SAS Statistics Research and Applications
Paper #2023-03

Bayesian Models with the Power Prior Using PROC BGLIMM

Yi Gong and Fang Chen, SAS Institute Inc., Cary, NC

The power prior, a general class of priors that are used in Bayesian analysis, provides a practical and dynamic approach to translate data information into distributional information about the model parameters. Since its introduction, the power prior has played an increasingly prominent role in many disciplines. As popular as this prior has become, a software problem persists: implementation of the power prior can be difficult using Bayesian software packages and often relies on programming solutions that are problem-specific, making it hard to generalize. In this paper, we introduce new features in the SAS® BGLIMM procedure that will enable you to fit the power prior to many models (repeated measurements models, random-effects models, missing data problems, etc.) with the simplest setup. We also discuss practical issues that arise in using the power prior, such as how to work with single and multiple historical data sets, how to choose the power parameter a_0, and the marginal power prior.

Introduction

In many situations that involve collecting and analyzing data, it is common to see sequential gathering of information. This can come in the form of availability of historical data or data gathered from similar research or studies. The Bayesian paradigm offers a convenient way to update the information and enables you to use data from the past to form a prior distribution that can be used in the current analysis. When used appropriately, properly constructed priors fully demonstrate the power of the paradigm, leading to efficient modeling and accurate predictions.

However, techniques and methods used in constructing informative priors are not easy, because they typically require translating domain knowledge (from experts, often in the area of data familiarity) to model uncertainties (in the parameter space). In addition, because of the potential heterogeneity of data from slightly different sources, it then becomes of paramount importanc to find a way to adaptively incorporate various sources of information in constructing and eliciting the prior distribution.

A systematic approach to constructing priors in the presence of historical data is to use power priors. Since its introduction (Ibrahim and Chen 2000), the power prior has been widely used in many areas of statistical and data science analysis, covering all areas of discipline that require the passing of information. For an extensive review of the power prior, including its theoretical properties, variations, and applications, see (Ibrahim et al. 2015) and references within.

This paper discusses the implementation of the power prior in SAS software. Although a number of Bayesian software packages have the capability to fit models by using the power prior, the convenience that the SAS BGLIMM procedure provides stands out. We lay out the details of how this can be done easily with the procedure, and we provide practical details about a number of important aspects of using this procedure.

This paper is organized in the following sections. “The Power Prior” provides the basic formulation of the power prior. “Implementing the Power Prior in Bayesian Software” shows how the power prior can be implemented in general Bayesian software and discusses the pros and cons of such an approach. The next section, which focuses on the BGLIMM procedure, introduces the procedure and demonstrates how to use the FREQ statement to fit a wide range of models with the power prior by using a prespecified weight parameter. “Power Prior Analysis Using PROC BGLIMM” uses multiple longitudinal data sets to demonstrate various aspects of Bayesian model fitting by using the power prior, including discussion of issues such as searching and identifying an optimal value for the weight parameter. “Marginal Power Prior” discusses another practical aspect of the problem: how to implement the marginal power prior in the presence of latent variables in the model. The final section offers some discussion and summarizes the paper.

The Power Prior

In an analysis setting, suppose that there is a current data set, denoted as D, that you want to analyze. It has a sample size n and the likelihood function $L(\theta|D)$, where θ is a vector of parameters and L is a general likelihood function for arbitrary models. Further suppose that there is a historical data set from a similar previous study, D_0, with sample size n_0 and the likelihood function $L(\theta|D_0)$.

The power prior is defined as

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

where $0 \leq a_0 \leq 1$ is a scalar parameter and $\pi_0(\theta)$ is the initial, often noninformative, prior for θ before the historical data set D_0 is observed. Note that this power prior has the form of a Bayesian posterior distribution (conditional on the historical data set D_0, when a_0 is set to 1). And the a_0 parameter, sometimes referred to as a discount parameter, down-weights the likelihood function and lessens its impact on the posterior distribution based on D_0.

Using the power prior in equation (1), we can obtain the posterior distribution of θ conditional on the current data set, which is the following:

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

Note that the value a_0 controls the amount of information that is passed from the historical data set D_0 to the current analysis, and it should be set between 0 and 1. You do not want to have a prior that exaggerates the amount of information in a data set.

Implementing the Power Prior in Bayesian Software

The formulation of the power prior lends itself to implementation in a Bayesian analysis in many software packages. But as we show in this section, it is not without limitations. To implement the power prior, a Bayesian software package needs to be able to define a general likelihood function, because in almost all cases, a likelihood function raised to a power becomes a nonstandard distribution. Without losing generality, here we use the MCMC procedure (Chen (2009), Chen, Brown, and Stokes (2016), Chen and Stokes (2017)) to show how to implement the power prior and discuss the pros and cons of such an approach.

One thing to recognize is that the historical data set (the weighted likelihood function portion of the power prior) can be combined with the current data set, and the Bayesian modeling becomes a noninformative analysis (using $\pi(\theta)$ and an enlarged data set with observations weighted differently). The posterior distribution can be rewritten as follows:

$$\begin{aligned} p(\theta|D, D_0, a_0) &\propto \prod_{i=1}^{n+n_0} f_i(y_i|\theta, x_i) \cdot \pi_0(\theta) \\ \text{where } f_i &= \begin{cases} f(y_i|\theta, x_i) & \text{for each } i \text{ in the current data set} \\ f(y_{0,i}|\theta, x_{0,i})^{a_0} & \text{for each } i \text{ in the historical data set} \end{cases} \end{aligned}$$

where y_i and x_i are the response variable and covariates in the data set D, respectively, and $y_{0,i}$ and $x_{0,i}$ are the response variable and covariates in the data set D_0, respectively.

We use a simple binomial model to illustrate this implementation. Let `hist` be the historical data set, which has a response variable `y`, the number of positive outcome from an experiment; a variable `n`, the total number of subjects in this experiment; and one explanatory variable, `dose`:

y	n	dose
9	86	0.0
3	50	1.0
18	50	10
34	48	100

Correspondingly, the `curr` data set has the same variables, `y`, `n`, and `dose`, although they take different values:

y	n	dose
5	75	0.0
1	49	1.4
3	50	7.1
12	49	71

We use a logistic regression to model the data:

$$\begin{aligned} p_i &= \text{logit}(\beta_0 + \beta_1 \cdot \texttt{dose}_i) \\ y_i &\sim \text{binomial}(\texttt{n}_i, p_i) \end{aligned}$$

In this example, we choose $a_0 = 0.3$. Next, we combine the data sets and create a new `a0` variable, giving it a value of 0.3 for observations in the `hist` data set and 1 for observations in the `curr` data set:

```
data combined;
   set Hist(in=i) Curr;
   a0 = 1;
   if i then a0 = 0.3;
run;
```

This data set, `combined`, is then used in PROC MCMC as follows:

```
proc mcmc data=combined nmc=10000;
   parm b0 0 b1 0;
   prior b: ~ general(0);
   p = logistic(b0 + b1 * dose);
   llike = a0 * logpdf("binomial", y, p, n);
   model y ~ general(llike);
run;
```

The initial prior on the β parameters is a flat prior, indicated here with the `general(0)` specification ($\pi(\beta) \propto 1; \log(1) = 0$, hence the flat prior on the logarithm is 0 in the `general` function). The `llike=` assignment statement defines the weighted binomial log-likelihood function with the response variable `y`, the success probability `p`, and the number of observation variable `n`. The MODEL statement assigns `llike` as the log likelihood. For the first four observations of the `combined` data set, the likelihood function is weighted by $a_0 = 0.3$, and the remaining observations have a weight of 1.

This approach is intuitive, easy to implement, and applicable to many model specifications. However, a couple of issues remain:

- The approach requires programming, such as the specification of the general and weighted likelihood function. This is not burdensome, but as a model gets more complex, such as in situations that involve random effects and/or missing data, this approach requires a greater level of programming sophistication.
- Many analyses would require the calculation of the deviance information criteria (DIC; Spiegelhalter et al. (2002)), which, for example, can be used to evaluate model fit and determine an optimal value of a_0. But because almost all software packages compute the DIC value on the basis of observations in the input data set (in the binomial example case, the `combined` data set, not the `curr` data set), you get incorrect DIC values. This means that you have to compute the DIC value separately, perhaps using the DATA step after sampling the posterior distribution. This can be cumbersome and not ideal for maintaining clean code (requiring dual maintenance of the model: once in the software, once in the DIC calculation).

In contrast to general Bayesian software packages that rely on programming inputs from users, PROC BGLIMM offers a convenient alternative. Not only does the procedure handle a large variety of Bayesian models, but it also has features that enable modeling by using the power prior. In the next section, we briefly introduce PROC BGLIMM and then show how to use it to fit models by using the power prior.

PROC BGLIMM for Generalized Linear Mixed-Effects Models

The BGLIMM procedure (Shi and Chen 2019) is designed specifically to fit Bayesian generalized linear mixed-effects models. The procedure builds its syntax on the basis of the popular GLIMMIX and MIXED procedures in SAS (SAS Institute Inc. 2022). The specification of the mixed-effects models and syntax should be familiar to many SAS users.

In SAS procedures, a linear mixed-effects model is specified as follows:

$$\begin{aligned} \mathbf{Y} &= \mathbf{X}\beta + \mathbf{Z}\gamma + \epsilon \\ \gamma &\sim N(\mathbf{0}, \mathbf{G}) \\ \epsilon &\sim N(\mathbf{0}, \mathbf{R}) \end{aligned}$$

where β are the fixed-effects parameters (regression coefficients that are the same for all observations in the data set) and γ are the random-effects parameters (coefficients that vary across different clusters). $\mathbf{Y}$ is the vectorized response variable, $\mathbf{X}$ are the vectorized fixed-effects covariates, and $\mathbf{Z}$ are the random-effects covariates. Both the prior distribution and the sampling distribution are assumed to be normal (often multivariate normal in nature). In addition, there are two sets of parameters, $\mathbf{G}$ and $\mathbf{R}$:

- $\mathbf{G}$, the G-side matrix, is the covariance matrix of the random effects.
- $\mathbf{R}$, the R-side matrix, is the covariance matrix of the residuals.

When you specify the identity link function and the normal likelihood function, PROC BGLIMM fits a linear mixed-effects model. It also fits generalized linear mixed-effects models, extending the linear model by assuming the following:

- that the model contains a linear predictor in

  $$\eta = \mathbf{X}\beta + \mathbf{Z}\gamma$$

  where β and γ are the same fixed- and random-effects parameters.
- a monotone link function $g(\cdot)$ that relates the linear predictor to the mean of the outcome:

  $$\mathrm{E}[Y|\beta, \gamma] = g^{-1}(\eta) = g^{-1}(\mathbf{X}\beta + \mathbf{Z}\gamma)$$

  where $g(\cdot)$ is a differentiable monotone link function and $g^{-1}(\cdot)$ is its inverse
- a sampling distribution in the exponential family (examples include the commonly encountered binary, binomial, Poisson, normal, gamma, and negative binomial distributions)

PROC BGLIMM is a statement-driven procedure, meaning that all model details (the likelihood, prior distributions, random effects, repeated observations, weights, etc.) can all be specified using statements and options in those statements. Some of the frequently used statements and their functionality are as follows:

- MODEL: Y, X, dist, and link function
- RANDOM: random effects (Z), the G-side covariance (TYPE=)

- REPEATED: the R-side covariance (TYPE=)
- CLASS: categorical variables
- ESTIMATE: linear combination of parameters
- FREQ: frequency

For example, the following statements fit a Bayesian logistic regression:

```
proc bglimm data=curr seed=985329;
   model y/n = dose / dist=binomial link=logic;
run;
```

What makes the procedure a fitting tool to implement the power prior is its FREQ statement, which "counts" the occurrence for each observation, based on an input data set variable. When the frequency variable takes values between 0 and 1, it effectively becomes the scalar parameter a_0 in the power prior formulation.

Power Prior Analysis Using PROC BGLIMM

In this section, we demonstrate the implementation of the power prior in a data analysis example by using PROC BGLIMM.

Data

The data set used in this example is a publicly available antidepressant drug trial data set in a longitudinal study. It was based on real clinical trial data and is made available by the DIA Working Group.[1]

There are a total number of 200 patients who were randomized into two groups; an active depression treatment group and a placebo control group. The primary endpoint is the Hamilton Depression 17-item (HAMD-17) total scores, which were measured six times: at week 0 (baseline) and then at weeks 1, 2, 4, 6, and 8. In addition to the longitudinal nature of the data, about 24% and 26% of patients in the active drug and placebo groups, respectively, dropped out before week 8 (the primary analysis time point). Although analyzing missing data is not of primary interest in this example, PROC BGLIMM models missing data by default, whether or not you use the power prior.

Figure 1 shows the mean value changes from baseline by treatment group for patients who dropped out in different weeks. The drug group and placebo group profiles are drawn using solid blue lines and dashed red lines, respectively. The numbers in parentheses (for weeks 1, 2, 4, and 6) are the numbers of patients from each group who stayed up to that week and then dropped out. There are 69 and 60 patients, respectively, from the two groups who completed the study (stayed until week 8). As in the HAMD-17 score, the downward trend indicates improvement in depression. The trend of the graph indicates that most patients saw improvement over

[1] The data set is accessible at https://www.lshtm.ac.uk/research/centres-projects-groups/missing-data.

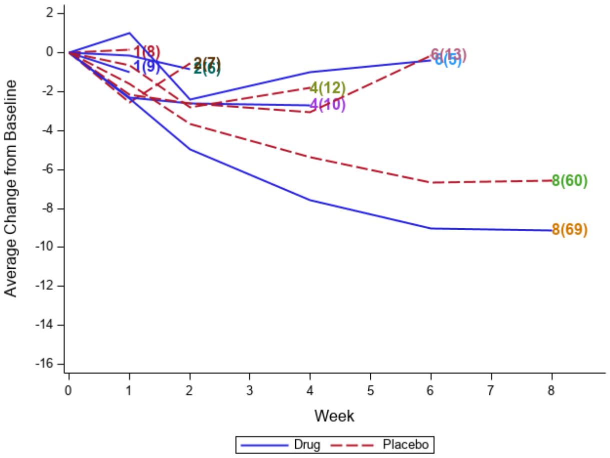

Figure 1: HAMD-17 depression scores by treatment: drug (solid blue) vs. placebo (dashed red).

time and that those who dropped out earlier perhaps experienced a smaller decrease in scores than those who stayed.

Two similarly constructed data sets are used as historical data sets. We name them historical data set 1 (HDS1) and data set 2 (HDS2). As Figure 2 indicates, there are some differences between the two historical data sets and the current data set. For example, effects (changes from baseline) appear to be greater in HDS1 than in the current data set, though they have the same trial duration; HDS2 exhibits a similar decreasing trend to that of the current data set, but the trials stopped earlier (the max is up to week 6).

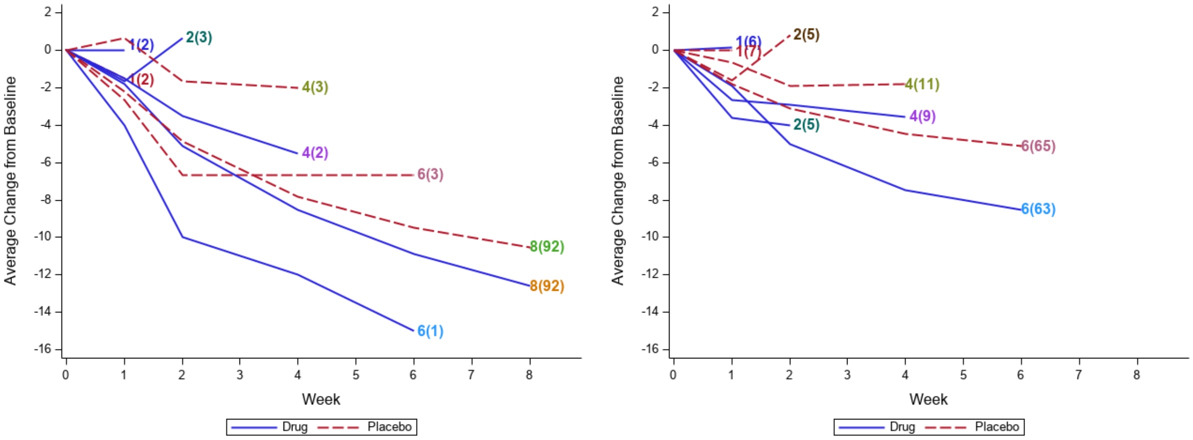

Figure 2: Patient score profiles for the two historical data sets: HDS1 (left) and HDS2 (right).

Model

To model the data, we use the following repeated measurement model, where each patient's observations over the weeks are considered to be longitudinally correlated (multivariate normal with unknown covariance matrix). Specifically, let Y_{ijk} be the outcome of interest for patient i

receiving treatment j at time k, where $i = 1, \ldots, n$, $j = 0, 1$; $k = 1, \ldots, T$. This leads to the following expectation:

$$E(Y_{ijk}) = \alpha_{jk} + \mathbf{X}_i\beta_k = \mu_{jk}$$

where α_{jk} is the intercept at time k for treatment group j and $\beta_k' = (\beta_{k1}, \ldots, \beta_{kL})$ are the regression coefficients. As for the covariates $\mathbf{X}_i$ for the ith patient, we use `therapy`, `week`, and the interaction terms `therapy` × `week` and `basval` × `week`.

Over the k time points, each patient's HAMD-17 scores are modeled using a multivariate normal distribution:

$$\mathbf{Y}_{ij} \sim \mathrm{N}(\mu_j, \Sigma)$$

where μ_j is the vectorized form of all k of the μ_{jk} and Σ is an unstructured covariance matrix.

The following program fits a repeated measurement model by using the `curr` data set and estimates the primary endpoint difference between the treatment group and the placebo group in week 8:

```
proc bglimm data=curr outpost=currOut seed=1215707 nmc=20000 nthreads=-1;
   class patient therapy week;
   model change = therapy week therapy*week basval*week;
   repeated week / subject=patient type=un;
   estimate "dp"   intercept 0 therapy 1 -1
                   therapy*week 0 0 0 0 1 0 0 0 0 -1
                   week 0 0 0 0 0
                   basval*week 0 0 0 0 0;
run;
```

PROC BGLIMM takes the `curr` data set as input and sets a Markov chain simulation size of 20,000. The NTHREADS= option specifies the number of threads (CPUs) on which to run the MCMC simulations simultaneously. Setting it to –1 uses all available threads on the system. The CLASS statement specifies the classification variables (note that the SUBJECT= variable in the REPEATED statement must be declared in the CLASS statement). The MODEL statement specifies the model. By default, PROC BGLIMM assumes a normal model with an identity link. The REPEATED statement specifies the $\mathbf{R}$ matrix in the model. By default, it is TYPE=VC.

In this analysis, the prior for the regression coefficient β is set to be a flat prior (controlled by the COEFFICIENT= option in the MODEL statement), and the default covariance matrix for $\mathbf{R}$ is an inverse-Wishart distribution with identity diagonal matrix and degrees of freedom equal to the dimension of $\mathbf{R}$ (five in this example) plus three. The prior on $\mathbf{R}$ is controlled by the COVPRIOR= option in the REPEATED statement.

The ESTIMATE statement computes linear combination of the parameter, and here it computes the end-of-study effect, drug versus placebo (`therapy` values at 1 versus –1), at week 8. The corresponding variable is named `dp`.

Figure 3 shows the posterior distributions of the effect (`dp`) at week 8 from two independent analyses: one based on the `curr` data set and the other based on the `HDS1` data set. The posterior

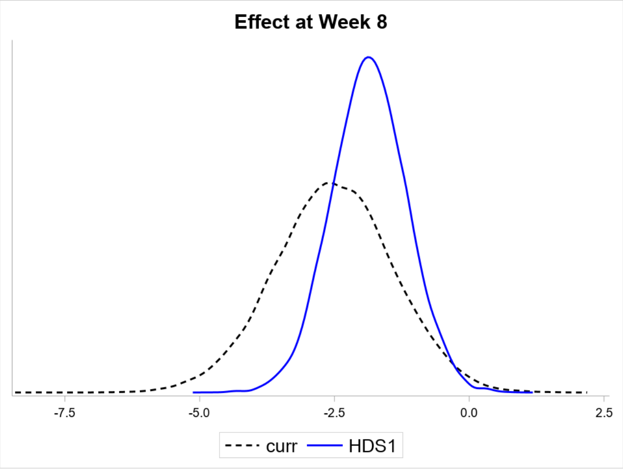

Figure 3: Posterior distributions of `dp`: `curr` (dashed) vs `HDS1` (solid) data set.

distribution from the historical data set shows a greater difference, as indicated by the profile plots in Figure 2.

Power Prior Using PROC BGLIMM

Fitting a model by using the power prior in PROC BGLIMM is similar to the approach described in the previous section using PROC MCMC, except that with PROC BGLIMM it is easier and more convenient. The first step remains the same: you combine the data sets and assign a predetermined a_0 value to observations in the historical data set:

```
data CurrHDS1;
   set hist1(in=i) curr;
   a0 = 1;
   if i then a0 = 0.3;
run;
```

The following code fits the same repeated-measurements model by using the power prior with weight $a_0 = 0.3$:

```
proc bglimm data=CurrHDS1 outpost=CurrHDS1Out seed=1215707
     nmc=20000 nthreads=-1;
   class patient therapy week;
   model change = therapy week therapy*week basval*week;
   repeated week / subject=patient type=un;
   freq a0 / notrunc;
   estimate "dp"   intercept 0 therapy 1 -1
                   therapy*week 0 0 0 0 1 0 0 0 0 -1
                   week 0 0 0 0 0
                   basval*week 0 0 0 0 0;
run;
```

The program is almost identical to that used in the noninformative analysis, with the added FREQ statement. By design, the FREQ statement “counts” the occurrence (indicated by the input variable) of each observation in the input data set. If n is the value of the FREQ variable for an observation, that observation’s likelihood function is raised to the power of n, which is the power prior formulation. If n is not an integer, then by default, the integer part of n is used as the weight. And the NOTRUNC option specifies that frequency values are not truncated to integers and instead used as they are.

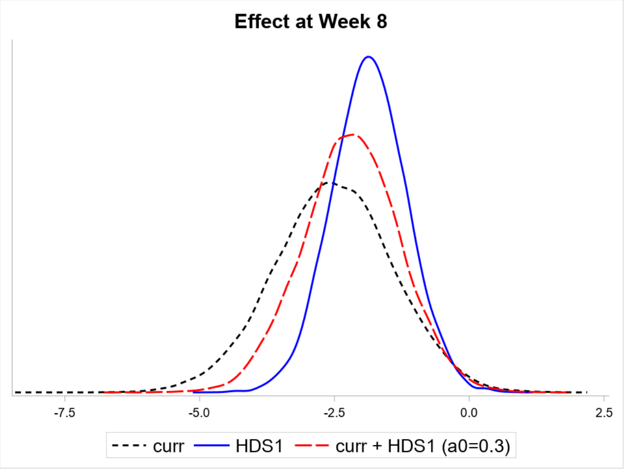

Figure 4: Posterior distributions of `dp`: `curr`-only (short-dashed line), `HDS1`-only (solid line), power prior ($a_0 = 0.3$) (long-dashed line).

Figure 4 shows an overlay of the posterior distributions from different analyses. The power prior analysis (with $a_0 = 0.3$, a weak borrowing from `HDS1`) leads to a posterior distribution that is weighted by the two independent analyses: it is between the two distributions (short-dashed line from the noninformative `curr` analysis, and solid line from the analysis using `HDS1` only), and the posterior variance also appears to be smaller than that of the `curr` analysis. This is an expected effect from the power prior, because it pulls in more information from the historical data set in the analysis, resulting in a weighted posterior with smaller variance.

Searching for an Optimal a_0 Value

One of the most important, and perhaps also most difficult, issues in using the power prior is how to choose the a_0 value. A number of approaches had been proposed. One is to treat a_0 as an unknown parameter and use the data sets to estimate it. This approach leads to a joint posterior distribution of the model parameters θ and a_0:

$$\pi(\theta, a_0|D_0) = \pi(\theta|D_0, a_0)\pi(a_0) = \frac{L(\theta|D_0)^{a_0}\pi_0(\theta)}{\int L(\theta|D_0)^{a_0}\pi_0(\theta)d\theta}\pi(a_0)$$

where $\pi_0(\theta)$ and $\pi(a_0)$ are noninformative initial prior distributions on θ and a_0. This type of power prior is sometimes referred to as the normalized power prior (Duan, Ye, and Smith (2006), Neuenschwander, Branson, and Spiegelhalter (2009)). This is a computationally intensive prior because it requires the integral with respect to θ. Except in the very simple cases of normal regression models, the analytical solution to $\int L(\theta|D_0)^{a_0}\pi_0(\theta)d\theta$ is not available, and a numerical integration routine is needed to implement this prior; this could lead to very long sampling time. Because of this difficulty, the normalizing power prior is not used much in practice.

A second approach is to take a_0 as fixed and use a model selection criterion to choose an optimal value (Ibrahim, Chen, and Sinha 2003). There are a number of criteria that you can consider, such as the penalized likelihood-type criterion, marginal likelihood criterion, deviance information criterion, and logarithm of the pseudo-marginal likelihood criterion. The most popular and computationally convenient approach is to use the deviance information criterion (DIC; Spiegelhalter et al. (2002)) to select an optimal a_0 value (Ibrahim, Chen, and Chu 2012).

When it comes to computing DIC values, as mentioned before, the important thing is that the DIC computation should use only the current data set and exclude the historical data set. The default DIC option in PROC BGLIMM computes the DIC value by using the input (the combined) data set, and thus it will produce an incorrect result.

To correct that problem, PROC BGLIMM introduces an INCLUDE= suboption in the DIC option to indicate which observations you want to use to compute the DIC value. You need to specify a data set variable in the input data set—call it dicIdx—which should take the value of 1 for observations in the curr data set and 0 for those in the hist data set. This variable is used to inform the procedure to exclude unneeded observations in the DIC computation:

```
proc bglimm data=CurrHDS1 outpost=CurrHDS1Out seed=1215707
     nmc=20000 nthreads=-1 dic(include=dicIdx);
   class patient therapy week;
   model change = therapy week therapy*week basval*week;
   repeated week / subject=patient type=un;
   freq a0 / notrunc;
run;
```

You can run the same analysis by using different a_0 values and choose the best a_0 that minimizes the DIC. This can be done using the BY statement, over copies of the CurrHDS1 data set with gridded a_0 values. The results are shown in the left panel of Figure 5, with the lowest DIC value achieved at $a_0 = 0.1$. This suggests that you want to minimize borrowing from the HDS1 data set in the analysis, which indicates potential disagreement between the historical data set and the current data set in measuring the effect of the depression drug.

With slightly more coding, you can search for optimal a_0 values based on two data sets. The results are shown in the right panel of Figure 5: you want to have less borrowing from HDS1 and more borrowing from HDS2. Higher weighting on HDS2 reflects what we observe in the data: that there is more similarity between HDS2 and the curr data set, and more borrowing can improve the fitting of the model.

Figure 6 overlays three posterior distributions of dp from different analyses of the curr data set by using different priors: noninformative priors (dashed line); the power prior using HDS1 with $a_0 = 0.1$ (solid line); and the power prior using HDS1 (with $a_0 = 0.1$) and HDS2 (with $a_0 = 0.9$; medium-dashed line). Here the posterior distribution from the power prior using HDS1 did not change much, because there is minimal amount of borrowing from that data set. The dissimilarities between the two data sets prevent an excessive amount of borrowing. The posterior distribution that uses two data sets has not shifted much; it more or less centers on the same area as the noninformative analysis. However, the posterior variance is smaller: the greater

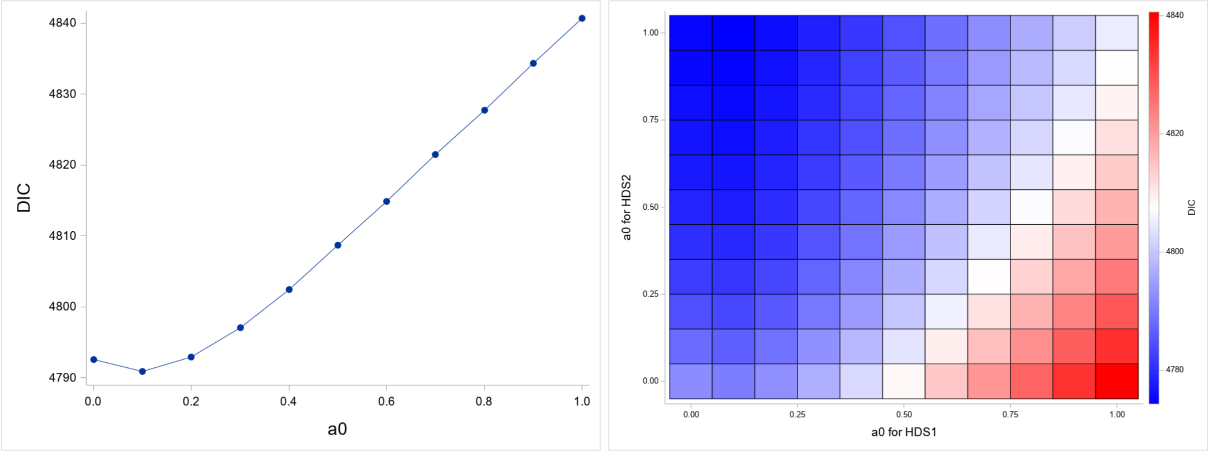

Figure 5: Left: DIC values over 11 gridded a_0 values using the HDS1 as the borrowing data set. Right: DIC values based on borrowing from two data sets. Darker blue regions indicate smaller DIC values.

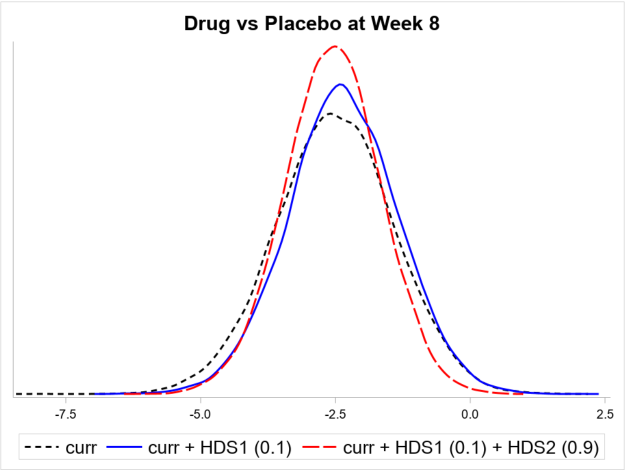

Figure 6: Posterior distributions of dp, drug vs. placebo, using optimal borrowing from two historical data sets.

amount of borrowing from a similar data set in HDS2 brings in more information to shrink the posterior distribution.

Marginal Power Prior

An underlying assumption in using the power prior is that the two likelihood functions for the current and historical data sets share the same set of parameters in θ. This could be a fairly strong assumption to make in an analysis. For example, in models with latent variables (random effects), this assumption implies that all latent variables from different data sets, or clusters, share the same characteristics. It is the same assumption that PROC BGLIMM makes in fitting a random-effects model by using the power prior. As an example, suppose that $\theta = \{\beta, \gamma\}$, where β are the fixed-effects parameters and γ the random-effects parameters. The joint posterior distribution becomes

$$\pi(\beta, \gamma | D, D_0, a_0) \propto L(\beta, \gamma | D) \cdot L(\beta, \gamma_0 | D_0)^{a_0} \cdot \pi_0(\beta, \gamma_0) \quad (3)$$

where γ_0 are random effects that are unique to the historical data set. Because γ and γ_0 are all part of a bigger model (for the combined data set), PROC BGLIMM considers them to be the same and assumes that they share the same hyperprior distribution (e.g., the same $\mathbf{G}$ covariance matrix). The impact of such an assumption is that random effects in the historical data set could influence the estimation of model parameters, such as $\mathbf{G}$, in ways that you might not want.

Alternatively, we can have a power prior that weighs on the marginal likelihood function of the fixed-effects parameters β only, which borrows on the fixed-effects parameters and not all parameters. A marginal power prior is defined as follows:

$$\begin{aligned} \pi(\beta | D_0, a_0) &\propto L(\beta | D_0)^{a_0} \pi_0(\beta) \\ &= \int L(\beta, \gamma_0 | D_0)^{a_0} \pi_0(\beta, \gamma_0) d\gamma_0 \end{aligned} \quad (4)$$

with the latent variables γ_0 integrated out.

Unfortunately, there is no direct way in PROC BGLIMM to fit a marginal power prior. This requires numerical integration over all the random-effects parameters in the part of the model that pertains to the historical data set. However, there is a relatively simple approximation workaround. You can use a multivariate normal distribution to approximate the marginal power prior on β, and then use that as the prior in an analysis that involves the current data set only. The steps are outlined as follows:

1. Fit a power prior by using the D_0 data set only, with a fixed a_0:

```
proc bglimm data=hds1 ...;
   /* same model syntax specification */
   class ...;
   model ...;
   random ...;
```

```
      freq a0 / notrunc;
   run;
```

This produces MCMC samples from the power prior directly, including the marginal distribution on β.

2. Compute the mean and covariance of β and save them to a data set `MargPrior`. Note that you should set `_type_="Mean"` and `_type_="Cov"` for the mean and covariance parameters.

3. Use the `MargPrior` data set as a prior distribution in a second PROC BGLIMM call to carry out an analysis on the `curr` data set:

```
   proc bglimm data=curr ...;
      class ...;
      model ... /
            cprior=normal(input=MargPrior);
      random ... ;
   run;
```

In most situations, the multivariate normal distribution serves well in approximating the marginal distribution of β in equation (4) and can be a good proxy to use in the analysis. As for other parameters in the model, such as the hyperparameter of the random-effects prior and/or the residual parameters on the R-side, you can use a similar approximation approach and use an inverse gamma prior, for example, to replace the marginal power prior on these parameters in an analysis.

Conclusion

Since its conception more than 20 years ago, the power prior has become a highly prominent Bayesian approach in using historical data sets as well as in eliciting informative prior. Although the power prior can be implemented in a number of general Bayesian software packages, such as the MCMC procedure, it often requires case-specific programming that can be quite complicated. The BGLIMM procedure, which specializes in fitting generalized linear mixed-effects models, has the capability to fit the power prior easily in a wide range of models. It is recommended when you need to borrow information from historical data sets in a Bayesian analysis setting. In addition to estimation and inference, we also demonstrated how to use the procedure to select an optimal a_0 value by using the DIC criterion. We also discussed the issue of the marginal power prior in the presence of latent variables and an approximation alternative. Current software does not have the capability to easily fit either the marginal power prior or the normalized power prior. This presents opportunities for future software development in Bayesian computation.

References

Chen, F. (2009). "Bayesian Modeling Using the MCMC Procedure." In *Proceedings of the SAS Global Forum 2009 Conference*. Cary, NC: SAS Institute Inc. http://support.sas.com/resources/papers/proceedings09/257-2009.pdf.

Chen, F., Brown, G., and Stokes, M. (2016). "Fitting Your Favorite Mixed Models with PROC MCMC." In *Proceedings of the SAS Global Forum 2016 Conference*. Cary, NC: SAS Institute Inc. https://support.sas.com/resources/papers/proceedings16/SAS5601-2016.pdf.

Chen, F., and Stokes, M. (2017). "Advanced Hierarchical Modeling with the MCMC Procedure." In *Proceedings of the SAS Global Forum 2017 Conference*. Cary, NC: SAS Institute Inc. https://support.sas.com/resources/papers/proceedings17/SAS478-2017.pdf.

Duan, Y., Ye, K., and Smith, E. P. (2006). "Evaluating Water Quality Using Power Priors to Incorporate Historical Information." *Environmetrics* 17:95–106.

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 Chu, H. (2012). "Bayesian Methods in Clinical Trials: A Bayesian Analysis of ECOG Trials E1684 and E1690." *BMC Medical Research Methodology* 12:183.

Ibrahim, J. G., Chen, M.-H., Gwon, Y., and Chen, F. (2015). "The Power Prior: Theory and Applications." *Statistics in Medicine* 34:3724–3749.

Ibrahim, J. G., Chen, M.-H., and Sinha, D. (2003). "On Optimality Properties of the Power Prior." *Journal of the American Statistical Association* 98:204–213.

Neuenschwander, B., Branson, M., and Spiegelhalter, D. J. (2009). "A Note on the Power Prior." *Statistics in Medicine* 28:3562–3566.

SAS Institute Inc. (2022). *SAS/STAT User's Guide*. Cary, NC: SAS Institute Inc. Revised March 2022. https://documentation.sas.com/doc/en/pgmsascdc/v_026/statug/titlepage.htm.

Shi, A., and Chen, F. (2019). "Introducing the BGLIMM Procedure for Bayesian Generalized Linear Mixed Models." In *Proceedings of the SAS Global Forum 2019 Conference*. Cary, NC: SAS Institute Inc. https://support.sas.com/resources/papers/proceedings19/3042-2019.pdf.

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (2002). "Bayesian Measures of Model Complexity and Fit." *Journal of the Royal Statistical Society, Series B* 64:583–616. With discussion.

Acknowledgment

The authors would like to thank Ed Huddleston for editing the paper.