dx \geq 0,$$

with equality held if and only if $f_1(x) = f_2(x)$, we conclude that

$$\frac{n}{n_0}h_1(D_0, D, \delta) = \int_\Theta \ln L(\theta|D)\{\pi(\theta|D_0, D, \delta) - \pi(\theta|D_0, \delta)\}d\theta$$
$$= \int_\Theta \ln\left\{\frac{\pi(\theta|D_0, D, \delta)}{\pi(\theta|D_0, \delta)}M(D_0, D|\delta)\right\}\{\pi(\theta|D_0, D, \delta) - \pi(\theta|D_0, \delta)\}d\theta$$
$$= \int \ln\frac{\pi(\theta|D_0, D, \delta)}{\pi(\theta|D_0, \delta)}\pi(\theta|D_0, D, \delta)\ d\theta + \int \ln\frac{\pi(\theta|D_0, \delta)}{\pi(\theta|D_0, D, \delta)}\pi(\theta|D_0, \delta)\ d\theta \geq 0, \quad (21)$$

with equality held if and only if $\pi(\theta|D_0, D, \delta) = \pi(\theta|D_0, \delta)$. In (21), $M(D_0, D|\delta)$ is a marginal density that does not depend on θ and hence its related term is 0 since both $\pi(\theta|D_0, D, \delta)$ and $\pi(\theta|D_0, \delta)$ are proper.

In order to show that the marginal posterior mode of δ is 1, it is sufficient to show that the derivative of $\pi(\delta|D_0, D)$ in (5) is non-negative. Using certain algebra, we obtain

$$\frac{d}{d\delta}\pi(\delta|D_0, D) = \frac{d}{d\delta}\{\ln\pi(\delta)\}\pi(\delta|D_0, D)$$
$$+ \pi(\delta|D_0, D)\int_\Theta \ln L(\theta|D_0)\{\pi(\theta|D_0, D, \delta) - \pi(\theta|D_0, \delta)\}d\theta. \quad (22)$$

Since we are dealing with the exponential family with form (9) and (11), the likelihood ratio

$$\ln L(\theta|D_0) = \ln h(D_0) + n_0\{\underline{C}(D_0)'\underline{w}(\theta) + \tau(\theta)\}$$
$$= \left\{lnh(D_0) - \frac{n_0}{n}\ln h(D)\right\} + \frac{n_0}{n}\ln L(\theta|D) + n_0\{\underline{C}(D_0) - \underline{C}(D)\}'\underline{w}(\theta). \quad (23)$$

Combining (21) and (23) into (22), we prove Theorem 2 by showing the condition (19).

Proof of Theorem 3:

Suppose that k is an arbitrary positive constant. We take the likelihood function of the form $L(\theta|x) = kf(x|\theta)$, then $L(\theta|D) = k^n f(D|\theta)$ and $L(\theta|D_0) = k^{n_0}f(D_0|\theta)$. For the original power prior approach, the marginal posterior distribution of δ can be rewritten as

$$\pi(\delta|D_0, D) \propto \pi(\delta)\int L(\theta|D)L(\theta|D_0)^\delta\pi(\theta)\ d\theta$$
$$\propto \pi(\delta)\int f(D|\theta)[k^{n_0}f(D_0|\theta)]^\delta\pi(\theta)\ d\theta. \quad (24)$$

To prove that the marginal posterior mode of δ is 0, it is sufficient to show that $\frac{\partial \pi(\delta|D_0,D)}{\partial \delta} \leq 0$ for any $\delta \in [0,1]$.

The derivative of $\pi(\delta|D_0, D)$ contains two parts. The first part is the derivative on $\pi(\delta)$. If $\pi(\delta)$ is non-increasing as described in the theorem, this part is non-positive. The second part is the derivative in the integral part in (24). This part is non-positive is equivalent to

$$\begin{aligned}
&\int f(D|\theta)\frac{\partial[k^{n_0}f(D_0|\theta)]^\delta}{\partial\delta}\pi(\theta)\ d\theta \leq 0 \\
&\iff k^{n_0\delta}\int \pi(\theta)f(D|\theta)f(D_0|\theta)^\delta\{n_0\ln k + \ln f(D_0|\theta)\}d\theta \leq 0 \\
&\iff \frac{\int \pi(\theta)f(D|\theta)f(D_0|\theta)^\delta \ln f(D_0|\theta)d\theta}{\int \pi(\theta)f(D|\theta)f(D_0|\theta)^\delta d\theta} \leq n_0 \ln\frac{1}{k}, \qquad (25)
\end{aligned}$$

assuming that the derivative and integral are interchangeable.

If we take

$$k_0 = \exp\left\{-\frac{1}{n_0}\max_{0\leq\delta\leq 1}\frac{\int \pi(\theta)f(D|\theta)f(D_0|\theta)^\delta \ln f(D_0|\theta)d\theta}{\int \pi(\theta)f(D|\theta)f(D_0|\theta)^\delta d\theta}\right\} > 0,$$

then the sufficient condition in (25) for the marginal posterior mode of δ being 0 is met for any δ.