A Study on the Power Parameter in Power Prior Bayesian Analysis

Zifei Han\textsuperscript{a}, Keying Ye\textsuperscript{b}, and Min Wang\textsuperscript{b}

\textsuperscript{a}School of Statistics, University of International Business and Economics, Beijing, China; \textsuperscript{b}Department of Management Science and Statistics, The University of Texas at San Antonio, San Antonio, TX

\section*{ABSTRACT}
The power prior and its variations have been proven to be a useful class of informative priors in Bayesian inference due to their flexibility in incorporating the historical information by raising the likelihood of the historical data to a fractional power $\delta$. The derivation of the marginal likelihood based on the original power prior, and its variation, the normalized power prior, introduces a scaling factor $C(\delta)$ in the form of a prior predictive distribution with powered likelihood. In this article, we show that the scaling factor might be infinite for some positive $\delta$ with conventionally used initial priors, which would change the admissible set of the power parameter. This result seems to have been almost completely ignored in the literature. We then illustrate that such a phenomenon may jeopardize the posterior inference under the power priors when the initial prior of the model parameters is improper. The main findings of this article suggest that special attention should be paid when the suggested level of borrowing is close to 0, while the actual optimum might be below the suggested value. We use a normal linear model as an example for illustrative purposes.

\section*{1. Introduction}
The power prior is a class of informative priors constructed from historical data in Bayesian inference. It allows researchers to incorporate historical information via the likelihood function of the historical data raised to a power. The basic principle is to use a power parameter $\delta$ ($0 \leq \delta \leq 1$) to control the influence of the historical data on the current study. This information borrowing strategy was introduced by Ibrahim and Chen (1998, 2000) with the formulation

$$
\pi(\theta|D_0, \delta) \propto L(\theta|D_0)^{\delta} \pi_0(\theta),
$$

where $L(\theta|D_0)$ is the likelihood based on the historical data $D_0$, $\pi_0(\theta)$ is an initial prior for the parameter of interest $\theta$, and $\delta$ may be assumed to be fixed. Despite the fact that $\pi(\theta|D_0, \delta)$ depends on $D_0$ and $\delta$, it is often treated as an informative prior for the current study and is essentially a powered posterior based on $D_0$. It is worth noting that the power prior in (1.1) and the initial prior $\pi_0(\theta)$ are not necessarily proper, so long as the resulting posterior is proper. In many practical applications, it seems natural to expect most information to come from the historical data rather than the initial prior, and thus we frequently adopt noninformative priors such as reference priors (Berger, Bernardo, and Sun 2009) or weakly informative priors as the initial prior in Bayesian analyses. Assuming that the likelihood based on the current data $D$ is of the form $L(\theta|D)$, using the power prior defined in (1.1), the posterior of $\theta$ when $\delta$ is fixed has the form

$$
\pi(\theta|D, D_0, \delta) \propto L(\theta|D)\pi(\theta|D_0, \delta),
$$

where $\pi(\theta|D_0, \delta)$ is considered as a prior based on the historical data.

An important issue in the use of the power prior is to determine the level of borrowing by selecting a sensible power parameter, which can usually be determined either by expert opinions in advance, or using criteria that can reflect the prior-data conflict. Ibrahim et al. (2015) proposed multiple information-based criteria to determine the level of borrowing based on the data compatibility. Among those criteria, the marginal likelihood criterion corresponding to the empirical Bayes solution has recently been extensively studied in the literature (Gravestock and Held 2017; Nikolakopoulos, van der Tweel, and Roes 2018; Gravestock and Held 2019; Wiesenfarth and Calderazzo 2020; Ollier et al. 2020; Pateras, Nikolakopoulos, and Roes 2021). To derive the marginal likelihood, one first integrates out the model parameter $\theta$ from the right-hand side of (1.2) to obtain

$$
\int_\Theta L(\theta|D)\pi(\theta|D_0, \delta)d\theta = \frac{\int_\Theta L(\theta|D) L(\theta|D_0)^{\delta} \pi_0(\theta) d\theta}{C(\delta)},
$$

where $\pi(\theta|D_0, \delta)$ is the power prior of the form (1.1), with a normalizing constant $C(\delta) = \int_\Theta L(\theta|D_0)^{\delta} \pi_0(\theta) d\theta$, and $\Theta$ is the parameter space of $\theta$. Since $C(\delta)$ is a function of $\delta$, for the purpose of selecting a power parameter, one cannot drop it when comparing different models indexed by $\delta$. To avoid an infinite $C(\delta)$, we denote the marginal likelihood as

$$
m(\delta|D_0) = \frac{\int_\Theta L(\theta|D) L(\theta|D_0)^{\delta} \pi_0(\theta) d\theta}{C(\delta)} I_A(\delta),
$$

where $I_A(\delta)$ is the normalizing constant.
in which the indicator function $I_A(\delta) = 1$ if $\delta \in A$ and 0 otherwise, and the nonempty set is given by
\[ A = \{ \delta | \delta \in [0, 1] \text{ and } C(\delta) < \infty \}. \tag{1.4} \]

We call $A$ the feasible set of the power parameter $\delta$, which is complete when $A = [0, 1]$, and semicomplete if $A = (0, 1]$ and otherwise incomplete. Then the suggested level of borrowing using the empirical Bayes solution could be written as $\delta_{EB} = \arg \max_{\delta \in A} m(\delta|D_0, D)$ (Gravestock and Held 2017), in which the condition of $\delta \in A$ to ensure a finite $C(\delta)$ in (1.4) seems to be ignored. Consequently, this might influence the marginal likelihood in (1.3) for quantifying the data compatibility. Although the formulation guarantees that $\delta_{EB} \in A$, the domain of $A$ might be incomplete, and can thus, preclude an optimal level of borrowing.

On the other hand, as a natural Bayesian response to the uncertainty of $\delta$, we can assign a hyperprior $\pi_0(\delta)$, which results in a hierarchical power prior. When a weakly-informative prior $\pi_0(\delta)$ is used, hypothetically, the posterior of $\delta$ would reflect the data compatibility in a semiautomatic way. Therefore, Ibrahim and Chen (2000) proposed the joint prior for $\delta$ jointly with the form
\[ \pi(\theta, \delta|D_0) \propto L(\theta|D_0)^{\delta} \pi_0(\theta) \pi_0(\delta). \tag{1.5} \]

It is noted that with the prior in (1.5), the likelihood principle (Bernbaum 1962) is violated, since using various forms of the historical likelihood differ by a multiplicative positive constant $c_0$ would result in different posteriors (Their posteriors would be differed by $c_0^\delta$, see, discussions in Neuenschwander, Branson, and Spiegelhalter 2009; Ye et al. 2022). We provide a simple example in the Appendix to illustrate its consequence. The prior in (1.5) simply specifies a joint prior of $(\theta, \delta)$ directly (Ibrahim et al. 2015). If one first specifies a prior $\pi_0(\delta)$, then specifies a conditional prior of $\theta$ given $\delta$ using the power prior in (1.1), the normalizing factor $C(\delta)$ in $\pi(\theta|D_0, \delta)$ should not be dropped as well. This is due to the fact that the power $\delta$ is treated as a parameter. Therefore, Duan, Ye, and Smith (2006) proposed the following modified power prior, named as the normalized power prior, given by
\[ \pi(\theta, \delta|D_0) \propto \frac{L(\theta|D_0)^{\delta} \pi_0(\theta) \pi_0(\delta) I_A(\delta)}{C(\delta)}, \tag{1.6} \]

which obviously obeys the likelihood principle since a multiplicative constant before $L(\theta|D_0)$ will be canceled out (Neuenschwander, Branson, and Spiegelhalter 2009). This prior exists under a nonrestrictive assumption of a nonempty set for $A$, since $1 \in A$ as long as the initial prior leads to a proper posterior in a conventional Bayesian analysis. Of particular note is that in the current literature, almost all normalized power prior formulas omit $I_A(\delta)$ and assume either $\delta \in [0, 1]$ with a proper prior $\pi_0(\theta)$ or $\delta \in (0, 1]$ with an improper prior $\pi_0(\theta)$. In this study, we will show that the latter may not be the case even for the commonly used improper priors that can yield proper posteriors, such as some reference priors. This finding justifies the importance of studying the range of the admissible $\delta$ and/or the feasible set $A$ in the marginal likelihood (1.3) and the normalized power prior (1.6). However, such importance has not been given sufficient consideration in the past.

The major contribution of this study is to mathematically examine an important but almost completely ignored key point that, with a commonly used improper initial prior that is believed to be objective, the feasible parameter space $A$ of $\delta$ might be restricted to only a subregion of $(0, 1]$, that is, $A$ may be incomplete. With the exception of Duan, Ye, and Smith (2006), almost all the research articles by default assume $\delta \in (0, 1]$. As a result, either the marginal likelihood in (1.3) or the normalized power prior (1.6) might not be able to accurately quantify the data compatibility. This is because $A$ could possibly exclude the optimal range of $\delta$. We further prove that, under certain conditions, the feasible set $A$ is a convex set with a lower limit $\delta^* \geq 0$, which suggests that the phenomenon described above might happen when the heterogeneity between $D_0$ and $D$ is strong. For instance, when the optimal level of borrowing under strong heterogeneity is close to 0, the set $A$ can exclude it by definition. Albeit its impact on the analysis could be mild in many scenarios, when using the empirical Bayes or the normalized power prior, researchers should be vigilant against the use of certain predominantly used improper priors as the initial objective prior, such as the reference prior in a normal linear model. The root cause is that, despite the improper prior with a full likelihood can yield a proper posterior, the same prior with a fractional likelihood may lead to an improper posterior for some $\delta$.

The rest of the article is organized as follows. In Section 2, we establish theoretical results regarding the propriety of the power priors in general. Explicit results under the normal linear model are provided in Section 3 with several commonly used initial priors. In Section 4, we conduct a numerical study to illustrate the undesirable behavior when using the power prior with empirical Bayes or the normalized power prior in a normal linear model. We further discuss the implications of our results in Section 5.

### 2. Propriety of the Power Priors

In this section, we discuss some fundamental issues regarding the propriety of the power priors in general, which shed light on the form of a feasible set $A$ defined in (1.4). Proofs of the theorems are provided in the Appendix. For all of the following results, we consider the power prior model with the form $\pi(\theta|D_0, \delta) \propto L(\theta|D_0)^{\delta} \pi_0(\theta)$ as described in (1.1).

**Theorem 1.** When the initial prior $\pi_0(\theta)$ is proper, the power prior is always proper with a nonnegative $\delta$. Therefore, the range of feasible $\delta$ defined in (1.3) is complete, that is, $A = [0, 1]$.

It is well known that with a proper prior and a nondegenerate (full) likelihood, the set of observations in which the posterior is improper is a Lebesgue null set. This guarantees that the propriety of the posterior holds almost surely with a proper prior. **Theorem 1** demonstrates that using a proper (initial) prior $\pi_0(\theta)$ with a fractional likelihood would be a safe choice since it can also almost surely guarantee posterior propriety. However, we show in the following theorem that, this may not be the case when an improper (initial) prior of $\theta$ is used.
Remark. When the initial prior \( \pi_0(\theta) \) is improper, the power prior may or may not be proper, even if the posterior with the corresponding full likelihood is proper. In other words, \( \int_\Theta \pi_0(\theta) L(\theta | D_0) d\theta < \infty \) does not necessarily indicate \( \int_\Theta \pi_0(\theta) L(\theta | D_0)^\delta d\theta < \infty \) for all \( \delta > 0 \).

Note that unlike the proof of Theorem 1, here Jensen’s inequality can no longer guarantee a finite upper bound for an improper density \( \pi_0(\theta) \). More details can be found in the Appendix. A primary example is also provided in Section 3 for the normal linear model. This suggests that when using the marginal likelihood criterion (1.3) or the normalized power prior (1.6) with an improper initial prior on \( \theta \), the feasible set \( A \) might be restricted to a subregion of \( (0, 1] \), which is incomplete. To better understand their properties, we have the following theorem.

**Theorem 2.** If the power prior with a positive power parameter \( \delta^* \) is proper, then for all \( \delta \geq \delta^* \), \( \pi_0(\theta) L(\theta | D_0)^\delta \) is integrable.

Based on the Theorem 2 and all the related results above, we come up with the following corollary regarding the feasible parameter space \( A \) of \( \delta \) in the formulation of the marginal likelihood (1.3) and the normalized power prior (1.6).

**Corollary 2.1.** The feasible parameter space \( A \) of \( \delta \) in the marginal likelihood (1.3) and the normalized power prior (1.6) is a convex set with upper limit 1 and lower limit \( \delta^* \), where \( 0 \leq \delta^* \leq 1 \).

3. **Investigation on the Feasible Set for the Normal Linear Model**

In this section we provide some results under the commonly used normal linear model, with detailed derivations in the Appendix. These results will serve as examples to illustrate some of the findings obtained in Section 2, and will be further used for a numerical study in Section 4. We derive the results for normal linear regression model with a common unknown variance \( \sigma^2 \), while the result for a known variance is given by Ibrahim et al. (2015). The model is specified as

\[
Y = X\beta + \epsilon, \quad \text{with} \quad \epsilon \sim N_n(0_n, \sigma^2 I_n),
\]

where \( I_n \) is the \( n \times n \) identity matrix, \( N_n(0_n, \sigma^2 I_n) \) denotes the \( n \)-dimensional multivariate normal distribution with mean \( 0_n \), the \( n \)-dimensional column vector of zeroes, and covariance matrix \( \sigma^2 I_n \). The dimension of the vector \( Y \) is also \( n \) and that of \( \beta \) is \( p \). Similarly, for historical data we assume \( Y_0 = X_0\beta + \epsilon_0 \), with \( \epsilon_0 \sim N_n(0_n, \sigma^2 I_n) \), and both \( X_0'X_0 \) and \( XX \) are positive definite. We adopt similar notations to Ye et al. (2022) but consider the prior with a more general form

\[
\pi_0(\beta, \sigma^2) \propto \frac{1}{(\sigma^2)^{\nu_0}} \exp \left\{ -\frac{1}{\sigma^2} \left[ b + \frac{k}{2} (\beta - \mu_0)' R (\beta - \mu_0) \right] \right\}, \quad \text{(3.1)}
\]

where \( R \) is a known \( p \times p \) real-valued positive-definite matrix, \( \mu_0 \) is a known \( p \times 1 \) real vector, \( t \) and \( b \) are non-negative real numbers, \( k = 0 \) or \( 1 \). This class of prior includes several commonly used ones in the linear model as special cases. For example, when \( t > 1 + p/2, b > 0, k = 1 \) and \( R \) is positive definite, the prior (3.1) is the proper conjugate multivariate normal-inverse-gamma prior denoted as \( N_p, \Gamma^{-1}(\mu_0, R, t - p/2 - 1, b) \). If \( t = 1 + p/2, b = 0, k = 1 \), and matrix \( R = g^{-1}(XX) \) with \( g \) is a positive constant, the prior has the form

\[
\pi_0(\beta, \sigma^2) \propto \left( \frac{1}{\sigma^2} \right)^{\nu_0 + k/2} \exp \left\{ -\frac{(\beta - \mu_0)' XX (\beta - \mu_0)}{2g\sigma^2} \right\},
\]

which is the well-known Zellner’s \( g \)-prior (Zellner 1986). When \( t = 1, k = b = 0 \), the general form reduces to a reference prior \( \pi_0(\beta, \sigma^2) \propto 1/\sigma^2 \) (Berger, Bernardo, and Sun 2009).

**Result 1.** Consider the initial prior of the form (3.1) and assume \( n_0 > p \). Then the power prior is proper when \( \delta > (2 - 2t + p)/n_0 \). Specifically,

(a) If we use the reference prior indicated above, then the power prior is proper only when \( \delta \in (p/n_0, 1] \).

(b) If the initial prior (3.1) satisfies \( t \geq 1 + p/2 \), then the power prior is proper for all \( \delta \in (0, 1] \). This includes Zellner’s \( g \)-prior and the proper conjugate normal-inverse-gamma prior.

Next, we provide some closed-form results to guide the choice of \( \delta \) when we use the power prior with a deterministic information-based criterion. In addition to the empirical Bayes approach described above based on the marginal likelihood criterion in (1.3), we include another information criterion, the deviance information criterion (DIC) (Spiegelhalter et al. 2002), which is extensively used for model selection in Bayesian statistics. For notational simplicity, we define

\[
\hat{\beta}_0 = (X_0'X_0)^{-1} X_0'Y_0, \quad S_0 = (Y_0 - X_0\hat{\beta}_0)' (Y_0 - X_0\hat{\beta}_0),
\]

\[
\hat{\beta} = (XX)^{-1} X'Y, \quad \text{and} \quad S = (Y - X\hat{\beta})'(Y - X\hat{\beta}).
\]

Then we have the following result.

**Result 2.** Consider the initial prior \( \pi_0(\beta, \sigma^2) \) of the form in (3.1) and assume \( n_0 > p \).

(a) The marginal likelihood in (1.3) is of the form

\[
m(\delta|D_0, D) \propto \frac{\Gamma(\nu)|\delta X_0'X_0 + kR|^{1/2} H_0(\delta)^{\nu_0}}{\Gamma(\nu_0)|XX + \delta X_0'X_0 + kR|^{1/2} H(\delta)^{\nu_0}}. \quad \text{(3.2)}
\]

(b) The DIC for a model with a certain value of \( \delta \), denoted as \( \text{DIC}(\delta|D_0, D) \), up to a constant, can be expressed as

\[
\text{DIC}(\delta|D_0, D) = n \left\{ \log (v - 1) + \log H(\delta) - 2\psi(v) \right\}
\]

\[
+ \left( v + 1 \right)^{1/2} \left\{ (XX) XX + \delta X_0'X_0 + kR \right\}^{-1}
\]

\[
+ 2\text{tr}(XX XX + \delta X_0'X_0 + kR)^{-1}.\quad \text{(3.3)}
\]

In both parts, \( \nu_0 = \frac{n_0 - p}{2} + t - 1, \quad v = \nu_0 + \frac{p}{2} \), \( \hat{\beta} = (XX_0 + kR)^{1/2} (XX_0 + kR)^{1/2}, \quad \nu = (XX + \delta X_0'X_0 + kR)^{-1} (XX + \delta X_0'X_0 + kR)^{-1} \).

\[
\pi_0(\beta, \sigma^2) \propto \left( \frac{1}{\sigma^2} \right)^{\nu_0 + k/2} \exp \left\{ -\frac{(\beta - \mu_0)' XX (\beta - \mu_0)}{2g\sigma^2} \right\},
\]
The selected Figure 1. erogeneity could impact the choice of We conduct two numerical studies to explore how data het-
is still appropriately defined, since after incorporating the full
likelihood of the current data D, the resulting posterior is still
proper. However, the normalized power prior is only defined
with $A = \{p/n_0, 1\}$. Likewise, when using the empirical Bayes
method to choose $\delta_{EB}$, it must be within the range of $A$, while
using other criterion, for example the DIC, can yield different
yet more suitable results. This will be illustrated via numerical
examples below.

4. Numerical Examples

We conduct two numerical studies to explore how data heter-
erogeneity could impact the choice of $\delta$ using different criteria
and initial priors, and assess its impact on making posterior
inference. We primarily illustrate our findings with the widely
used marginal likelihood criterion, which is also equivalent to
the use of the posterior mode of $\delta_{EB}$ as a function of the sample statistics $y_0 - \bar{y}$.

In an intercept-only model (normal population without
covariates), the criteria (3.2) and (3.3) can be simplified to a
function of the sample statistics $y_0 - \bar{y}$ when the sample variances of $D_0$ and $D$ are the same. We set $y_0 - \bar{y}$ at various levels
which creates different degrees of heterogeneity, and plot its
relationship with the selected $\delta$ in Figure 1. In this experiment,
the sample sizes of $D_0$ and $D$ are $n_0 = n = 10$, with standard
deviations $s = s_0 = 0.5$, respectively, and the sample mean
of the current data is $\bar{y} = 0$. The solid curve (labeled as EB1)
is the $\delta_{EB}$ using reference (initial) prior $\pi_0(\mu, \sigma^2) \propto 1/\sigma^2$, and each point on the dotted curve is the $\delta$ associated with
the minimum DIC with the same reference prior. The dashed curve
(labeled as EB2) is the $\delta_{EB}$ with $\pi_0(\mu, \sigma^2) = \pi_0(\sigma^2)\pi_0(\mu|\sigma^2)$,
where $\pi_0(\sigma^2) \propto 1/\sigma^2$ and $\pi_0(\mu|\sigma^2) \sim N(0, 10^4 \sigma^2)$. Note
that the initial prior used in EB1 is equivalent to the use of
the prior of the form (3.1) when $p = 1$, with $t = 1$ and
$k = b = 0$. The initial prior used in EB2 is also a special
case of (3.1) when $p = 1$, with $\mu_0 = b = 0$, $t = 1.5$, $k = 1$, and
$R = 10^{-4}$.

The general trend in Figure 1 shows that the optimal $\delta$ will
decrease with the increase of $y_0 - \bar{y}$, which reflects the prior-
data conflict in an expected way. However, the desired $\delta$ should
be very close to 0 when the discrepancy between $y_0$ and $\bar{y}$ is
large. This is not the case when using an empirical Bayes with
reference prior (EB1), since the range of feasible $\delta$ is $(0, 1, 1)$,
which precludes values below 0.1. In other words, enforcing
$\delta \in (0.1, 1]$ will possibly result in borrowing more information
than the optimal choice of $\delta$ under strong heterogeneity. This
suggests that both the empirical Bayes and the normalized power
prior should be cautiously used with improper initial priors.

To assess its impact on the inferential results for model
parameters, in the following experiment, we consider a linear
regression model with an intercept and three covariates so the
regression parameter is $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)'$, and the variance
parameter is $\sigma^2$. To generate different levels of heterogeneity
between the historical and the current data, we simulate current
data $D$ with $\beta = (\beta_1, \beta_2, \beta_3, \beta_4)' = (1, 1, 1, 1)'$ and simulate
historical data $D_0$ with $\beta = (1, 1, 1, \beta_{04})'$, where $\beta_{04}$ takes a
grid of the values between 1 and 3. We generate $10^4$ datasets
for each scenario, with sample sizes $n = n_0 = 20$ and the
covariates are generated from the uniform distribution on
$(0, 1)$. For each dataset, we first choose the optimal $\delta$ with
criteria similar to those in Figure 1, and display the results of
the average $\delta$ over the $10^4$ samples in Figure 2 (left).
For each dataset, we also calculate the posterior mean of $\beta_4$ with
the power priors using the corresponding $\delta$, and report the
logarithm of the mean squared error (logMSE) in Figure 2
(right). Likewise, the empirical Bayes with the reference prior of the form $\pi_0(\beta, \sigma^2) \propto 1/\sigma^2$ is denoted as EB1, and the empirical
Bayes with the prior of the form $\pi_0(\beta, \sigma^2) \propto \pi_0(\sigma^2)\pi_0(\beta|\sigma^2)$,
where $\pi_0(\sigma^2) \propto 1/\sigma^2$ and $\pi_0(\beta|\sigma^2) \sim N_4(0_4, 10^4 \sigma^2I_4)$, is
denoted as EB2. The DIC is obtained by using the reference prior
described above as the initial prior, but selecting the optimal $\delta$
via the DIC. Setting $\delta$ as random and using the normalized
power prior (with $\pi_0(\delta) \sim \text{unif}(0, 1)$) with (initial) reference
prior provides similar results to the EB1, so it is not shown in
the plots.
Similar to Figure 1, we observe from Figure 2 (left) that on average, the borrowing strength will decrease when the level of heterogeneity between $D_0$ and $D$ increases, but the lower limit is $p/n_0 = 0.2$ when a reference prior is used with the empirical Bayes (this is also the lower bound of $\delta$ if using the normalized power prior). Therefore, when estimating the parameter $\beta_4$, at least 20% of the information will be borrowed on average from historical data regardless of the strength of heterogeneity. As a result, the logMSE increases monotonically with $\beta_{0A} - \beta_4$. On the other hand, the lower limit of $\delta$ in the other two approaches is 0, so eventually their logMSE can be maintained to the same level as no borrowing.

Overall we can conclude that, when using either the empirical Bayes with a fixed $\delta$, or the normalized power prior in a normal linear model, an improper initial prior may prevent the selection of an optimal $\delta$. The primary example is the reference prior, which is commonly used in Bayesian analysis with the full likelihood. On the other hand, under these scenarios, the DIC and other information criteria could be better alternatives.

5. Concluding Remarks

In this study, we have discussed a critical yet largely ignored key point for the use of the power priors or its modified form, the normalized power prior, to borrow information from historical data in conducting a new study. By establishing general results regarding the propriety of a powered posterior (i.e., using a fractional likelihood with power $\delta$) under various initial priors, we showed that the lower limit for $\mathcal{A}$, the feasible set of $\delta$, is not necessarily 0 with an improper initial prior, even if an initial prior can yield a proper posterior in conventional Bayesian inferences (i.e., $\delta = 1$). We thus, advocated the use of a more rigorous formulation for the normalized power prior, as well as for the formula of the marginal likelihood. These formulations are provided in (1.3) and (1.6), which account for the aforementioned restriction on the power parameter.

What is the influence of this result on Bayesian inference? We showed in a normal linear model, with the reference prior as the initial prior, the lower limit of feasible $\delta$ is $p/n_0$ when using the normalized power prior or the marginal likelihood criterion. Such impact could have a strong bearing on parameter estimates especially when the sample size of the historical data is small or moderate with some covariates, whereas the heterogeneity between the historical and the current data is strong. Therefore, an ideal level of borrowing may fall below the lower limit $p/n_0$. This might be encountered especially in the case when one splits the whole historical dataset into multiple small ones and borrows each of them individually (Banbelta et al. 2019). In these cases, we should avoid using either the marginal likelihood criterion or the normalized power prior and choose other criteria instead. In reality, it is also suggested to use multiple information criteria when possible, while another option is to use a (vague) proper initial prior.

More generally, sampling from the posterior based on a fractional likelihood is not only used for informative prior elicitation, but also a technique widely used in Bayesian computation for more generic problems. For instance, Friel and Pettitt (2008) calculated the normalizing constant based on ideas of thermodynamic integration or path sampling (Gelman and Meng 1998) with the identity

$$\log z = \int_\mathcal{A} E_{t_1}(\theta|D,\delta^*) \log L(\theta|D) \, d\delta^*, $$

where the feasible set $\mathcal{A}$ is similarly defined as in (1.3). We use the same notation $z$ as in Gelman and Meng (1998) to denote the normalizing constant in a general Bayesian model, and $L(\theta|D)$ is the likelihood based on the data $D$. Note that under this general setting we do not consider the historical data $D_0$, so essentially $D$ takes the role of $D_0$ in (1.1), and now (1.1) becomes a powered posterior. Then $\pi(\theta|D,\delta^*) \propto \pi(\theta)L(\theta|D)\delta^*$, where $\pi(\theta)$ denotes an arbitrary prior for $\theta$ such that $z = \int_\mathcal{A} \pi(\theta) L(\theta|D) \, d\theta$. The integrand is approximated by sampling from a sequence of the posterior densities based on different powered likelihoods with power $t = \{t_i\}_{i=1} \in \mathcal{A}$, where $0 \leq t_1 < \cdots < t_i \leq 1$, and the integral is approximated by the trapezoidal rule. Often $\pi(\theta)$ is assumed to be proper such that the sequence of the corresponding powered posteriors is believed to be proper, while Theorem 1 provided the evidence toward this belief. When an improper prior is used, our result indicated that the starting point $t_1$ is not necessarily very close to 0. If an improper prior that can yield a proper posterior based on the powered likelihood with $t_1 > 0$, Theorem 2 demonstrated that sampling from all the subsequent powered posteriors with $t$ is valid.
Appendix: Additional Examples and Proofs of Theorems

An example to illustrate the differences between (1.5) and (1.6) For independent Bernoulli trials with \( y_0 \) successes out of the \( n_0 \) trials in \( D_0 \), suppose the probability of success is \( \theta \). Then the historical likelihood based on the product of independent Bernoulli densities is \( L(\theta|D_0) = \theta^{n_0}(1-\theta)^{n_0-y_0} \). Assuming that \( \pi_0(\theta) \sim \text{Beta}(a_1, a_2) \), with \( a_1 > 0 \), \( a_2 > 0 \), and \( \pi_0(\delta) \) is a proper prior for \( \delta \), the joint power prior using (1.5) is of the form

\[
\pi_J(\theta, \delta|D_0) \propto \pi_0(\delta)^{b_0} \theta^{a_1} (1-\theta)^{a_2},
\]

where \( \pi_J(\theta, \delta|D_0) \) stands for a joint power prior. If we use the normalized power prior (1.6), the prior is of the form

\[
\pi_N(\theta, \delta|D_0) \propto \pi_0(\delta)^{\beta_0} \theta^{a_1} (1-\theta)^{a_2}. \]

These two priors differ by a multiplicative factor \( B(\delta_0+1, a_1, (n_0-y_0)+a_2) \) which is not a constant. Furthermore, if we use the likelihood based on the sufficient statistics \( y_0 \), which follows a binomial distribution, the joint power prior denoted as \( \pi_J^*(\theta, \delta|D_0) \) is of the form

\[
\pi_J^*(\theta, \delta|D_0) \propto \left( \frac{n_0}{\delta} \right)^{\beta_0} \pi_0(\delta|D_0).
\]

This is clearly different from \( \pi_J(\theta, \delta|D_0) \), which indicates a violation of the likelihood principle (Birnbaum 1962; Duan, Ye, and Smith 2006; Neuenschwander, Branson, and Spiegelhalter 2009). The normalized power prior remains unchanged since the extra term cancels in the numerator and denominator.

**Proof of Theorem 1.** Assume regularity conditions hold including \( L(\theta|D_0) \) is nonnegative and finite, \( \pi_0(\cdot) \geq 0 \) and is proper, and \( P(L(\theta|D_0) > 0) \) is positive. Since the function \( g(x) = x^{\delta} \) is concave \((\delta < 1)\), from Jensen's inequality we have

\[
C(\delta) = E_{\pi_0(\theta)} \left[ L(\theta|D_0)^{\delta} \right] \leq \left[ E_{\pi_0(\theta)} [L(\theta|D_0)] \right]^{\delta},
\]

which indicates

\[
\int \pi_0(\theta) L(\theta|D_0)^{\delta} d\theta \leq \left( \int \pi_0(\theta) L(\theta|D_0) d\theta \right)^{\delta}. \tag{A.1}
\]

Since \( \pi_0(\theta) \) is proper, \( C(0) = 1 \). Also the posterior based on the historical data is proper almost surely, that is, \( \int \pi_0(\theta) L(\theta|D_0) d\theta < \infty \). Therefore, \( C(\delta) \) is finite when \( \delta \in [0, 1] \). Note that this proof is used in the proof of Theorem 1 of Carvalho and Ibrahim (2001), however, its generalization in the (Carvalho and Ibrahim 2017, Remark 3) may not hold, for the reason stated below.

Some explanations on the Remark in Section 2. We first show that the inequality in (A.1) in the proof of Theorem 1 is no longer valid if the density \( \pi_0(\theta) \) is not a valid (normalized) probability distribution. Suppose \( \pi_0(\theta) = c \theta^{\alpha}(\theta) \), where \( \int \theta^{\alpha}(\theta) d\theta = 1 \), and \( c \) is a positive number. The Jensen's inequality indicates that

\[
E_{\pi^{\ast}} \left[ L(\theta|D_0)^{\delta} \right] \leq \left[ E_{\pi^{\ast}} [L(\theta|D_0)] \right]^{\delta},
\]

which implies

\[
\int \pi_0(\theta) L(\theta|D_0)^{\delta} d\theta \leq c \left( \int \pi_0(\theta) L(\theta|D_0) d\theta \right)^{\delta}.
\]

This does not satisfy the inequality in (A.1). Moreover, when \( c \) is not finite, Jensen's inequality fails to provide a finite upper bound. Therefore, to use the Jensen's inequality, one has to consider the normalizing constant, and therefore, the result in (A.1) cannot be generalized to an improper prior \( \pi_0(\theta) \). A primary example to illustrate that the powered posterior based on \( D_0 \) can be either proper or improper with an improper initial prior is given in the last paragraph of Section 3 in a normal linear model.

**Proof of Theorem 2.** Let \( \pi(\theta|D_0, \delta^*) = \pi_0(\theta) L(\theta|D_0)^{\delta^*} C(\delta^*)^{-1} \), which is assumed to be proper, and \( C(\delta^*) = \int \pi_0(\theta) L(\theta|D_0)^{\delta^*} d\theta < \infty \). Now set \( \delta = \delta - \delta^* \), where \( \delta^* \in (0, 1) \). We have

\[
\int \pi_0(\theta) L(\theta|D_0)^{\delta} d\theta = \int \pi_0(\theta) L(\theta|D_0)^{\delta^*} L(\theta|D_0)^{\delta^*} d\theta = C(\delta^*) E_{\pi(\theta|D_0, \delta^*)} \left( \frac{L(\theta|D_0)^{\delta}}{C(\delta^*)} \right).
\]

Since \( \pi(\theta|D_0, \delta^*) \) is the probability density function and \( \delta^* \in (0, 1) \), from Theorem 1, \( E_{\pi(\theta|D_0, \delta^*)} \left( \frac{L(\theta|D_0)^{\delta}}{C(\delta^*)} \right) \) is finite. Therefore, \( \int \pi_0(\theta) L(\theta|D_0)^{\delta} d\theta \) is also finite, which completed the proof.

**Proof of Result 1.** With likelihood of the form

\[
L(\beta, \sigma^2|D_0) \propto \frac{1}{(2\pi \sigma^2)^{n/2}} \exp \left\{ -\frac{1}{2\sigma^2} \left[ S_0 + (\beta - \hat{\beta}_0)^T X_0' X_0 (\beta - \hat{\beta}_0) \right] \right\},
\]

where \( \hat{\beta}_0 = (X_0' X_0)^{-1} X_0' Y_0 \) and \( S_0 = (Y_0 - X_0 \hat{\beta}_0)' (Y_0 - X_0 \hat{\beta}_0) \), we have

\[
C(\delta) \propto \left[ \int \frac{d\beta}{(2\pi)^{n/2}} \right] \exp \left\{ -\frac{(\beta - \hat{\beta}_0)^T (X_0' X_0 + k R)(\beta - \hat{\beta}_0) + 2H_0(\delta)}{2\sigma^2} \right\},
\]

where \( v_0, k, \delta \) and \( H_0(\delta) \) are defined as per Result 2. The second line follows from completing the squares of the form

\[
(\beta - \mu_0)' k R (\beta - \mu_0) + (\beta - \hat{\beta}_0)' X_0' X_0 (\beta - \hat{\beta}_0)
\]

\[
= (\beta - \hat{\beta}_0)' (X_0' X_0 + k R) (\beta - \hat{\beta}_0) + 2 k (\mu_0 - \hat{\beta}_0)' X_0' X_0 (\beta - \hat{\beta}_0) + 2H_0(\delta)
\]

and the last line follows from using the multivariate normal-inverse-gamma integral. Clearly, \( C(\delta) \) is finite when \( v_0 > 0 \). Thus, we have \( A(\delta) > 2 - 2t + 1/\nu_0 \). When \( t = 1 \) (as in the reference prior), clearly \( \delta \) is defined only when \( \delta > p/n_0 \). When \( t \geq 1 + p/2, \delta \) is defined on \( (0, 1) \).

**Proofs of Result 2.** The normalized power prior \( \pi(\beta, \sigma^2, \delta|D_0) \) is proportional to

\[
\left( \frac{1}{\sigma^2} \right)^{n/2 + t} \frac{\pi_0(\delta) H_0(\delta)^{\nu_0/2}}{\Gamma(\nu_0)/\delta X_0' X_0 + k R} \exp \left\{ -\frac{(\beta - \hat{\beta}_0)' (X_0' X_0 + k R)(\beta - \hat{\beta}_0) + 2H_0(\delta)}{2\sigma^2} \right\}.
\]

Multiplying by the likelihood of the current data \( L(\beta, \sigma^2|D_0) \), and by a similar argument, the full posterior \( \pi(\beta, \sigma^2, \delta|D_0) \) is proportional to

\[
\left( \frac{1}{\sigma^2} \right)^{n/2 + t} \frac{\pi_0(\delta) H_0(\delta)^{\nu_0/2}}{\Gamma(\nu_0)/\delta X_0' X_0 + k R} \exp \left\{ -\frac{(\beta - \hat{\beta}_0)' (X_0' X_0 + k R)(\beta - \hat{\beta}_0) + 2H_0(\delta)}{2\sigma^2} \right\}.
\]

(A.2)
where \( H(\delta) \) and \( \beta^* \) are defined in Result 2. To get the marginal posterior \( \pi(\delta|D_0, D) \) (where \( \delta \in \mathcal{A} \)), we integrate \( (\beta, \sigma^2) \) out from (A.2), which is of the form

\[
\pi(\delta|D_0, D) \propto \frac{\Gamma(v)|\delta X_0^T X_0 + k R|^{\frac{1}{2}} H_0(\delta)^{\nu_0} \pi_0(\delta)}{\Gamma(\nu_0)|X^T X + \delta X_0^T X_0 + k R|^2 H(\delta)^v},
\]

where \( v \) is defined in Result 2. Setting \( \pi_0(\delta) = 1 \) we can easily derive the marginal likelihood \( m(\delta|D_0, D) \).

The DIC calculation: For given \( \delta \in \mathcal{A} \), the conditional posterior of \( (\beta, \sigma^2|D, D_0, \delta) \) is

\[
\pi(\beta, \sigma^2|D, D_0, \delta) \propto \left( \frac{1}{\sigma^2} \right)^{\frac{v+1+\delta}{2}} \exp \left\{ -\frac{(\beta - \beta^*)^T (X^T X + \delta X_0^T X_0 + k R) (\sigma - \beta^*) + 2H(\delta)}{2\sigma^2} \right\},
\]

which is the normal-inverse-gamma kernel with \( N_\nu(\delta|\beta^*, X^T X + \delta X_0^T X_0 + k R, v, H(\delta)) \). Therefore,

\[
E(\beta|D, D_0, \delta) = \beta^*, \quad E(\sigma^2|D, D_0, \delta) = \frac{H(\delta)}{v - 1}.
\]

Recall the definition of deviance for parameter \( \theta \), defined as \( \text{Dev}_L(\theta|D) = -2 \log L(\theta|D) \). Hereafter we use the subscript \( L(\theta|D) \) to clarify that the deviance is for the current data model. Then

\[
\text{Dev}_{L(\beta, \sigma^2|D)}(E(\beta, \sigma^2|D, D_0, \delta)) = n \left\{ \log(2\pi) + \log H(\delta) - \log(v - 1) \right\} + \frac{v - 1}{H(\delta)} \left\{ (\beta^* - \tilde{\beta})^T X^T X (\sigma - \beta^*) + S \right\}.
\]

Also

\[
E \left\{ \text{Dev}_{L(\beta, \sigma^2|D)}(E(\beta, \sigma^2|D, D_0, \delta)) \right\}
= -2E_{\pi(\beta, \sigma^2|D, D_0, \delta)} \left\{ \log L(\beta, \sigma^2|D) \right\}
= n \log(2\pi) + E_{\pi(\beta, \sigma^2|D, D_0, \delta)} \left\{ n \log(\sigma^2) + \frac{S}{\sigma^2} + \frac{(\beta - \tilde{\beta})^T X^T X (\beta - \tilde{\beta})}{\sigma^2} \right\}
= n \left\{ \log(2\pi) + \log H(\delta) - \psi(v) \right\} + \frac{v}{H(\delta)} \left\{ (\beta^* - \tilde{\beta})^T X^T X (\beta^* - \tilde{\beta}) + S \right\} + \text{tr}(X^T X^T + \delta X_0^T X_0 + k R)^{-1},
\]

where \( \psi(\cdot) \) is the digamma function. The first two terms in (A.4) other than the \( n \log(2\pi) \) can be derived using the fact that the marginal posterior \( \pi(\sigma^2|D_0, D) \) follows an inverse gamma distribution with shape \( v \) and scale \( H(\delta) \). For the last term \( E_{\pi(\beta, \sigma^2|D, D_0, \delta)} \left\{ (\beta - \tilde{\beta})^T X^T X (\beta - \tilde{\beta})/\sigma^2 \right\} \), the integrand is

\[
\frac{1}{(2\pi)^{\frac{v+1+\delta}{2}}} \left( \frac{1}{\sigma^2} \right)^{\frac{v+1+\delta}{2}} \exp \left\{ -\frac{(\beta - \beta^*)^T \Lambda (\beta - \beta^*) + 2H(\delta)}{2\sigma^2} \right\} (\beta - \tilde{\beta})^T X^T X (\beta - \tilde{\beta}),
\]

where \( \Lambda = X^T X + \delta X_0^T X_0 + k R \). We first integrate \( \sigma^2 \) out from (A.5) which results in

\[
|\Lambda|^{\frac{1}{2}} \frac{1}{2} H(\delta)^{\frac{v+1+\delta}{2}} \Gamma(v) \left\{ 1 + \frac{2H(\delta)}{2H(\delta)} \right\}^{-(v+\frac{1}{2}+1)} (\beta - \tilde{\beta})^T X^T X (\beta - \tilde{\beta}).
\]

Let \( \Sigma = \frac{H(\delta)}{v + 1} \Lambda^{-1} \) and \( v^* = 2v + 2 \). Equation (A.6) can be expressed as

\[
\left| \Sigma \right|^{-\frac{1}{2}} \frac{1}{2} \Gamma \left( \frac{v^* + 1}{2} \right) \left\{ 1 + \frac{1}{v^*} (\beta - \beta^*)^T \Sigma^{-1} (\beta - \beta^*) \right\}^{\frac{v^*+1}{2}} \frac{v}{H(\delta)} (\beta - \tilde{\beta})^T X^T X (\beta - \tilde{\beta}),
\]

which is \( v H(\delta)^{-1} (\beta - \tilde{\beta})^T X^T X (\beta - \tilde{\beta}) \) multiplied by a multivariate Student-\( t \) density with location parameter \( \beta^* \), shape matrix \( \Sigma \) and the degree of freedom \( v^* \). Applying the expectation of a quadratic form we get (A.4).

Combining (A.3) and (A.4) we can easily derive an analytical form of the DIC. The effective number of parameters (Spiegelhalter et al. 2002) in our model is denoted by

\[
P_D(\delta|D_0, D) = E \left\{ \text{Dev}_{L(\beta, \sigma^2|D)}(E(\beta, \sigma^2|D, D_0, \delta)) \right\} - \text{Dev}_{L(\beta, \sigma^2|D)}(E(\beta, \sigma^2|D, D_0, \delta)) \]

\[
= n \log(2\pi) + \log H(\delta) - \psi(v) \]

\[
+ \frac{v + 1}{H(\delta)} \left\{ (\beta^* - \tilde{\beta})^T X^T X (\beta^* - \tilde{\beta}) + S \right\}
+ 2\text{tr}(X^T X^T + \delta X_0^T X_0 + k R)^{-1}. \]

Now the DIC for a model with specific \( \delta \), up to a constant, is given by

\[
\text{DIC}(\delta|D_0, D) = \text{Dev}_{L(\beta, \sigma^2|D)}(E(\beta, \sigma^2|D, D_0, \delta)) + 2P_D(\delta|D_0, D)
= n \log(2\pi) + \log H(\delta) - 2\psi(v) \]

\[
+ \frac{v + 1}{H(\delta)} \left\{ (\beta^* - \tilde{\beta})^T X^T X (\beta^* - \tilde{\beta}) + S \right\}
+ 2\text{tr}(X^T X^T + \delta X_0^T X_0 + k R)^{-1}. \]

This completed the derivation of Result 2.

Acknowledgments

The authors would like to express our deep appreciation for the Associate Editor and the anonymous reviewer for their comments and suggestions, which lead to a much improved article.

Funding

Research of Keying Ye and Min Wang were partially supported by grants from the Alvarez College of Business at the University of Texas at San Antonio.

References

Banbata, A., van Rosmalen, J., Dejardin, D., and Lesaffre, E. (2019), “Modified Power Prior with Multiple Historical Trials for Binary Endpoints,” Statistics in Medicine, 38, 1147–1169. [16]
Berger, J. O., Bernardo, J. M., and Sun, D. (2009), “The Formal Definition of Reference Priors,” The Annals of Statistics, 37, 905–938. [12, 14]
Binbaum, A. (1962), “On the Foundations of Statistical Inference,” Journal of the American Statistical Association, 57, 269–306. [13, 17]
Carvalho, L. M., and Ibrahim, J. G. (2021), “On the Normalized Power Prior,” Statistics in Medicine, 40, 5251–5275. [17]
Duan, Y., Ye, K., and Smith, E. P. (2006), “Evaluating Water Quality Using Power Priors to Incorporate Historical Information,” Environmetrics, 17, 95–106. [13, 17]
Friel, N., and Pettitt, A. N. (2008), “Marginal Likelihood Estimation via Power Posteriors,” Journal of the Royal Statistical Society, Series B, 70, 589–607. [16]
Gelman, A., and Meng, X.-L. (1998), "Simulating Normalizing Constants: From Importance Sampling to Bridge Sampling to Path Sampling," Statistical Science, 13, 163–185. [16]
Gravestock, I., and Held, L. (2017), "Adaptive Power Priors with Empirical Bayes for Clinical Trials," Pharmaceutical Statistics, 16, 349–360. [12,13]
Ibrahim, J. G., and Chen, M.-H. (1998), "Prior Distributions and Bayesian Computation for Proportional Hazards Models," Sankhya: The Indian Journal of Statistics, Series B, 60, 48–64. [12]
Ibrahim, J. G., Chen, M.-H., Gwon, Y., and Chen, F. (2015), "The Power Prior: Theory and Applications," Statistics in Medicine, 34, 3724–3749. [12,13,14]
Neuenschwander, B., Branson, M., and Spiegelhalter, D. I. (2009), "A Note on the Power Prior," Statistics in Medicine, 28, 3562–3566. [13,17]
Nikolakopoulos, S., van der Tweel, I., and Roes, K. C. B. (2018), "Dynamic Borrowing through Empirical Power Priors that Control Type I Error," Biometrics, 74, 874–880. [12]
Ollier, A., Morita, S., Ursino, M., and Zohar, S. (2020), "An Adaptive Power Prior for Sequential Clinical Trials – Application to Bridging Studies," Statistical Methods in Medical Research, 29, 2282–2294. [12]
Pateras, K., Nikolakopoulos, S., and Roes, K. C. B. (2021), "Combined Assessment of Early and Late-Phase Outcomes in Orphan Drug Development," Statistics in Medicine, 40, 2957–2974. [12]
Spiegelhalter, D., Best, N., Carlin, B., and Linde, A. (2002), "Bayesian Measures of Model Complexity and Fit," Journal of Royal Statistical Society, Series B, 64, 583–639. [14,18]
Wiesenfarth, M., and Calderazzo, S. (2020), "Quantification of Prior Impact in Terms of Effective Current Sample Size," Biometrics, 76, 326–336. [12]
Ye, K., Han, Z., Duan, Y., and Bai, T. (2022), "Normalized Power Prior Bayesian Analysis," Journal of Statistical Planning and Inference, 216, 29–50. [13,14]
Zellner, A. (1986), "On Assessing Prior Distributions and Bayesian Regression Analysis with g-Prior Distributions," in Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti. Studies, Volume 6 of Studies in Bayesian Econometrics and Statistics, eds. P. K. Goel and A. Zellner, pp. 233–243, Amsterdam: North-Holland. [14]