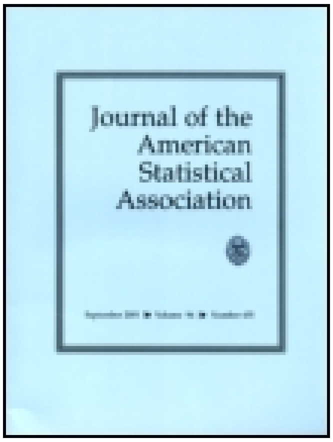

Journal of the American Statistical Association

Journal of the American Statistical Association
Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uasa20

On Optimality Properties of the Power Prior

Joseph G Ibrahim[a], Ming-Hui Chen[a] & Debajyoti Sinha[a]

[a] Joseph G. Ibrahim is Professor, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599 . Ming-Hui Chen is Associate Professor, Department of Statistics, University of Connecticut, Storrs, CT 06269 . Debajyoti Sinha is Associate Professor, Department of Biometry and Epidemiology, Medical University of South Carolina, Charleston, SC 29425 . The authors thank the editor, associate editor, and two referees for several suggestions that have greatly improved the article. Ibrahim's research was partially supported by National Institutes of Health (NIH) grants CA 70101 and CA 74015, Chen's research was partially supported by National Science Foundation grant DMS-9702172 and NIH grants CA 70101 and CA 74015, and Sinha's research was partially supported by National Cancer Institute grant R29-CA69222.
Published online: 31 Dec 2011.

To cite this article: Joseph G Ibrahim, Ming-Hui Chen & Debajyoti Sinha (2003) On Optimality Properties of the Power Prior, Journal of the American Statistical Association, 98:461, 204-213, DOI: 10.1198/016214503388619229

To link to this article: http://dx.doi.org/10.1198/016214503388619229

On Optimality Properties of the Power Prior

Joseph G. IBRAHIM, Ming-Hui CHEN, and Debajyoti SINHA

The *power prior* is a useful general class of priors that can be used for arbitrary classes of regression models, including generalized linear models, generalized linear mixed models semiparametric survival models with censored data, frailty models, multivariate models, and nonlinear models. The power prior specification for the regression coefficients focuses on observable quantities in that the elicitation is based on historical data, D_0, and a scalar quantity, a_0, quantifying the heterogeneity between the current data, D, and the historical data D_0. The power prior distribution is then constructed by raising the likelihood function of the historical data to the power a_0, where $0 \le a_0 \le 1$. The scalar a_0 is a precision parameter that can be viewed as a measure of compatibility between the historical and current data. In this article we give a formal justification of the power prior and show that it is an optimal class of informative priors in the sense that it minimizes a convex sum of Kullback–Leibler (KL) divergences between two specific posterior densities, in which one density is based on no incorporation of historical data and the other density is based on pooling the historical and current data. This result provides a strong motivation for using the power prior as an informative prior in Bayesian inference. In addition, we derive a formal relationship between this convex sum of KL divergences and the information-processing rules proposed by others. Specifically, we show that the power prior is a 100% efficient information-processing rule in the sense defined earlier. Several examples involving simulations as well as real datasets are examined to demonstrate the proposed methodology.

KEY WORDS: Generalized linear model; Gibbs sampling; Historical data; Kullback–Leibler divergence; Power prior; Prior elicitation; Proportional-hazards model; Random-effects model; Sensitivity analyses.

1. INTRODUCTION

Prior elicitation plays a crucial role in Bayesian inference. Noninformative priors are attractive in many situations in which there is lack of prior information or complete ignorance about the parameters in the model. Although noninformative and improper priors may be useful and easier to specify for certain problems, they cannot be used in all applications (e.g., model selection or model comparison), because it is well known that proper priors are required to compute Bayes factors and posterior model probabilities. Moreover, noninformative priors do not make use of real prior information that one may have for a specific application. Informative priors are often essential in these situations, and in general they are useful in settings where the investigator has access to historical data. For example, in many cancer and AIDS clinical trials, current studies often use treatments that are very similar to 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.

One way of constructing an informative prior in the presence of historical data is to write out the likelihood based on the historical data, and then raise the likelihood to a suitable power to discount the historical data relative to the current data. The idea of raising a likelihood to a power to construct a prior has been discussed by several authors in various contexts and applications. One of the first was Zellner (1988, 1997, 2002), who proposed the idea of raising a likelihood to a power in the context of information processing and optimal information processing rules. Ibrahim and Chen (2000) called such priors *power priors*, and used a different formulation than that of Zellner to motivate them.

We caution here, however, that one should not use historical data blindly or in a semiautomatic fashion when constructing informative priors, because the information contained in the historical data may be inappropriate for the research problem at hand. Thus extreme care must be taken when constructing informative priors.

The power prior provides a useful class of informative priors for Bayesian inference. To fix ideas, suppose that we have historical data from a similar previous study, denoted by $D_0 = (n_0, y_0, X_0)$, where n_0 is the sample size of the historical data, y_0 is the $n_0 \times 1$ response vector, and X_0 is the $n_0 \times p$ matrix of covariates based on the historical data. The power prior is defined as the likelihood function based on the historical data, D_0, raised to a power a_0, where $0 \le a_0 \le 1$ is a scalar parameter that controls the influence of the historical data on the current data. One of the most useful applications of the power prior is in model selection problems, because these priors inherently automate the informative prior specification for all possible models in the model space. They are quite attractive in this context, because specifying meaningful informative prior distributions for the parameters in each model is a difficult task, requiring contextual interpretations of a large number of parameters. In variable subset selection, for example, the prior distributions for all possible subset models are automatically determined once the historical data D_0 and a_0 are specified.

Three issues immediately arise when using the power prior. The first issue concerns propriety of the power prior under various models; the second concerns the choice of a_0; the third issue concerns the theoretical justification of the power

Joseph G. Ibrahim is Professor, Department of Biostatistics, University of North Carolina, Chapel Hill, NC 27599 (E-mail: *ibrahim@bios.unc.edu*). Ming-Hui Chen is Associate Professor, Department of Statistics, University of Connecticut, Storrs, CT 06269 (E-mail: *mhchen@stat.uconn.edu*). Debajyoti Sinha is Associate Professor, Department of Biometry and Epidemiology, Medical University of South Carolina, Charleston, SC 29425 (E-mail: *sinhad@musc.edu*). The authors thank the editor, associate editor, and two referees for several suggestions that have greatly improved the article. Ibrahim's research was partially supported by National Institutes of Health (NIH) grants CA 70101 and CA 74015, Chen's research was partially supported by National Science Foundation grant DMS-9702172 and NIH grants CA 70101 and CA 74015, and Sinha's research was partially supported by National Cancer Institute grant R29-CA69222.

Journal of the American Statistical Association
March 2003, Vol. 98, No. 461, Theory and Methods
DOI 10.1198/016214503388619229

prior. For the first issue, Ibrahim and Chen (2000) and Chen, Ibrahim, and Shao (2000) showed that the power prior is proper for a wide variety of models under some very general conditions that typically are easily satisfied. For the second issue, one might argue that $a_0 = 1$ should be used so that a coherent Bayesian updating scheme is followed when constructing priors. That is, the posterior of the parameters based on the historical data serves as the prior of the parameters for the current experiment. Using $a_0 = 1$ results in equal weighting of the historical and current data, and thus implies a pooling of D and D_0 given the parameters. This is a common and perhaps more traditional approach to incorporating historical data. Such pooling might not be desirable in many applications when the investigator wants to discount the historical data relative to the current data; in these cases, a choice of a_0 less than 1 is desirable.

For the third issue, we provide in this article a formal justification of the power prior as an optimal class of priors for Bayesian inference when historical data are available and show that it is closely related the optimal information processing rules of Zellner (1988, 2002). Specifically, we obtain an important result showing that the power prior is indeed optimal in the sense that it minimizes a convex sum of Kullback–Leibler (KL) divergences between posterior densities based on pooling the historical data ($a_0 = 1$) and a posterior density not using the historical data ($a_0 = 0$). This KL divergence is closely linked to the optimal information processing rules of Zellner (1988). In particular, following Zellner (1988), we show that the power prior is 100% efficient in the sense that the ratio of the output to input information is equal to 1.

The rest of the article is organized as follows. We review the power prior in Section 2, and give the first theorem for the optimality result in Section 2.1. In Section 2.2 we discuss relationships of the power prior to other priors, such as those considered by Spiegelhalter, Freedman, and Parmar (1994), and in Section 2.3 we extend our results to multiple historical datasets. In Section 3 we consider connections with Zellner's work on optimal information processing. In Section 4 we present several numerical examples involving simulated and real datasets that illustrate the properties of the power prior under various scenarios. We conclude the article with a brief discussion in Section 5.

2. THE POWER PRIOR

We consider the power prior for an arbitrary regression model. Let the data from the current study be denoted by $D = (n, y, X)$, where n denotes the sample size, y denotes the $n \times 1$ response vector, and X denotes the $n \times p$ matrix of covariates. Further, denote the likelihood for the current study by $L(\theta|D)$, where θ is a vector of indexing parameters. Thus $L(\theta|D)$ is a general likelihood function for an arbitrary regression model, such as a generalized linear model, random-effects model, nonlinear model, or a semiparametric survival model with censored data. Now suppose we that have historical data from a similar previous study, denoted by $D_0 = (n_0, y_0, X_0)$. Further, let $\pi_0(\theta)$ denote the prior distribution for θ before the historical data, D_0, is observed. We call $\pi_0(\theta)$ the *initial prior* distribution for θ. Given a_0, we define the *power prior* distribution of θ for the current study as

$$\pi(\theta|D_0, a_0) \propto L(\theta|D_0)^{a_0} \pi_0(\theta), \tag{1}$$

where a_0 is a scalar prior parameter that weights the historical data relative to the likelihood of the current study. In Section 3 we justify (1) an optimal prior in the sense that is yields a 100% information processing rule in the sense defined by Zellner (1988). The parameter a_0 controls the influence of the historical data on $\pi(\theta|D_0, a_0)$. The parameter a_0 can be interpreted as a precision parameter for the historical data. For example in the normal linear model (1) implies that a priori, $\theta \sim N((X_0'X_0)^{-1}X_0'y_0, a_0^{-1}(X_0'X_0)^{-1})$, so that a_0 is the precision parameter. For models other than the normal model, it can be shown that as $n_0 \to \infty$, a_0 is a precision parameter for θ (see Ibrahim, Ryan, and Chen 1998). It is reasonable to restrict the range of a_0 to be between 0 and 1, and thus we take $0 \le a_0 \le 1$. One of the main roles of a_0 is to control the heaviness of the tails of the prior for θ. As a_0 becomes smaller, the tails of (1) become heavier—that is, the prior variance of θ becomes larger. Setting $a_0 = 1$, (1) corresponds to the update of $\pi_0(\theta)$ using Bayes's theorem. That is, with $a_0 = 1$, (1) corresponds to the posterior distribution of θ based on the historical data. When $a_0 = 0$, then the prior does not depend on the historical data D_0, and in this case $\pi(\theta|D_0, a_0 = 0) \equiv \pi_0(\theta)$. Thus $a_0 = 0$ is equivalent to a prior specification with no incorporation of historical data. Therefore, (1) can be viewed as a generalization of the usual Bayesian update of $\pi_0(\theta)$. The parameter a_0 allows the investigator to control the influence of the historical data on the current study. Such control is important in cases where there is heterogeneity between the previous and current study, or when the sample sizes of the two studies are quite different. Finally, we mention that we may not want to restrict a_0 to be less than or equal to 1. In this case, the only requirement is that $a_0 \ge 0$. Then in this situation, we see that as $a_0 \to \infty$, (1) becomes a degenerate prior, with a point mass.

Under the power prior, the posterior distribution of θ can be written as

$$\pi(\theta|D, D_0, a_0) \propto L(\theta|D)L(\theta|D_0)^{a_0} \pi_0(\theta). \tag{2}$$

We mention here that (2) might suggest that a_0 is a Box–Cox–type transformation. The Box–Cox transformation is typically defined as a transformation on the response variable or the covariates, whereas a_0 in (2) is a transformation on an entire likelihood and thus is not a Box–Cox transformation in the usual sense. Moreover, there typically will be very little information in D and D_0 for estimating a_0, so estimation of a_0 via maximum likelihood likely will not be fruitful.

There are two interesting special cases of (2). These special cases are at the extremes $a_0 = 0$ and $a_0 = 1$. First, when $a_0 = 0$, this leads to the posterior

$$\pi(\theta|D, D_0, a_0 = 0) \propto L(\theta|D)\pi_0(\theta). \tag{3}$$

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

$$\pi(\theta|D, D_0, a_0 = 1) \propto L(\theta|D)L(\theta|D_0)\pi_0(\theta). \tag{4}$$

Thus (3) and (4) represent the two extremes. In one case, no historical data are used, in the other case, the historical and current data are equally weighted, and thus (4) corresponds to pooling the historical and current data. We mention here that Min and Zellner (1993) considered issues of how to combine datasets in the contexts of forecasts in econometrics. Although their methodology and applications are quite different than those that we present here, they have presented some similar ideas in the context of normal linear models.

2.1 Optimality Result

We now give a formal justification of the class of power priors. Here we assume that the power parameter a_0 is fixed; we consider a_0 random in Section 2.4. The power prior can be justified as the minimizer of the convex sum of the KL divergences between the posterior densities given in (3) and (4). Toward this goal, recall the definition of the KL divergence. Suppose that f_0 and f_1 are two densities with respect to Lebesgue measure. Then the KL-directed divergence between f_0 and f_1 is defined as

$$K(f_0, f_1) = \int \log\left(\frac{f_0(\theta)}{f_1(\theta)}\right) f_0(\theta)\, d\theta. \tag{5}$$

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

$$K_g = (1 - a_0)K(g, f_0) + a_0 K(g, f_1), \tag{6}$$

where $0 \le a_0 \le 1$. It turns out that the density $g \equiv g(\theta)$ that minimizes K_g, denoted by g_{opt}, is $g_{opt} = \pi(\theta|D, D_0, a_0) \propto L(\theta|D)L(\theta|D_0)^{a_0}\pi_0(\theta)$. This tells us that the power prior is the unique prior that minimizes (6). We state this as a formal theorem.

Theorem 1. Let f_0 denote the density in (3) and f_1 denote the density in (4). The density $g \equiv g(\theta)$ that minimizes

$$K_g = (1 - a_0)K(g, f_0) + a_0 K(g, f_1) \tag{7}$$

is

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

Proof. We have

$$\begin{aligned}
(1 - a_0)&K(g, f_0) + a_0 K(g, f_1)\\
&= (1 - a_0)\int \log(g(\theta)/f_0(\theta))g(\theta)\, d\theta\\
&\quad + a_0 \int \log(g(\theta)/f_1(\theta))g(\theta)\, d\theta\\
&= \int \log[g(\theta)/f_0(\theta)]^{1-a_0} g(\theta)\, d\theta\\
&\quad + \int \log[g(\theta)/f_1(\theta)]^{a_0} g(\theta)\, d\theta\\
&= \int \log[g(\theta)/(f_0(\theta)^{1-a_0} f_1(\theta)^{a_0})]g(\theta)\, d\theta\\
&= K\left(g, \frac{f_0^{1-a_0} f_1^{a_0}}{h(a_0)}\right) - \log(h(a_0)),
\end{aligned} \tag{8}$$

where $h(a_0) = \int f_0(\theta)^{1-a_0} f_1(\theta)^{a_0}\, d\theta$ is the normalizing constant of $f_0^{1-a_0} f_1^{a_0}$. Now clearly $K(g, \frac{f_0^{1-a_0} f_1^{a_0}}{h(a_0)})$ is minimized and equal to 0 when

$$g = g_{opt} = \frac{f_0^{1-a_0} f_1^{a_0}}{h(a_0)} \propto f_0^{1-a_0} f_1^{a_0}.$$

Because f_0 is (3) and f_1 is (4), we have

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

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 in this sense the power prior is 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 $g \propto (f_0 f_1)^{1/2} \propto L(\theta|D)L(\theta|D_0)^{1/2}\pi_0(\theta)$. 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 $a_0 = .5$. This result tells us that a choice of $a_0 = .5$ is a reasonable starting value to use in an analysis and on which to base sensitivity analyses.

2.2 Properties of K_g

Recall that

$$K_g = (1 - a_0)K(g, f_0) + a_0 K(g, f_1). \tag{9}$$

If we view K_g as a function of g, then it is easy to see that K_g is nonnegative, because each of $K(g, f_0)$ and $K(g, f_1)$ is a KL-directed divergence. Furthermore, we can show that K_g is convex in g. We formally state this result in the following theorem.

Theorem 2. Assume that g_1 and g_2 are any proper densities, that is,

$$\int g_1(\theta)\, d\theta = \int g_2(\theta)\, d\theta = 1.$$

Then, for $0 \le \alpha \le 1$, we have

$$K_{\alpha g_1 + (1-\alpha)g_2} \le \alpha K_{g_1} + (1 - \alpha)K_{g_2}. \tag{10}$$

Proof. After some algebra, K_g in (9) can be written as

$$\begin{aligned}
K_g = \int g(\theta)\log g(\theta)\, d\theta - (1 - a_0)\int g(\theta)\log f_0(\theta)\, d\theta&\\
- a_0 \int g(\theta)\log f_1(\theta)\, d\theta.&
\end{aligned} \tag{11}$$

Using (11), we have

$$\begin{aligned}
K_{\alpha g_1 + (1-\alpha)g_2} = \int &\{\alpha g_1(\theta) + (1 - \alpha)g_2(\theta)\}\\
&\times \log[\alpha g_1(\theta) + (1 - \alpha)g_2(\theta)]\, d\theta
\end{aligned}$$

$$\begin{aligned}
&- (1-a_0)\int\{\alpha g_1(\theta)+(1-\alpha)g_2(\theta)\}\\
&\quad\times \log f_0(\theta)\,d\theta\\
&- a_0\int\{\alpha g_1(\theta)+(1-\alpha)g_2(\theta)\}\log f_1(\theta)\,d\theta\\
={}&\int\{\alpha g_1(\theta)+(1-\alpha)g_2(\theta)\}\\
&\quad\times\log[\alpha g_1(\theta)+(1-\alpha)g_2(\theta)]\,d\theta\\
&-\alpha\bigg[(1-a_0)\int g_1(\theta)\log f_0(\theta)\,d\theta\\
&\qquad +a_0\int g_1(\theta)\log f_1(\theta)\,d\theta\bigg]\\
&-(1-\alpha)\bigg[(1-a_0)\int g_1(\theta)\log f_0(\theta)\,d\theta\\
&\qquad +a_0\int g_1(\theta)\log f_1(\theta)\,d\theta\bigg]. \qquad (12)
\end{aligned}$$

Define $u(t)=t\log t$ for $t>0$. It can be shown that $u''(t)=1/t>0$ for $t>0$. Thus $u(t)$ is convex in t for $t>0$. This result leads directly to

$$\begin{aligned}
&\{\alpha g_1(\theta)+(1-\alpha)g_2(\theta)\}\log[\alpha g_1(\theta)+(1-\alpha)g_2(\theta)]\\
&\qquad\le \alpha g_1(\theta)\log g_1(\theta)+(1-\alpha)g_2(\theta)\log g_2(\theta). \qquad (13)
\end{aligned}$$

Combining (12) and (13) yields

$$\begin{aligned}
K_{\alpha g_1+(1-\alpha)g_2} \le{}& \alpha\int g_1(\theta)\log g_1(\theta)\,d\theta\\
&+(1-\alpha)\int g_2(\theta)\log g_2(\theta)\,d\theta\\
&-\alpha\bigg[(1-a_0)\int g_1(\theta)\log f_0(\theta)\,d\theta\\
&\qquad+a_0\int g_1(\theta)\log f_1(\theta)\,d\theta\bigg]\\
&-(1-\alpha)\bigg[(1-a_0)\int g_1(\theta)\log f_0(\theta)\,d\theta\\
&\qquad+a_0\int g_1(\theta)\log f_1(\theta)\,d\theta\bigg]\\
={}&\alpha K_{g_1}+(1-\alpha)K_{g_2},
\end{aligned}$$

which proves the theorem.

Remark 1. Theorem 2 directly implies that in terms of a function of g, (a) the minimum of K_g exists and (b) K_g has a unique minimizer.

Soofi and Retzer (2002) have provided an excellent overview on information theoretic measures, their properties, and their relationships. We now give several examples that yield closed-form expressions for g_{opt}.

Example 1: Normal Linear Model. Consider historical data $y_0=X_0\beta+\epsilon_0$, where $\epsilon_0\sim N_n(0,\sigma^2 I)$, X_0 is $n_0\times p$ of rank p, and β is $p\times 1$. We assume here that σ^2 is known and that the initial prior is $\pi_0(\beta)\propto 1$. In this case the power prior is given by

$$\begin{aligned}
\pi(\beta|D_0,a_0)&\propto\exp\left\{-\frac{a_0}{2\sigma^2}(y_0-X_0\beta)'(y_0-X_0\beta)\right\}\\
&\propto\exp\left\{-\frac{a_0}{2\sigma^2}(\beta-\hat\beta_0)'(X_0'X_0)(\beta-\hat\beta_0)\right\},
\end{aligned}$$

where $\hat\beta_0=(X_0'X_0)^{-1}X_0'y_0$. Thus we see in this case that

$$\pi(\beta|D_0,a_0)=N_p(\hat\beta_0,a_0^{-1}\sigma^2(X_0'X_0)^{-1}).$$

Also, consider the current data, which follow the linear model $y=X\beta+\epsilon$, where $\epsilon\sim N_n(0,\sigma^2 I)$. Without loss of generality, and for ease of exposition, suppose that $\sigma^2=1$. Let $D=(n,y,X)$ and $D_0=(n_0,y_0,X_0)$. It is easily shown that g_{opt} is a multivariate normal distribution $N_p(\mu,\Sigma)$, where

$$\mu=\Lambda\hat\beta_0+(I-\Lambda)\hat\beta, \qquad (14)$$

$$\begin{aligned}
\hat\beta=(X'X)^{-1}X'y,\ \Lambda=(X'X+a_0X_0'X_0)^{-1}(a_0X_0'X_0),\quad\text{and}\\
\Sigma=\left(X'X+a_0X_0'X_0\right)^{-1}. \qquad (15)
\end{aligned}$$

From this, it follows that $\pi(\beta|D,D_0,a_0=0)=N(\mu_0,\Sigma_0)$, where $\mu_0=\hat\beta$ and $\Sigma_0=(X'X)^{-1}$. Also, $\pi(\beta|D,D_0,a_0=1)=N(\mu_1,\Sigma_1)$, where $\Lambda_1=(X'X+X_0'X_0)^{-1}(X_0'X_0)$, $\mu_1=\Lambda_1\hat\beta_0+(I-\Lambda_1)\hat\beta$, and $\Sigma_1=(X'X+X_0'X_0)^{-1}$. We note here that if σ^2 priors is assumed to be unknown and we specify an inverse gamma prior for it, then a straightforward derivation shows that $g_{opt}(\beta)$, the marginal posterior distribution of β, is a multivariate t distribution.

Example 2: Exponential Model. Suppose that the current data y_i have an exponential distribution with mean $1/\theta$, $i=1,\ldots,n$, and that the y_i's are iid. Let the historical data y_{0i} have the same distribution as y_i, and let $y_0=(y_{01},\ldots,y_{0n_0})'$ and $y=(y_1,\ldots,y_n)'$. Further, let $D=(n,y)$ and $D_0=(n_0,y_0)$. Take $\pi_0(\theta)\propto\theta^{-1}$. Then $\pi(\theta|D,D_0,a_0=1)=\mathcal{G}(n+n_0,n\bar y+n_0\bar y_0)$, where $\mathcal{G}(a,b)$ denotes the gamma distribution with shape parameter a and scale parameter b, $\bar y_0=\frac{1}{n_0}\sum_{i=1}^{n_0}y_{0i}$ and $\bar y=\frac{1}{n}\sum_{i=1}^{n}y_i$. Also, $\pi(\theta|D,D_0,a_0=0)=\mathcal{G}(n,n\bar y)$. A straightforward derivation yields that g_{opt} is $\mathcal{G}(n+a_0n_0,n\bar y+a_0n_0\bar y_0)$. Similar formulas are obtained when we have right-censored data for this model.

Example 3: Poisson Model. Suppose that the current data y_i has a Poisson distribution with mean θ, $i=1,\ldots,n$, and the y_i's are iid. Let the historical data y_{0i} have the same distribution as y_i, and let $y_0=(y_{01},\ldots,y_{0n_0})'$ and $y=(y_1,\ldots,y_n)'$. Further, let $D=(n,y)$ and $D_0=(n_0,y_0)$. Take $\pi_0(\theta)\propto\theta^{-1}$. Then $\pi(\theta|D,D_0,a_0=1)=\mathcal{G}(n\bar y+n_0\bar y_0,n+n_0)$, and $\pi(\theta|D,D_0,a_0=0)=\mathcal{G}(n\bar y,n)$. A straightforward derivation yields that g_{opt} is $\mathcal{G}(n\bar y+a_0n_0\bar y_0,n+a_0n_0)$.

2.3 Relationships With Other Priors

The power prior is related to the class of priors discussed by Spiegelhalter et al. (1994), that are motivated from applications in clinical trials. Spiegelhalter et al. discussed a prior for θ of the form $\theta \sim N(\theta_0, \sigma^2/n_0)$ for the normal model. The power prior is related to this prior specification. Suppose that we have historical data $y_0 = (y_{01}, \ldots, y_{0n_0})$, and the y_{0i}'s are iid. $N(\theta, \sigma^2)$. Assuming that the initial prior for θ is $\pi_0(\theta) \propto 1$, the power prior for θ is $\pi(\theta|D_0, a_0) \propto L(\theta|D_0)^{a_0} \propto \exp\{-\frac{n_0 a_0}{2\sigma^2}(\bar{y}_0 - \theta)^2\}$, where $\bar{y}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} y_{0i}$. So the power prior for θ is $N(\bar{y}_0, a_0^{-1}(\sigma^2/n_0))$. Thus if $\theta_0 = \bar{y}_0$ and $a_0 = 1$, then the power prior and the Spiegelhalter et al. prior are identical. The main difference between the two priors is in the prior variance. The power prior has the extra parameter a_0, which gives the investigator more control over the impact of the historical data. Thus the power prior has an extra multiple a_0^{-1} for σ^2/n_0, and we see that a_0 plays the role of a precision parameter in the power prior. Further, we see that the Spiegelhalter et al. prior is in fact a special case of the power prior.

2.4 Extension to Multiple Historical Datasets

The power prior defined in (1) can be easily generalized to multiple historical datasets. If there are L_0 historical studies, then we define $D_{0k} = (n_{0k}, X_{0k}, y_{0k})$ to be the historical data based on the kth study, $j = 1, \ldots, L_0$, and $D_0 = (D_{01}, \ldots, D_{0L_0})$. Letting $a_0 = (a_{01}, \ldots, a_{0L_0})$, the prior in (1) can be generalized as

$$\pi(\theta, a_0|D_0) \propto \left(\prod_{k=1}^{L_0} [L(\theta|D_{0k})]^{a_{0k}}\right)\pi_0(\theta). \tag{16}$$

Under the power prior, the posterior distribution of θ can be written as

$$\pi(\theta|D, D_0, a_0) \propto L(\theta|D)\left(\prod_{k=1}^{L_0} [L(\theta|D_{0k})]^{a_{0k}}\right)\pi_0(\theta), \tag{17}$$

where $L(\theta|D)$ denotes the likelihood function of θ given the current data D. There are $(L_0 + 1)$ interesting special cases of (17). These special cases are at the extremes $a_0 = (0, 0, \ldots, 0)$ and

$$a_0 = e_k \equiv (a_{01} = 0, \ldots, a_{0,k-1} = 0,$$
$$a_{0k} = 1, a_{0,k+1} = 0, \ldots, a_{0L_0} = 0),$$

where e_k is a vector with a 1 in the kth position and 0s elsewhere, for $k = 1, 2, \ldots, L_0$. First, when $a_0 = (0, 0, \ldots, 0) \equiv 0$ this leads to the posterior

$$\pi(\theta|D, D_0, a_0 = 0) \propto L(\theta|D)\pi_0(\theta). \tag{18}$$

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

$$\pi(\theta|D, D_0, a_0 = e_k) \propto L(\theta|D)L(\theta|D_{0k})\pi_0(\theta), \tag{19}$$

for $k = 1, 2, \ldots, L_0$. Equations (18) and (19) represent the $(L_0 + 1)$ extremes. In (18) no historical data are used, and in (19) only the kth historical dataset is used for $k = 1, 2, \ldots, L_0$.

Let $g(\theta)$ denote an arbitrary density function of θ. Also, let

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

and

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

for $k = 1, 2, \ldots, L_0$. Now we define

$$K_{mg} = \left(1 - \sum_{k=1}^{L_0} a_{0k}\right)K(g, f_0) + \sum_{k=1}^{L_0} a_{0k}K(g, f_k). \tag{20}$$

We are led to the following theorem.

Theorem 3. Assume that $a_{0k} \geq 0$ for $k = 1, 2, \ldots, L_0$ and $\sum_{k=1}^{L_0} a_{0k} \leq 1$. The density that minimizes K_{mg} defined by (20) is

$$g_{m,\,opt} = \pi(\theta|D, D_0, a_0) \propto L(\theta|D)\left(\prod_{k=1}^{L_0} [L(\theta|D_{0k})]^{a_{0k}}\right)\pi_0(\theta).$$

Proof. We can write

$$\begin{aligned} K_{mg} &= \left(1 - \sum_{k=1}^{L_0} a_{0k}\right)K(g, f_0) + \sum_{k=1}^{L_0} a_{0k}K(g, f_k) \\ &= \int \log\left(\frac{g^{1-\sum_{k=1}^{L_0} a_{0k}}}{f_0^{1-\sum_{k=1}^{L_0} a_{0k}}} \cdot \prod_{k=1}^{L_0} \frac{g^{a_{0k}}}{f_k^{a_{0k}}}\right) g\, d\theta \\ &= \int \log\left(\frac{g}{f_0^{1-\sum_{k=1}^{L_0} a_{0k}} \prod_{k=1}^{L_0} f_k^{a_{0k}}}\right) g\, d\theta \\ &= K\left(g, \frac{f_0^{1-\sum_{k=1}^{L_0} a_{0k}} \prod_{k=1}^{L_0} f_k^{a_{0l}}}{h_m(a_0)}\right) - \log(h_m(a_0)), \end{aligned}$$

where $h_m(a_0) = \int f_0^{1-\sum_{k=1}^{L_0} a_{0k}} \prod_{k=1}^{L_0} f_k^{a_{0k}}\, d\theta$. Similar to the proof of Theorem 1, K_{mg} is minimized and equal to $-\log(h_m(a_0))$ when

$$g = g_{m,\,opt} \propto L(\theta|D)\left(\prod_{k=1}^{L_0} [L(\theta|D_{0k})]^{a_{0k}}\right)\pi_0(\theta),$$

as desired.

Finally, we mention here that Theorems 1–3 can be extended to the case in which a_0 is random and has a beta distribution for the single historical dataset case and a Dirichlet distribution for the multiple historical dataset case. Specifically, when a_0 is random with these priors, it can be shown that the power prior minimizes $E(K_g)$ or $E(K_{mg})$, where the expectation is taken with respect to the prior of a_0.

3. RELATIONSHIP TO INFORMATION PROCESSING

The optimality result is related to optimal information processing as discussed by Zellner (1988). Zellner discussed formulations similar to the power prior discussed here, and examined weighting the prior as well as the likelihood by a power parameter. Zellner's development is follows: He defined the information-processing rule,

$$\begin{aligned}\Delta[g(\theta)] &= \text{output information} - \text{input information}\\ &= \int g(\theta)\log(g(\theta))\,d\theta + \int g(\theta)\log(m)\,d\theta\\ &\quad - \int g(\theta)\log(\pi(\theta))\,d\theta - \int g(\theta)\log(L(\theta))\,d\theta\\ &\equiv \int g(\theta)\log\left\{\frac{g(\theta)}{L(\theta)\pi(\theta)/m}\right\}d\theta, \qquad (21)\end{aligned}$$

where $g(\theta)$ is proper probability density function, $m = \int L(\theta)\pi(\theta)\,d\theta$ is the marginal density of the data, $L(\theta)$ is the likelihood function of θ, and $\pi(\theta)$ is the prior distribution for θ. Zellner defined a rule to be 100% efficient if the g^* that minimizes (21) yields output information = input information; that is, the ratio of output to input information is 1. Because the KL divergence is always nonnegative, it is clear from (21) that the function that minimizes (21) is

$$g^*(\theta) = \frac{L(\theta)\pi(\theta)}{m}, \qquad (22)$$

and thus g^* yields a 100% efficient information-processing rule.

To show that the power prior results in a 100% efficient processing rule, we consider a weighted version of (21) as in Zellner (1997, 2002). Let $m(D, D_0) = \int L(\theta|D_0)L(\theta|D)\pi_0(\theta)\,d\theta$ and $m(D) = \int L(\theta|D)\pi_0(\theta)\,d\theta$. Thus the weighted version of (21) is given by

$$\begin{aligned}\Delta[g(\theta)] &= \int g(\theta)\log(g(\theta))\,d\theta + \int g(\theta)\log(m(D, D_0))\,d\theta\\ &\quad - w_0\int g(\theta)\log(\pi_0(\theta))\,d\theta\\ &\quad - w_D\int g(\theta)\log(L(\theta|D))\,d\theta\\ &\quad - w_{D_0}\int g(\theta)\log(L(\theta|D_0))\,d\theta, \qquad (23)\end{aligned}$$

where $0 \le w_D \le 1$ and $0 \le w_{D_0} \le 1$. In our scenario, it is desirable to choose $w_0 = w_D = 1$ and $w_{D_0} = a_0$, because we would like to give a unit weight to both the current data and the initial prior $\pi_0(\theta)$ and give weight a_0 $(0 \le a_0 \le 1)$ to the historical data. We are now led to the following theorem, which directly relates (23) to K_g in (6).

Theorem 4. If we choose $w_0 = w_D = 1$ and $w_{D_0} = a_0$, then

$$K_g = \Delta[g(\theta)] + C, \qquad (24)$$

where C is a constant free of g, given by $C = (1 - a_0)(\log(m(D)) - \log(m(D, D_0)))$.

Proof. Because $m(D, D_0)$ is a constant free of θ and $\int g(\theta)\,d\theta = 1$, we have

$$\int g(\theta)\log(m(D, D_0))\,d\theta = \log(m(D, D_0)).$$

Thus we can write

$$\begin{aligned}\Delta[g(\theta)] &= \int g(\theta)\log(g(\theta))\,d\theta + \log(m(D, D_0))\\ &\quad - w_0\int g(\theta)\log(\pi_0(\theta))\,d\theta\\ &\quad - w_D\int g(\theta)\log(L(\theta|D))\,d\theta\\ &\quad - w_{D_0}\int g(\theta)\log(L(\theta|D_0))\,d\theta.\end{aligned}$$

Now we can write K_g in (6) as

$$\begin{aligned}K_g &= (1 - a_0)\int g(\theta)\log\left(\frac{g(\theta)m(D)}{L(\theta)\pi_0(\theta)}\right)d\theta\\ &\quad + a_0\int g(\theta)\log\left(\frac{g(\theta)m(D, D_0)}{L(\theta)L(\theta|D_0)\pi_0(\theta)}\right)\\ &= \int g(\theta)\log(g(\theta))\,d\theta + (1 - a_0)\log(m(D))\\ &\quad + a_0\log(m(D, D_0)) - \int g(\theta)\log(L(\theta|D))\,d\theta\\ &\quad - a_0\int g(\theta)\log(L(\theta|D_0))\,d\theta - \int g(\theta)\log(\pi_0(\theta))\,d\theta.\end{aligned}$$

Now letting $w_0 = w_D = 1$ and $w_{D_0} = a_0$ yields (24).

Remark 2. Equation (24) implies that K_g and $\Delta(g)$ have the same minimizer, and thus the two criteria are equivalent.

Remark 3. Theorem 4 implies that the power prior yields a 100% efficient information processing rule.

The relationship between K_g and $\Delta(g)$ can also be made in the case of multiple historical datasets. If we have several historical datasets, $D_{01}, \ldots, D_{0L_0}$, and an initial prior, π_0, then we need to consider the $L_0 + 1$ possible sets of input information, $(D, \pi_0), (D, D_{01}, \pi_0), \ldots, (D, D_{0L_0}, \pi_0)$. In this situation, we are not sure which (if any) of these L_0 historical datasets are compatible with the current data D to be included in the input. Thus in this case we are interested in finding an output density $g(\theta)$ such that it minimizes a convex combination of $L_0 + 1$ possible information losses, given by

$$K_{mg} = \left(1 - \sum_{k=1}^{L_0} a_{0k}\right)K(g, f_0) + \sum_{k=1}^{L_0} a_{0k}K(g, f_k). \qquad (25)$$

Thus in the case of multiple historical datasets, we can write

$$\begin{aligned}K_{mg} &= \left(1 - \sum_{k=1}^{L_0} a_{0k}\right)\int g(\theta)\log\left(\frac{g(\theta)m(D)}{L(\theta|D)\pi_0(\theta)}\right)d\theta\\ &\quad + \sum_{k=1}^{L_0} a_{0k}\int g(\theta)\log\left(\frac{g(\theta)m(D, D_{0k})}{L(\theta|D)L(\theta|D_{0k})\pi_0(\theta)}\right)d\theta\\ &= \int g(\theta)\log(g(\theta))\,d\theta + \left(1 - \sum_{k=1}^{L_0} a_{0k}\log(m(D))\right)\\ &\quad + \sum_{k=1}^{L_0} a_{0k}\log(m(D, D_{0k}))\\ &\quad - \int g(\theta)\log(\pi_0(\theta))\,d\theta - \int g(\theta)\log(L(\theta|D))\,d\theta\\ &\quad - \sum_{k=1}^{L_0} a_{0k}\int g(\theta)\log(L(\theta|D_{0k}))\,d\theta. \qquad (26)\end{aligned}$$

Now we can extend Zellner's $\Delta(g)$ to the multiple historical dataset case as

$$\begin{aligned}\Delta[g(\theta)] = &\int g(\theta)\log(g(\theta))\,d\theta \\ &+\log(m(D, D_{01}, D_{02}, \ldots, D_{0,L_0})) \\ &- w_0\int g(\theta)\log(\pi_0(\theta))\,d\theta \\ &- w_D\int g(\theta)\log(L(\theta|D_0))\,d\theta \\ &- \sum_{k=1}^{L_0} w_{D_{0k}}\int g(\theta)\log(L(\theta|D_{0k}))\,d\theta, \qquad (27)\end{aligned}$$

where $m(D, D_{0k}) = \int L(\theta|D_{0k})L(\theta|D_{0k})\pi_0(\theta)\,d\theta$ and

$$m(D, D_{01}, \ldots, D_{0k}) = \prod_{k=1}^{L_0}\int L(\theta|D_{0k})\pi_0(\theta)\,d\theta.$$

Now taking $w_0 = w_D = 1$ and $w_{D_{0k}} = a_{0k}$, we have the relation

$$K_{mg} = \Delta[g(\theta)] + C, \qquad (28)$$

where $C = (1 - \sum_{k=1}^{L_0} a_{0k})\log(m(D)) + \sum_{k=1}^{L_0} a_{0k}\log(m(D, D_{0k})) - \log(m(D, D_{01}, \ldots, D_{0,L_0}))$.

4. ILLUSTRATIVE EXAMPLES

We consider several examples to illustrate the power prior and the choice of a_0. To help facilitate the choice of a_0, we determine a *guide value* for a_0, denoted as a_g, by minimizing a penalized likelihood-type criterion taking the form

$$G(a_0) = -2\log(h^*(a_0)) + \frac{\log(n_0)}{a_0}, \qquad (29)$$

where $h^*(a_0) = \int L(\theta|D)L(\theta|D_0)^{a_0}\pi_0(\theta)\,d\theta$. Thus the minimizer to (29) is a_g. The criterion (29) has nice properties and appears to work quite well in practice for determining a_g. We emphasize here that the guide value derived from (29) serves only as a starting point for the analysis, and we recommend doing several analyses in the range of the guide value. We emphasize that we do not recommend doing an analysis based on a single a_0 value, nor do we propose specifying a_0 by a one-time automated procedure.

Example 4: Normal Models

We consider an example to demonstrate the behavior of a_0 as the historical data D_0 and the current data D are varied. Consider iid normally distributed observations in which $y_i \sim N(\theta, 1)$, and $y_{i0} \sim N(\theta, 1)$.

Tables 1 and 2 show a_g derived from (29) based on several different values of the current data sample size n, the historical data sample size n_0, $\bar{y} = \frac{1}{n}\sum_{i=1}^n y_i$, $\bar{y}_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} y_{0i}$, $s^2 = \frac{1}{n-1}\sum_{i=1}^n (y_i - \bar{y})^2$, and $s_0^2 = \frac{1}{n_0-1}\sum_{i=1}^{n_0}(y_{0i} - \bar{y}_0)^2$. The tables reveal many interesting features regarding the behavior of a_g as n, n_0, $\bar{y}_0$, $\bar{y}$, s^2, and s_0^2 are varied. First, we observe that as n is varied, a fairly constant moderate weight of $a_g = .127$ is obtained when the sample means and standard deviations are the same between historical and current datasets and $n_0 = 100$. When we take $s_0^2 = .5$ for the historical data as in Table 1, we see that a_g increases slightly. Table 1 further shows the behavior of a_g as n_0 increases and n is held fixed. We see that in this case larger values of a_g are obtained when n_0 is made smaller, and that in general a_g decreases as n_0 increases. This is yet another feature of the conservative nature of the criterion (29) for obtaining a_g. From Table 1, we see that the historical data are particularly downweighted when the historical data sample size n_0 is larger than the current data sample size n. Thus the criterion for obtaining a_g always ensures that the historical data can never dominate the posterior or have more influence than the current data. This feature is further illustrated in Figure 1, which shows a plot of the criterion function (29) plotted for several values of n_0, assuming $n = 100$, $s_0^2 = .5$, $s^2 = 1$, and $\bar{y} = \bar{y}_0 = 10$. We see from Figure 1 that the smaller the n_0, the larger the a_g. Moreover, in Figure 1 we see how the convexity of (29) changes as n_0 increases.

Table 1. a_g for $s^2 = 1$ and $\bar{y}_0 = \bar{y} = 10$

	$n_0 = 100$			$n = 100$	
	$s_0^2 = 1$	$s_0^2 = .5$		$s_0^2 = 1$	$s_0^2 = .5$
n	a_g	a_g	n_0	a_g	a_g
5	.126	.139	5	.349	.378
10	.127	.139	10	.289	.317
50	.127	.140	50	.166	.183
100	.127	.140	100	.127	.140
200	.128	.140	200	.097	.106
1,000	.128	.140	1,000	.049	.054
100,000	.128	.141	100,000	.006	.007

Table 2 shows how the historical data can be heavily discounted when there is a lack of compatibility between the current and historical data. For example, we see in Table 2 that when $\bar{y}_0 = 50$ and $\bar{y} = 10$, the value of a_g is quite small and is approximately equal to .005 for various values of n. The same pattern emerges when n is held fixed and n_0 is varied.

Example 5: Normal Linear Regression

We consider an example to demonstrate the properties of a_0 in a linear regression context. We consider a simulation with a normal linear regression model. The historical data are generated from

$$y_{0i} = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_{0i},$$

Table 2. a_g for $s^2 = s_0^2 = 1$ and $\bar{y} = 10$

	$n_0 = 100$			$n = 100$	
	$\bar{y}_0 = 15$	$\bar{y}_0 = 50$		$\bar{y}_0 = 15$	$\bar{y}_0 = 50$
n	a_g	a_g	n_0	a_g	a_g
5	.085	.006	5	.108	.014
10	.060	.006	10	.092	.012
50	.044	.005	50	.054	.007
100	.042	.005	100	.042	.005
200	.041	.005	200	.033	.004
1,000	.041	.005	1,000	.018	.002
100,000	.041	.005	100,000	.006	.0004

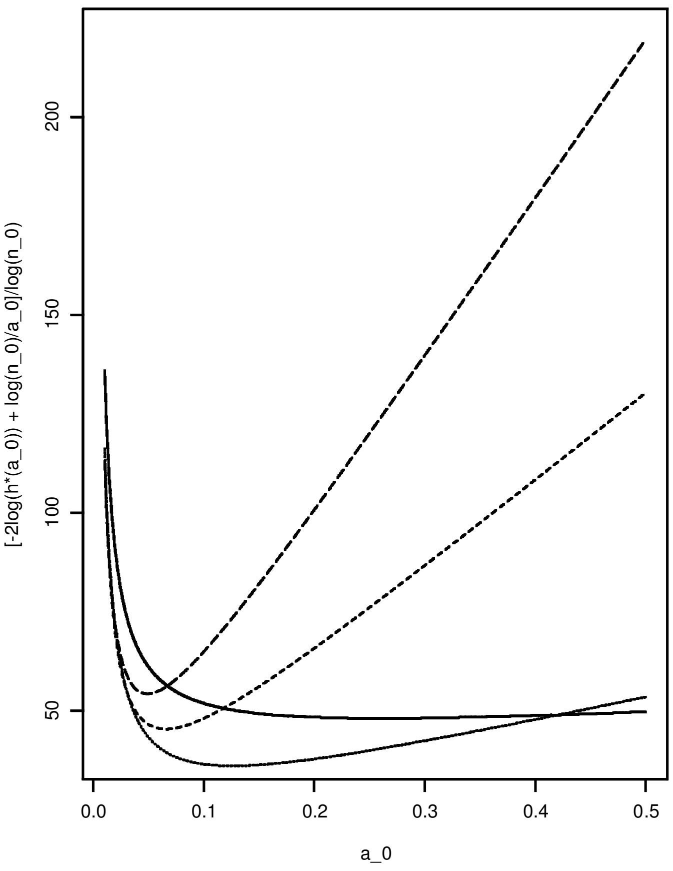

Figure 1. Criterion for Several Values of n_0 (—— $n_0 = 10$, ... $n_0 = 100$, - - - - - - $n_0 = 500$, – – – – $n_0 = 1{,}000$).

where $\epsilon_{0i} \sim N(0, \sigma_0^2)$, and the current data are simulated from

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i,$$

where $\epsilon_i \sim N(0, 1)$ and $(\beta_0, \beta_1, \beta_2) = (1, .5, -1)$. Furthermore, $x_{i2} \sim N(0, .25)$, and $x_{i1}|x_{i2}$ has a Bernoulli distribution with success probability

$$P(x_{i1} = 1|x_{i2}) = \frac{\exp(\alpha_0 + \alpha_1 x_{i2})}{1 + \exp(\alpha_0 + \alpha_1 x_{i2})},$$

where $(\alpha_0, \alpha_1) = (1, -1)$. Further, we take $\sigma^2 \sim IG(.1, .1)$, where IG denotes the inverse gamma distribution. Let $D = (n, y, X)$, $y = (y_1, \ldots, y_n)'$, and X is the $n \times 3$ $(p = 3)$ matrix of covariates with ith row $x_i' = (1, x_{i1}, x_{i2})$. The likelihood function of β (given σ^2) based on the current data is given by

$$L(\beta|\sigma^2, D) = \exp\left\{-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta)\right\},$$

and the power prior for β given σ_0^2 is given by

$$\pi(\beta|D_0, a_0, \sigma_0^2) \propto \exp\left\{-\frac{a_0}{2\sigma_0^2}(y_0 - X_0\beta)'(y_0 - X_0\beta)\right\},$$

where $y_0 = (y_{01}, \ldots, y_{0n_0})'$, X_0 is the $n_0 \times 3$ matrix of covariates based on the historical data, and $D_0 = (n_0, y_0, X_0)$.

Table 3 shows results for a_g derived from (29) based on several values of n, n_0, and σ_0^2. For ease of presentation and a clearer exposition of the sensitivity analyses in Table 3, we took σ_0^2 as a fixed value and did not specify a prior for it. However, we took a $IG(.1, .1)$ prior for σ^2. From Table 3, we see that as the historical data error variance σ_0^2 increases, the value of a_g decreases, which is a desirable property of a_g because an increasing σ_0^2 implies that the historical and current data become less compatible, and thus leads to a smaller value of a_g. Table 3 also shows the behavior of a_g for fixed n and varying n_0. We see in this table, as we saw in Example 4, that a_g decreases monotonically as n_0 increases, and takes the value .126 when $n = n_0 = 100$, as in Example 4. We further see that as σ_0^2 increases, the value of a_g decreases for each value of n_0. For example, when $\sigma_0^2 = 1$ and $n_0 = 10$, $a_g = .261$; and when $\sigma_0^2 = 10$ and $n_0 = 10$, $a_g = .125$.

Table 3. Values of a_g Based on Various n and n_0

$n_0 = 100$			$n = 100$		
σ_0^2	n	a_g	σ_0^2	n_0	a_g
1	10	.123	1	10	.261
1	100	.126	1	100	.126
1	500	.128	1	500	.066
1	1,000	.128	1	1,000	.049
10	10	.092	10	10	.125
10	100	.072	10	100	.072
10	500	.066	10	500	.043
10	1,000	.065	10	1,000	.034
100	10	.076	100	10	.052
100	100	.045	100	100	.045
100	500	.026	100	500	.034
100	1,000	.024	100	1,000	.027

Example 6: Logistic Regression

We demonstrate the role and interpretation of a_g in the context of logistic regression. We consider the two AIDS studies, ACTG019 and ACTG036, where ACTG036 represents the current data and ACTG019 represents the historical data.

The ACTG019 study was a double-blind, placebo-controlled clinical trial comparing zidovudine (AZT) to placebo in persons with CD4 counts below 500. The sample size for this study, excluding cases with missing data, was $n_0 = 823$. The response variable (y_0) for these data is binary, with 1 indicating death, development of AIDS, or AIDS-related complex (ARC) and 0 indicating otherwise. Several covariates were also measured. The ones that we use here are CD4 count (i.e., cell count per mm^3 of serum) (x_{01}), age (x_{02}), and treatment, (x_{03}). The covariates CD4 count and age are continuous, whereas the treatment covariate is binary. The ACTG036 study was also a placebo-controlled clinical trial comparing AZT with placebo in patients with hereditary coagulation disorders. The sample size in this study, excluding cases with missing data, was $n = 183$. The response variable (y) for these data is binary, with 1 indicating death, development of AIDS, or ARC and 0 indicating otherwise. Several covariates were measured for these data; we use CD4 count (x_1), age (x_2), and treatment (x_3).

The likelihood function of β for the current study is given by

$$L(\beta|D) = \exp\{y'X\beta - J'Q\},$$

Table 4. Posterior Estimates for AIDS Data

a_0	Variable	Posterior mean	Posterior standard deviation	95% HPD interval
0	Intercept	−4.78	.85	(−6.46, −3.22)
	CD4 count	−1.64	.45	(−2.54, −.79)
	Age	.12	.23	(−.33, .58)
	Treatment	−.06	.38	(−.80, .70)
$a_g = .130$	Intercept	−3.60	.42	(−4.44, −2.80)
	CD4 count	−1.02	.26	(−1.53, −.51)
	Age	.20	.18	(−.17, .55)
	Treatment	−.26	.28	(−.81, .28)
1.0	Intercept	−3.04	.17	(−3.38, −2.72)
	CD4 count	−.68	.12	(−.92, −.44)
	Age	.30	.11	(.08, .51)
	Treatment	−.38	.14	(−.65, −.11)

where $y = (y_1, \ldots, y_n)'$ denotes the $n \times 1$ vector of binary responses, J is an $n \times 1$ vector of 1s, and X is an $n \times 4$ matrix of covariates. Also, Q is an $n \times 1$ vector with jth element $\log(1+\exp(x_j'\beta))$, where x_j' denotes the jth row of X. Finally, $D = (n, y, X)$ denotes the data for the current study. For this model, the power prior is given by

$$\pi(\beta|D_0, a_0) \propto \exp\{a_0(y_0'X_0\beta - J_0'Q_0)\},$$

where y_0, X_0, J_0, and Q_0 have similar definitions as y, X, J, and Q; but based on the historical data. Using (29), the guide value for these data was estimated as $a_g = .130$. Table 4 gives posterior estimates of the regression coefficients based on the guide value as well as $a_0 = 0$ and $a_0 = 1$.

From Table 4, we see that the guide value $a_g = .130$ gives 13% weight to the historical data relative to the current data. This is consistent with earlier examples in light of the fact that the historical data sample size is $n_0 = 823$ and the current data sample size is $n = 183$. Thus, because a_g is a conservative (or skeptical) weight estimate, it takes this information into account and appropriately downweights the historical data relative to the current data.

Example 7: Semiparametric Cure Rate Model

To further demonstrate the role of a_g, we consider the power prior for right-censored survival data. We examine two melanoma clinical trials conducted by the Eastern Cooperative Oncology Group (ECOG). The first trial, denoted E1684, was a two-arm phase III clinical trial involving high-dose interferon alpha-2b (IFN) versus observation (OBS). The sample size for E1684 was $n_0 = 286$. A second trial in malignant melanoma, denoted E1690, was also conducted by ECOG. This study had $n = 427$ patients on the IFN arm and OBS arm combined. ECOG initiated this trial right after the completion of E1684 in an attempt to confirm the results of E1684. The E1690 trial was designed for exactly the same patient population as E1684, and the high-dose IFN arm in E1690 was identical to that of E1684.

We demonstrate a Bayesian analysis of E1690 using E1684 as historical data incorporated via the power prior. In addition, using (29), we compute a_g for this analysis and present the results based on this value. The response variable is taken to be relapse-free survival, and we use treatment (IFN vs. OBS) as the only covariate. Because E1684 has longer follow-up than E1690, using E1684 as historical data may help improve accuracy in the survival estimates for this disease. We use a semiparametric cure rate model here for inference (see Ibrahim, Chen, and Sinha 2001). Suppose that for an individual in the population, we let N denote the number of *metastasis-competent* tumor cells for that individual left active after the initial treatment. A metastasis-competent tumor cell is a tumor cell that has the potential to metastasize. Further, we assume that N has a Poisson distribution with mean θ. We let Z_i denote the random time for the ith metastasis-competent tumor cell to produce detectable metastatic disease. That is, Z_i can be viewed as promotion time for the ith tumor cell. The variables Z_i, $i = 1, 2, \ldots$, are assumed to be iid with a common distribution function $F(t) = 1 - S(t)$ that is independent of N. The time to relapse of cancer can be defined by the random variable $T = \min\{Z_i, 0 \le i \le N\}$, where $P(Z_0 = \infty) = 1$. Here Z_0 denotes the time to relapse of cancer given $N = 0$ metastasis-competent tumor cells, and thus $P(Z_0 = \infty) = 1$. The survival function for T, and hence the survival function for the population, is given by

$$S_p(t) = \exp(-\theta F(t)). \tag{30}$$

Because $S_p(\infty) = \exp(-\theta) > 0$, this not a proper survival function. We also see from (30) that the cure fraction (i.e., cure rate) is given by $S_p(\infty) \equiv P(N = 0) = \exp(-\theta)$. As $\theta \to \infty$, the cure fraction tends to 0, whereas as $\theta \to 0$, the cure fraction tends to 1. The hazard function is given by $h_p(t) = \theta f(t)$, leading to a proportional-hazards model if the covariates are modeled through θ.

To construct the likelihood, we first construct a finite partition of the time axis, $0 < s_1 < \cdots < s_J$, with $s_J > y_i$ for all $i = 1, \ldots, n$. Thus we have the J intervals $(0, s_1]$, $(s_1, s_2]$, $\ldots$, $(s_{J-1}, s_J]$. In the jth interval, we assume a constant hazard equal to λ_j. Throughout, we let $D = (n, y, X, \nu)$ denote the observed data for the current study, where $y = (y_1, \ldots, y_n)'$, $\nu = (\nu_1, \ldots, \nu_n)'$, and X is the $n \times p$ matrix of covariates with ith row x_i'. Letting $\lambda = (\lambda_1, \ldots, \lambda_J)$, we can write the likelihood function of (β, λ) for all n subjects as

$$\begin{aligned} L(\beta, \lambda|D_{comp}) &= \prod_{i=1}^{n}\prod_{j=1}^{J} \exp\left\{-(N_i - \nu_i)\delta_{ij}\left[\lambda_j(y_i - s_{j-1}) + \sum_{g=1}^{j-1}\lambda_g(s_g - s_{g-1})\right]\right\} \\ &\quad \times \prod_{i=1}^{n}\prod_{j=1}^{J}(N_i\lambda_j)^{\delta_{ij}\nu_i}\exp\left\{-\nu_i\delta_{ij}\left[\lambda_j(y_i - s_{j-1}) + \sum_{g=1}^{j-1}\lambda_g(s_g - s_{g-1})\right]\right\} \\ &\quad \times \exp\left\{\sum_{i=1}^{n}\left[N_i x_i'\beta - \log(N_i!) - \exp(x_i'\beta)\right]\right\}, \end{aligned} \tag{31}$$

where $D_{comp} = (n, y, X, \nu, N)$, $N = (N_1, \ldots, N_n)$, and $\lambda = (\lambda_1, \ldots, \lambda_J)$. If we take $J = 1$ in (31), then the model reduces

Table 5. Posterior Estimates of E1690

a_0	HR	SD	HPD	CR (OBS)	CR (IFN)
0	1.29	.17	(.97, 1.63)	.34	.43
a_g	1.30	.16	(.99, 1.62)	.33	.42
1	1.34	.13	(1.10, 1.60)	.28	.39

to a fully parametric cure rate model with promotion time cdf $F(t|\lambda) = 1 - \exp(-\lambda t)$.

Using the power prior based on (31) along with (29), the guide value for a_0 is estimated as $a_g = .085$. Table 5 shows estimates of the posterior hazard ratio (HR) of OBS vs. IFN, the posterior standard deviations, 95% highest posterior density (HPD) intervals for the HR, and posterior estimates of the cure rate (denoted by CR) for the OBS and IFN arms. From Table 5, we see that a_g provides a suitable starting point for further sensitivity analyses, and that it gives similar results to those based on $a_0 = 0$.

5. DISCUSSION

We have provided a formal justification of the power prior for Bayesian inference. The power prior is optimal in the sense that it minimizes the convex sum of KL divergences between two specific densities. The first density involves the posterior density of θ when $a_0 = 0$, and the second density involves the posterior density of θ when $a_0 = 1$. This optimality result is appealing, because it provides a justification for the precise form of the informative prior to use when historical data is available. This form is precisely the likelihood function of the historical data raised to a power a_0, where $0 \le a_0 \le 1$.

The procedure for finding a guide value for a_0 given by (29) appears to work quite well, but it needs further theoretical justification. This is a topic of current research. The motivation for (29) is based on penalized likelihood and is a criterion similar to the Bayesian Information Criterion (BIC). This guide value serves as a useful tool for a benchmark analysis and provides us with a conservative (skeptical) choice of a_0. The criterion in (29) is conservative in the sense that it gives us the most skeptical estimate of compatibility between historical and current data. In the examples of the previous section, we examined the behavior of a_g under various scenarios and observed that it gives very sound and interpretable results. The choice of a_0 ideally should depend how close we would like our posterior belief g to be with f_1 relative to f_0. For example, when the sample sizes for the historical and current data are comparable and we know that the datasets are fairly compatible, then $a_0 = .5$ is a suitable value, and this value is also justified by our earlier discussion in Section 2.1. If n_0 is much larger than n and we are unsure of the full compatibility between the historical and current data, then it is preferable to have g closer to f_0 rather than f_1, and hence in this case a_0 should be smaller than .5. In the absence of any definite knowledge about the compatibility between the historical and current data, we are in need of a formal assessment of a *guide value* for a_0, given here by a_g. A further motivation and interpretation of a_g, as derived from (29), is as follows. Substituting a_g into g_{opt} yields $\hat{g}_{opt} = L(\theta|D)L(\theta|D)^{a_g}\pi_0(\theta)$. Here $\hat{g}_{opt}$ represents the optimal "compromise posterior," which minimizes (6) and (29) when we give the "worst" possible relative importance to the historical data. That is, the "worst" possible relative importance is given to the posterior $\pi(\theta|D, D_0, a_0 = 1) \propto L(\theta|D)L(\theta|D_0)\pi_0(\theta)$ compared with $\pi(\theta|D, D_0, a_0 = 0)$. Thus a_g can be viewed as a conservative choice of a_0, and it allows a conservative incorporation of the historical data into the current study. This is motivated by the notion that without any concrete information about the relative importance of $\pi(\theta|D, D_0, a_0 = 1)$ compared with $\pi(\theta|D, D_0, a_0 = 0)$, the most conservative approach is to give the "worst" possible relative importance to $\pi(\theta|D, D_0, a_0 = 1)$, and this is precisely what we are doing by minimizing (29). Further research is needed to examine the theoretical properties of (29) and its relationship to optimal information processing.

Another research problem currently under investigation is the invariance properties of the power prior under transformations of the parameters. Of particular interest are theoretical properties of a_0 under one-to-one transformations of the parameters. Local invariance properties are currently being investigated.

[Received August 2001. Revised July 2002.]

REFERENCES

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.

Ibrahim, J. G., and Chen, M.-H. (2000), "Power Prior Distributions for Regression Models," *Statistical Science*, 15, 46–60.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2001), *Bayesian Survival Analysis*, New York: Springer-Verlag.

Ibrahim, J. G., Ryan, L.-M., and Chen, M.-H. (1998), "Use of Historical Controls to Adjust for Covariates in Trend Tests for Binary Data," *Journal of the American Statistical Association*, 93, 1282–1293.

Min, C. K., and Zellner, A. (1993), "Bayesian and Non-Bayesian Methods for Combining Models Forecasts With Applications to Forecasting International Growth Rates," *Journal of Econometrics*, 56, 89–118.

Soofi, E. S., and Retzer, J. J. (2002), "Information Indices: Unification and Applications," *Journal of Econometrics*, 107, 17–40.

Spiegelhalter, D. J., Freedman, L. S., and Parmar, M. K. B. (1994), "Bayesian Approaches to Randomized Trials" (with discussion), *Journal of the Royal Statistical Society*, Ser. A, 157, 357–416.

Zellner, A. (1988), "Optimal Information Processing and Bayes's Theorem" (with discussion), *The American Statistician*, 42, 278–284.

——— (1997), *Bayesian Analysis in Econometrics and Statistics*, Cheltenham, U.K.: Edward Elgar.

——— (2002), "Information Processing and Bayesian Analysis," *Journal of Econometrics*, 107, 41–50.