Biostatistics (2022) **00**, 00, *pp.* 1–17
https://doi.org/10.1093/biostatistics/kxac022

A hierarchical prior for generalized linear models based on predictions for the mean response

ETHAN M. ALT*

Division of Pharmacoepidemiology and Pharmacoeconomics, Brigham and Women's Hospital and Harvard Medical School, 1620 Tremont St., Suite 3030, Boston, MA 02120, USA
ealt1@bwh.harvard.edu

MATTHEW A. PSIODA, JOSEPH G. IBRAHIM

Department of Biostatistics, University of North Carolina, 135 Dauer Drive, Chapel Hill, NC 27599, USA

Summary

There has been increased interest in using prior information in statistical analyses. For example, in rare diseases, it can be difficult to establish treatment efficacy based solely on data from a prospective study due to low sample sizes. To overcome this issue, an informative prior to the treatment effect may be elicited. We develop a novel extension of the conjugate prior of Chen and Ibrahim (2003) that enables practitioners to elicit a prior prediction for the mean response for generalized linear models, treating the prediction as random. We refer to the hierarchical prior as the hierarchical prediction prior (HPP). For independent and identically distributed settings and the normal linear model, we derive cases for which the hyperprior is a conjugate prior. We also develop an extension of the HPP in situations where summary statistics from a previous study are available. The HPP allows for discounting based on the quality of individual level predictions, and simulation results suggest that, compared to the conjugate prior and the power prior, the HPP efficiency gains (e.g., lower mean squared error) where predictions are incompatible with the data. An efficient Monte Carlo Markov chain algorithm is developed. Applications illustrate that inferences under the HPP are more robust to prior-data conflict compared to selected nonhierarchical priors.

Keywords: Bayesian inference; Generalized linear models; Hierarchical prior; Hyperprior.

1. Introduction

Exponential family models, which include distributions for the binary, count, and continuous data, are among the most utilized models for statistical analysis. In many application areas, it is desirable to incorporate prior information in an analysis. In the Bayesian paradigm, such information may be incorporated through an informative prior distribution on the parameters of interest and, possibly, nuisance parameters. For example, in rare disease clinical trials, it can be difficult to establish treatment efficacy based solely on data from a prospective study, so an informative prior may be elicited.

*To whom correspondence should be addressed.

When previous studies have been conducted, it is often desirable to construct a prior based on these data. Three popular priors that have been proposed for this setting are the power prior (PP) (Ibrahim and Chen, 2000), commensurate priors (CPs) (Hobbs *and others*, 2012), and meta-analytic-predictive (MAP) priors (e.g., Schmidli *and others*, 2014). The primary limitation of the former two priors is that they require access to an entire historical set in order to elicit a joint prior for correlated regression coefficients. MAP priors are limited in that they have only been developed under the context where there is a single parameter of interest, and it is unclear how to specify the mixture weights in such priors (Egidi *and others*, 2021). In many data applications, it is desirable to elicit a prior for regression coefficients whose components are correlated. For example, in clinical trials and observational studies, it is often of interest to determine whether an intervention is more/less efficacious for certain groups (i.e., effect heterogeneity). Another example arises when prediction is of paramount interest, such as is often the case in precision medicine.

In this article, we develop a novel extension of the conjugate prior of Chen and Ibrahim (2003), where the prediction of the mean response is treated as random. Treatment of the prior prediction as random is intuitive because the prior prediction will typically be made on the basis of summary statistics or expert opinion, both of which have some degree of uncertainty. We refer to the hierarchical prior as the "hierarchical prediction prior" (HPP). In regression models, the HPP induces a correlation structure on the regression coefficients *a priori*. The conjugate priors of Diaconis and Ylvisaker (1979) (the DY prior) and Chen and Ibrahim (2003) (the CI prior) may be cast as special cases of the HPP for independent and identically distributed (i.i.d.) and regression models, respectively. Moreover, the HPP is quite flexible, enabling practitioners to elicit predictions on the mean response based on one or more covariates.

The posterior density under the HPP is robust to incompatible prior predictions for the mean response. In particular, we show in i.i.d. settings and for the normal linear model that the posterior means of the predictions fall between the predicted values based on the maximum likelihood estimator (MLE) and the elicited prior prediction. Moreover, under some limiting cases, we show in i.i.d. data settings and the normal linear model that the HPP is a conjugate prior with the posterior maintaining the same marginal and conditional structures as the prior.

We illustrate how to utilize the HPP for settings in which a previous study was conducted, but only the point estimates and standard errors for the regression coefficients are available, comparing the posterior distribution of the parameter of interest under the HPP against that from an analysis using the CI prior, an approximation to the asymptotic PP (Ibrahim *and others*, 2015a), and the PP (Ibrahim and Chen, 2000), the latter for which the individual participant data set (IPD) was utilized by necessity. When the historical data set is incompatible with the current data set (e.g., when the historical data provide evidence that treatment is beneficial, and the current data suggest treatment is not beneficial), the posterior of the treatment effect under the HPP consistently placed more mass in the region associated with nonbeneficial effects, while the posterior distributions for the treatment effect under the other priors suggested a high probability that treatment is efficacious.

One of the most attractive features of the HPP is its versatility. The HPP may be utilized (1) to elicit a prior based on expert opinion; (2) to elicit a prior on the basis of summary statistics such that the regression coefficients are correlated *a priori*; and (3) as an alternative to existing approaches for prior elicitation in the presence of a historical data set. While we motivate the usage of the HPP in clinical trials, the prior is useful in other contexts, such as in comparative effectiveness research in observational data settings, where it is difficult to detect an effect based off of data alone due to a small number of events.

In i.i.d. settings, samples from the posterior density under the HPP ar straightforward utilizing any Markov chain Monte Carlo (MCMC) algorithm. For regression settings excluding the normal linear model, sampling is more involved because there is no general closed-form of the normalizing constant for the CI prior. We utilize a fast and accurate Laplace approximation to the normalizing constant, which enables efficient sampling using MCMC methods.

The remainder of the article is organized as follows. In Section 2, we review the conjugate prior of Diaconis and Ylvisaker (1979) and develop the HPP in i.i.d. data settings. In Section 3, we develop the HPP for generalized linear models (GLMs). In Section 4, we illustrate how the HPP may be utilized for settings in which summary statistics are available (e.g., in a publication). In Section 5, we conduct a data analysis, and in Section 6, a simulation study to compare the performance of the HPP with selected nonhierarchical priors, showing that the HPP performs favorably compared to the nonhierarchical priors when the prior prediction is incompatible with the observed data. In Section 7, we close with some discussion.

2. The independent and identically distributed case

We begin by considering the i.i.d. data setting for exponential family models. We develop the HPP for the i.i.d. case and establish theoretical connections to the conjugate prior of Diaconis and Ylvisaker (1979).

2.1. *Hyperprior motivation and construction*

Here, we discuss the motivation for the HPP and some of its properties. Suppose we observe $\{y_i, i = 1, \ldots, n\}$ with likelihood function in the exponential family, given by

$$f(\boldsymbol{y}|\theta, \phi) = \prod_{i=1}^{n} \exp\left[\frac{1}{a_i(\phi)}\{y_i\theta - b(\theta)\} + c(y_i; \phi)\right], \tag{2.1}$$

where a_i is a positive function and typically $a_i(\phi) = w_i\phi$, $\boldsymbol{y} = (y_1, \ldots, y_n)'$, $\theta \in \Theta$ is referred to as the canonical parameter, where Θ is the domain of θ, and the functions b and c index the density or mass function. We assume $a_i(\phi) = \phi$ is known and fixed, and we may suppose without loss of generality that $\phi = 1$. Diaconis and Ylvisaker (1979) showed that each distribution in the exponential family admits a conjugate prior of the form

$$\pi_{\text{DY}}(\theta|\lambda, m) = \frac{1}{Z(\lambda, \lambda m)} \exp\left[\lambda\{\theta m - b(\theta)\}\right], \theta \in \Theta, \tag{2.2}$$

where $Z(\lambda, \lambda m) = \int_\Theta \exp\left[\lambda\{\theta m - b(\theta)\}\right] d\theta$ is a normalizing constant, λ is a precision parameter typically chosen so that $\lambda \in (0, n]$, and $m \in \dot{b}(\Theta)$ is a location parameter, where $\dot{b}$ is the first derivative of the function b and $\dot{b}(\Theta)$ is the image of the set Θ under the function $\dot{b}$. Henceforth, we refer to the prior (2.2) simply as the "DY prior." Diaconis and Ylvisaker (1979) showed that, under the DY prior, $E(y) = m$, that is, m may be interpreted as a prior prediction (or "guess") for the mean of y. The hyperparameter λ controls for the level of informativeness in the prior. Let $\boldsymbol{y} = (y_1, \ldots, y_n)'$. The posterior density utilizing the likelihood (2.1) and prior (2.2) is given by $p(\theta|\boldsymbol{y}, \lambda, m) = \pi_{\text{DY}}\left(\theta \left| n + \lambda, \frac{n\bar{y} + \lambda m}{n + \lambda}\right.\right)$, where $\bar{y} = n^{-1}\sum_{i=1}^{n} y_i$ is the sample mean. The primary disadvantage of the DY prior is that the posterior is sensitive to the prior prediction m when λ is large. For example, let $\mu = \dot{b}(\theta)$ denote the mean parameter of the distribution of the response variables. It can be shown that the posterior mean of μ under the DY prior is given by $E(\mu|\boldsymbol{y}, \lambda, m) = (n\bar{y} + \lambda m)/(n + \lambda)$, i.e., the posterior mean of μ under the DY prior is a convex combination of the observed sample mean $\bar{y}$ and the prior prediction m, with higher values of λ putting more weight on the value of m in the posterior mean of μ. Thus, if m is an inaccurate prediction for $E(y)$, the posterior under the DY prior could give misleading results.

Moreover, the hyperparameter λ in the DY prior cannot simultaneously control the uncertainty surrounding m and the level of borrowing from the prior. For example, in a Bernoulli model, suppose an expert believes with 95% probability that the success probability is between 0.2 and 0.4. This may be accomplished by eliciting $m = 0.3$ and $\lambda = 78.8$ in the DY prior (2.2). This imposes a restriction on the

sample size, namely, $n \geq 79$. Conversely, for fixed values of n, λ, and m, we cannot adjust the DY prior to account for uncertainty surrounding the value of m.

Since m is typically elicited on the basis of summary statistics or expert opinion, both of which are measured with uncertainty, it is natural to view m as having a probability distribution. Thus, we propose to treat the hyperparameter m in the DY prior as random. We refer to the joint prior of (θ, m) as the HPP, and the prior on m simply as the hyperprior.

We now develop the hyperprior. Let $\nu = \dot{b}^{-1}(m) \in \Theta$. We elicit

$$\pi(\nu|\lambda_0, \mu_0) \propto \exp\left[\lambda_0\{\nu\mu_0 - b(\nu)\}\right], \nu \in \Theta, \tag{2.3}$$

which we may write in terms of the shape parameter $m = \dot{b}(\nu)$ as

$$\pi_{\text{HPP}}(m|\lambda_0, \mu_0) \propto \exp\left[\lambda_0\{\dot{b}^{-1}(m)\mu_0 - b(\dot{b}^{-1}(m))\}\right] \frac{1}{v(m)}, m \in \dot{b}(\Theta), \tag{2.4}$$

where $v(m) = \ddot{b} \circ \dot{b}^{-1}(m)$ is the variance function associated with the exponential family model (2.1). For each exponential family model, the hyperprior is a recognizable density (e.g., a beta density for binomial models, a gamma density for Poisson models, an inverse-gamma density for gamma models, and a normal density for normal models). The hyperparameter λ_0 is a precision parameter controlling for the level of certainty surrounding the prior prediction.

The HPP is obtained by combining the DY prior (2.2) and the hyperprior (2.4), giving

$$\pi_{\text{HPP}}(\theta, m|\lambda, \lambda_0\mu_0) = \pi_{\text{DY}}(\theta|\lambda, m)\, \pi_{\text{HPP}}(m|\lambda_0, \mu_0), \theta \in \Theta, m \in \dot{b}(\Theta). \tag{2.5}$$

It can be shown that the prior mean of m in the hyperprior is $E(m) = \mu_0$. Thus, $E(y) = E_\theta\{E(y|\theta)\} = E_\theta\{\dot{b}(\theta)\} = E_m\{E_{\theta|m}(\dot{b}(\theta))\} = E(m) = \mu_0$, so that, similar to the DY prior, μ_0 may be interpreted as a prior prediction for $E(y)$. However, unlike the DY prior, the HPP allows practitioners to directly elicit uncertainty surrounding the prior prediction μ_0 for any fixed λ. In the Bernoulli example above, we may elicit $\mu_0 = 0.30$ and $\lambda_0 = 78.8$, so that, *a priori*, $P(0.2 \leq m \leq 0.4) = 0.95$ for any value of λ.

Combining (2.1), (2.2), and (2.4) yields the joint posterior density

$$p(\theta, m|\boldsymbol{y}, \lambda, \lambda_0, \mu_0) \propto \pi_{\text{DY}}\left(\theta \,\middle|\, n + \lambda, \frac{n\bar{y} + \lambda m}{n + \lambda}\right) p(m|\boldsymbol{y}, \lambda, \lambda_0, \mu_0)\,. \tag{2.6}$$

Note that as $\lambda_0 \to \infty$, $p(m|\boldsymbol{y}, \lambda, \lambda_0, \mu_0)$ is simply a point mass at $m = \mu_0$, and the posterior distribution of θ thus converges to the posterior under the DY prior. Thus, we may view the HPP as a flexible generalization of the DY prior.

We may write the marginal posterior distribution of θ as

$$p(\theta|\boldsymbol{y}, \lambda, \lambda_0, \mu_0) \propto \int \pi_{\text{DY}}\left(\theta \,\middle|\, n + \lambda, \frac{n\bar{y} + \lambda m}{n + \lambda}\right) p(m|\boldsymbol{y}, \lambda, \lambda_0, \mu_0)\mathrm{d}m. \tag{2.7}$$

The relationship (2.7) indicates that the marginal posterior density of θ under the HPP may be interpreted as the posterior utilizing the DY prior with a fixed value of m averaged over the posterior distribution of m. In Section 2.2, we show that the posterior mean of m is between $\bar{y}$ and μ_0 as $\lambda_0 \to \infty$. Simulation results suggest that the result holds in general. In effect, the marginal posterior distribution of θ depends on the data $\boldsymbol{y}$ in two ways: through the shape parameter in the posterior of the DY prior conditional on m, and in the marginal posterior distribution for m. Conversely, the posterior density of the DY prior treats $m = \mu_0$ as fixed.

2.2. *Limiting posterior distributions*

In this section, we discuss an interesting limiting case of the HPP. Namely, we establish results for the prior when $\lambda \to \infty$. We may write the posterior distribution of m as

$$p(m|\boldsymbol{y}, \lambda, \lambda_0, \mu_0) \propto \frac{Z(\lambda + n, \lambda m + n\bar{y})}{Z(\lambda, \lambda m)} \pi_{\text{HPP}}(m|\lambda_0, \mu_0), \tag{2.8}$$

where $Z(a, c) = \int_\Theta \exp[a\{\theta(c/a) - b(\theta)\}]\mathrm{d}\theta$ is the normalizing constant of the DY prior with precision parameter c and shape parameter a and $\pi_{\text{HPP}}(m|\lambda_0, \mu_0)$ is defined in (2.4). As $\lambda \to \infty$, the hyperprior becomes a conjugate density, whose posterior mean is a convex combination of the sample mean and the prior prediction. We state this formally in the following theorem.

THEOREM 2.1 Let $y_1, \ldots, y_n$ be observations from a binomial, Poisson, normal, or gamma distribution. Let the prior for (θ, m) be given by the HPP (2.5). Then, $p(m|\boldsymbol{y}, \lambda, \lambda_0, \mu_0) \to \pi_{\text{HPP}}\left(m \middle| n + \lambda_0, \frac{n\bar{y}+\lambda_0\mu_0}{n+\lambda_0}\right)$ as $\lambda \to \infty$.

In words, Theorem 2.1 states that as $\lambda \to \infty$, the marginal prior of m under the HPP is conjugate density. Hence, even as $\lambda \to \infty$, the posterior distribution of m is nondegenerate. Because the marginal prior for m in the HPP is conjugate in the limit, Theorem 2.1 provides an approximation for the posterior of m for large values of λ in terms of recognizable densities (e.g., a gamma distribution for Poisson models).

An important consequence of Theorem 2.1 is $E(m|\boldsymbol{y}, \lambda, \lambda_0, \mu_0) \to (n\bar{y} + \lambda_0\mu_0)/(n + \lambda_0)$ as $\lambda \to \infty$, so that the posterior mean of m converges to a convex combination of the observed sample mean, $\bar{y}$, and the elicited prior prediction, μ_0. This is closely related to, but quite different than, the posterior distribution of μ using the DY prior. The parameter μ is a model parameter, while the hierarchical parameter m is a prediction for μ. We provide a formal proof for Theorem 2.1 in Section 1 of the Supplementary material available at *Biostatistics* online.

COROLLARY 2.2 Given the same setup as Theorem 2.1, $\lim_{\lambda\to\infty} p(\mu|\boldsymbol{y}, \lambda, \lambda_0, \mu_0) \to \pi_{\text{HPP}}\left(\mu \middle| n + \lambda_0, \frac{n\bar{y}+\lambda_0\mu_0}{n+\lambda_0}\right)$.

The proof of Corollary 2.2 is obtained from noting that as $\lambda \to \infty$, $p(\mu|m, \boldsymbol{y}, \lambda, \lambda_0, \mu_0)$ converges to a point mass at m and $p(m|\boldsymbol{y}, \lambda, \lambda_0, \mu_0)$ converges to $\pi_{\text{HPP}}(m|n + \lambda_0, (n\bar{y} + \lambda_0\mu_0)/(n + \lambda_0))$ by Theorem 2.1. In words, Corollary 2.2 states that, as $\lambda \to \infty$, the posterior distribution for the mean parameter, μ, under the HPP converges to the posterior density under the DY prior with precision parameter $n + \lambda_0$ and mean parameter $(n\bar{y} + \lambda_0\mu_0)/(n + \lambda_0)$, which is precisely the same posterior distribution obtained by eliciting $\lambda = \lambda_0$ and $m = \mu_0$ in the DY prior (2.2). When $\lambda \leq n$ (as is typically the case), the HPP cannot have more influence on the posterior than the likelihood. This fact and Corollary 2.2 imply that if $\lambda \to \infty$ and $\lambda_0 \leq n$ or $\lambda_0 \to \infty$ and $\lambda \leq n$, the posterior distribution of μ cannot depend more on the prior than the data.

REMARK 2.3 By transforming the mean parameter, μ, to the canonical parameter θ, it is easy to see from Corollary 2.2 that $p(\theta|\boldsymbol{y}, \lambda, \lambda_0, \mu_0) \to \pi_{\text{DY}}(\theta|n + \lambda_0, (n\bar{y} + \lambda_0\mu_0)/(n + \lambda_0))$ as $\lambda \to \infty$, therefore, the posterior density of θ under the HPP converges to that under the DY prior, with shape parameter equal to a convex combination of the sample mean and prior prediction.

REMARK 2.4 For the i.i.d. normal case, the HPP induces a conjugate prior on $\theta = \mu$ for finite λ. The difference between the DY prior and the HPP for the normal case is that the prior variance of μ under the HPP is larger than the DY prior for $\lambda_0 < \infty$.

3. THE REGRESSION CASE

In this section, we develop the HPP and illustrate its properties for GLMs. We also derive the posterior distribution of the regression coefficients for the normal linear model under the HPP. Furthermore, we discuss the posterior distribution of the hierarchical parameter $\boldsymbol{m}$, which is a vector for GLMs, and how possible efficiency gains may be achieved even when some of the components of the prior prediction, $\boldsymbol{\mu}_0$, are inaccurate. Finally, we discuss how to implement the HPP computationally to obtain posterior samples.

3.1. *The HPP for GLMs*

Throughout this section, suppose we observe $\{(y_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$, where y_i is a response variable and $\boldsymbol{x}_i$ is a $p \times 1$ vector of covariates associated with subject i, which may include an intercept term. Suppose $E(y_i) = \mu_i$, where $g(\mu_i) = \boldsymbol{x}_i'\boldsymbol{\beta}$, where $\boldsymbol{\beta}$ is a $p \times 1$ vector of regression coefficients. The function g is referred to as the μ-link function. The likelihood function of $\boldsymbol{y} = (y_1, \ldots, y_n)'$ may be written as

$$f(\boldsymbol{y}|\phi, \boldsymbol{\beta}, \boldsymbol{X}) = \prod_{i=1}^{n} \exp\left[\frac{1}{a_i(\phi)}\left\{y_i\theta(\boldsymbol{x}_i'\boldsymbol{\beta}) - b(\theta(\boldsymbol{x}_i'\boldsymbol{\beta}))\right\} + c(y_i, \phi)\right], \tag{3.9}$$

where $\theta(\cdot) = \dot{b}^{-1} \circ g^{-1}(\cdot)$ is referred to as the θ-link function and the functions b and c index the density or mass function. We assume $a_i(\phi) = \phi$ for $i = 1, \ldots, n$ is known and fixed, and we may suppose without loss of generality that $\phi = 1$. Chen and Ibrahim (2003) showed that the likelihood (3.9) admits a conjugate prior of the form

$$\pi_{\text{CI}}(\boldsymbol{\beta}|\lambda, \boldsymbol{m}) = \frac{1}{Z(\lambda, \lambda\boldsymbol{m})} \exp\left[\lambda\left\{\boldsymbol{m}'\theta(\boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{J}'b(\theta(\boldsymbol{X}\boldsymbol{\beta}))\right\}\right], \boldsymbol{\beta} \in \mathbb{R}^p, \tag{3.10}$$

where $Z(\lambda, \lambda\boldsymbol{m}) = \int_{\mathbb{R}^p} \exp\left[\lambda\left\{\boldsymbol{m}'\theta(\boldsymbol{X}\boldsymbol{\beta}) - \boldsymbol{J}'b(\theta(\boldsymbol{X}\boldsymbol{\beta}))\right\}\right] \mathrm{d}\boldsymbol{\beta}$ is the normalizing constant, which has no closed-form expression in general, $\lambda > 0$ is a precision parameter controlling for the informativeness of the prior, $\boldsymbol{m}$ is a n-dimensional shape parameter that may be interpreted as a prior prediction for $E(\boldsymbol{y})$, and the function b is taken componentwise. The prior (3.10), which we refer to as the "CI prior," is proper when the design matrix has full rank (Chen and Ibrahim, 2003). An alternative characterization of propriety is if, treating $\boldsymbol{m}$ as observed data and using a uniform improper prior, the posterior is proper. When $\boldsymbol{m}$ is fixed, Chen and Ibrahim (2003) showed that the posterior is $p(\boldsymbol{\beta}|\lambda, \boldsymbol{m}) = \pi_{\text{CI}}(\boldsymbol{\beta}|1 + \lambda, (\boldsymbol{y} + \lambda\boldsymbol{m})/(1 + \lambda))$. Typically, $\lambda \in (0, 1]$ so that the effective sample size contributed by the prior does not exceed that of the data. Note that, similar to the i.i.d. case, the posterior distribution of $\boldsymbol{\beta}$ for fixed $\boldsymbol{m}$ has a shape parameter that is a convex combination of the observed data $\boldsymbol{y}$ and prior prediction $\boldsymbol{m}$. Thus, it is clear that the posterior under the CI prior is sensitive to the choice of $\boldsymbol{m}$.

One of the challenges encountered when using the CI prior (3.10) is that it is difficult to quantify uncertainty surrounding the value of $\boldsymbol{m}$. That is, the CI prior treats $\boldsymbol{m}$ as observed data. While the precision hyperparameter λ controls for the degree of influence of the prior on the posterior, it does not explicitly reflect uncertainty in the prior prediction. Moreover, elicitation of $\boldsymbol{m}$ is typically based on expert opinion or summary statistics, and, thus, it is natural to view $\boldsymbol{m}$ as having a probability distribution.

To this end, we now derive the hyperprior. Let $\boldsymbol{\nu} = \dot{b}^{-1}(\boldsymbol{m})$, where the function $\dot{b}^{-1}$ is taken component-wise. We may elicit $\pi(\boldsymbol{\nu}|\lambda_0, \boldsymbol{\mu}_0) \propto \exp\left[\lambda_0 \left\{\boldsymbol{\nu}'\boldsymbol{\mu}_0 - \boldsymbol{J}'b(\boldsymbol{\nu})\right\}\right] = \prod_{i=1}^{n} \exp\left[\lambda_0 \left\{\nu_i\mu_{0i} - b(\nu_i)\right\}\right]$, $\boldsymbol{\nu} \in \Theta^n$, where $\lambda_0 > 0$ is a precision parameter and $\boldsymbol{\mu}_0 \in [\dot{b}(\Theta)]^n$ is a n-dimensional vector giving the prior prediction for $E(\boldsymbol{y})$, which may depend on covariates. In general, one may utilize a separate precision parameter for each component of $\boldsymbol{\nu}$, but we will proceed with a common precision parameter for notational convenience (allowing each component to have its own precision is discussed in Section 4).

Using $\pi(\boldsymbol{\nu}|\lambda_0, \boldsymbol{\mu}_0)$, the hyperprior is obtained using the transformation $\boldsymbol{m} = \dot{b}(\boldsymbol{\nu})$, i.e.,

$$\pi_{\text{HPP}}(\boldsymbol{m}|\lambda_0, \boldsymbol{\mu}_0) \propto \prod_{i=1}^{n} \exp\left[\lambda_0 \left\{\dot{b}^{-1}(m_i)'\mu_{0i} - b(\dot{b}^{-1}(m_i))\right\}\right] \frac{1}{v(m_i)}, \boldsymbol{m} \in \dot{b}(\Theta)^n, \tag{3.11}$$

where $v(x) = \ddot{b} \circ \dot{b}^{-1}(x)$ is the variance function of the family. Note that the hyperprior (3.11) is a product of n-independent densities, each having the same form as the i.i.d. case (2.4), where each component of $\boldsymbol{m}$ has its own mean. The hyperprior is thus a product of n-independent recognizable densities (e.g., beta for binomial models, gamma for Poisson models, inverse-gamma for gamma models, and normal for normal models). An attractive feature of the HPP is that, if covariates and regression coefficients are ignored, the HPP is a conjugate prior for each component of $\boldsymbol{y}$. For example, if $\boldsymbol{y} = (y_1, \ldots, y_n)'$ is a collection of Bernoulli random variables each with mean μ_i, $\{m_i, i = 1, \ldots, n\}$ is *a priori* a collection of n-independent beta random variables with mean μ_{0i} and dispersion parameter λ_0.

The HPP for regression models is thus given by

$$\pi_{\text{HPP}}(\boldsymbol{\beta}, \boldsymbol{m}|\lambda, \lambda_0, \boldsymbol{\mu}_0) = \pi_{\text{CI}}(\boldsymbol{\beta}|\lambda, \boldsymbol{m})\, \pi_{\text{HPP}}(\boldsymbol{m}|\lambda_0, \boldsymbol{\mu}_0). \tag{3.12}$$

The marginal prior of $\boldsymbol{\beta}$ under the HPP is thus the CI prior conditional on $\boldsymbol{m}$ averaged over a distribution on $\boldsymbol{m}$. As uncertainty surrounding the prior prediction decreases, i.e., for larger values of λ_0, the marginal prior of $\boldsymbol{\beta}$ under the HPP will become more similar to the CI prior. In particular, as $\lambda_0 \to \infty$, the HPP and the CI priors coincide.

Using the HPP (3.12), the joint posterior density may be written as

$$p(\boldsymbol{\beta}, \boldsymbol{m}|\boldsymbol{y}, \lambda, \lambda_0, \boldsymbol{\mu}_0) \propto \pi_{\text{CI}}\left(\boldsymbol{\beta} \middle| 1 + \lambda, \frac{\boldsymbol{y} + \lambda\boldsymbol{m}}{1 + \lambda}\right) \frac{Z(1 + \lambda, \boldsymbol{y} + \lambda\boldsymbol{m})}{Z(\lambda, \lambda\boldsymbol{m})}\, \pi_{\text{HPP}}(\boldsymbol{m}|\lambda_0, \boldsymbol{\mu}_0). \tag{3.13}$$

Thus, the joint posterior density may be expressed as the product of the posterior under the CI prior with shape parameter $(\boldsymbol{y} + \lambda\boldsymbol{m})/(1 + \lambda)$ and a density over $\boldsymbol{m}$, which depends on the observed data $\boldsymbol{y}$ and the prior prediction $\boldsymbol{\mu}_0$.

The HPP for GLMs is similar to a hierarchical conditional means prior (CMP) for GLMs (Bedrick *and others*, 1996). For the CMP, p *potential* response and covariate pairs $(\tilde{y}_i, \tilde{\boldsymbol{x}}_i)$ are elicited, where p is the number of regression coefficients. The $\tilde{y}_i$'s may be interpreted as a prior prediction for the mean response based on covariate $\tilde{\boldsymbol{x}}_i$. A prior inducing an *a priori* correlation structure on the regression coefficients may be obtained by treating each $\tilde{y}_i, i = 1, \ldots, p$ as random (e.g., by utilizing the hyperprior of the HPP). However, it may be difficult to justify a choice for the p potential covariate vectors $\tilde{\boldsymbol{x}}_i$. By contrast, specification of $\boldsymbol{\mu}_0$ is straightforward because it is simply a prior prediction of the mean response for each of the n observations, potentially based on *observed* (rather than potential) covariates.

While closed forms for the prior and posterior distributions under the HPP are not available in general, they are available for the normal linear model since the normalizing constant of the CI prior has a closed-form solution. We summarize the main results in the following theorem:

THEOREM 3.1 Suppose we possess data $D = \{(y_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$, where $y_i \sim N(\boldsymbol{x}_i'\boldsymbol{\beta}, \sigma^2)$. We assume σ^2 is known, and we may assume without loss of generality that $\sigma^2 = 1$. Let the HPP be given by (3.12). Then, if $\boldsymbol{X}$ is full column rank,

$$(1)\ \pi_{\text{HPP}}(\boldsymbol{\beta}, \boldsymbol{m}|\lambda, \lambda_0, \boldsymbol{\mu}_0) = \phi_{p+n}\left(\begin{pmatrix}\boldsymbol{\beta}\\ \boldsymbol{m}\end{pmatrix}\middle|\begin{pmatrix}\hat{\boldsymbol{\beta}}_{\boldsymbol{\mu}_0}\\ \boldsymbol{\mu}_0\end{pmatrix}, \begin{pmatrix}(\lambda^{-1}+\lambda_0^{-1})(\boldsymbol{X}'\boldsymbol{X})^{-1} & \lambda_0^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\\ \lambda_0^{-1}\boldsymbol{X}'(\boldsymbol{X}'\boldsymbol{X})^{-1} & \lambda_0^{-1}\boldsymbol{I}_n\end{pmatrix}\right),$$

$$(2)\ p(\boldsymbol{\beta}|\lambda, \lambda_0, \boldsymbol{\mu}_0) = \phi_p(\boldsymbol{\beta}|\boldsymbol{\mu}_{\boldsymbol{\beta}}, (1+\lambda_{\text{H}})^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}),$$

$$(3)\ p(\boldsymbol{m}|\lambda, \lambda_0, \boldsymbol{\mu}_0) = \phi_n\left(\boldsymbol{m}\middle|\boldsymbol{\mu}_{\boldsymbol{m}}, \left(\lambda_0\boldsymbol{I}_n + \tfrac{\lambda}{1+\lambda}\boldsymbol{H}\right)^{-1}\right),$$

where $\phi_p(\cdot|\boldsymbol{a}, \boldsymbol{C})$ is the p-dimensional multivariate normal density function with mean $\boldsymbol{a}$ and covariance matrix $\boldsymbol{C}$, $\hat{\boldsymbol{\beta}}_{\boldsymbol{\mu}_0} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_{\boldsymbol{\beta}} = (1+\lambda_{\text{H}})^{-1}\hat{\boldsymbol{\beta}} + [1-(1+\lambda_{\text{H}})^{-1}]\hat{\boldsymbol{\beta}}_{\boldsymbol{\mu}_0}$, $\lambda_{\text{H}} = (\lambda_0\lambda)/(\lambda_0+\lambda)$, $\boldsymbol{\mu}_{\boldsymbol{m}} = \boldsymbol{\Lambda}\boldsymbol{\mu}_0 + (\boldsymbol{I}_n - \boldsymbol{\Lambda})\hat{\boldsymbol{y}}$, $\boldsymbol{\Lambda} = (\lambda_0\boldsymbol{I}_n + \frac{\lambda}{1+\lambda}\boldsymbol{H})^{-1}$, and $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

The proof of Theorem 3.1 is presented in Section 2 of the Supplementary material available at *Biostatistics* online. Part 1 of Theorem 3.1 indicates that the prior mean of $\boldsymbol{\beta}$ is the MLE based on treating the prior prediction $\boldsymbol{\mu}_0$ as data. Part 2 of Theorem 3.1 says that the posterior mean of $\boldsymbol{\beta}$ under the HPP is a convex combination of the MLE $\hat{\boldsymbol{\beta}}$ and $\hat{\boldsymbol{\beta}}_{\boldsymbol{\mu}_0}$. Note that whenever $\lambda \leq 1$, $\lambda_{\text{H}} < 1$, so that the posterior mean of $\boldsymbol{\beta}$ is closer to the MLE than $\hat{\boldsymbol{\beta}}_{\boldsymbol{\mu}_0}$.

Part 3 of Theorem 3.1 says that the posterior distribution of $\boldsymbol{m}$ under the HPP is a convex combination of the prior prediction $\boldsymbol{\mu}_0$ and the predicted values $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$. We discuss in detail the important role that the posterior distribution of $\boldsymbol{m}$ has in Section 3.2. Note that, as $\lambda \to \infty$, $\lambda_{\text{H}} \to \lambda_0$, and thus $p(\boldsymbol{m}|\lambda, \lambda_0, \boldsymbol{\mu}_0) \to \phi_n\left(\boldsymbol{m}|\tilde{\boldsymbol{\Lambda}}\boldsymbol{\mu}_0 + (\boldsymbol{I}_n - \tilde{\boldsymbol{\Lambda}})\hat{\boldsymbol{y}}, (\lambda_0\boldsymbol{I}_n + \boldsymbol{H})^{-1}\right)$ as $\lambda \to \infty$, which is a nondegenerate density. Since $\lambda_{\text{H}} \to \lambda_0$ as $\lambda \to \infty$, we have $p(\boldsymbol{\beta}|\lambda, \lambda_0, \boldsymbol{\mu}_0) \to \phi_p(\boldsymbol{\beta}|(\hat{\boldsymbol{\beta}} + \lambda_0\hat{\boldsymbol{\beta}}_0)/(1+\lambda_0), (1+\lambda_0)^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1})$ as $\lambda \to \infty$, which is equal to the posterior density under the prior $\pi_{\text{CI}}(\boldsymbol{\beta}|\lambda_0, \boldsymbol{\mu}_0)$. Part 3 of Theorem 3.1 generalizes Theorem 2.1 for the normal linear model.

For more details about the HPP for the normal linear model, we refer the reader to Section 2 of the Supplementary material available at *Biostatistics* online. In particular, we establish a formal relationship between the HPP, the CI prior, and a multivariate normal prior for the regression coefficients.

3.2. *The posterior distribution of the hierarchical parameter*

We now discuss the posterior distribution of $\boldsymbol{m}$. We will see that the posterior distribution of $\boldsymbol{m}$ has important implications on the posterior distribution of the regression coefficients.

We may write the posterior distribution of $\boldsymbol{m}$ under the HPP as

$$p(\boldsymbol{m}|\boldsymbol{y}, \lambda, \lambda_0, \boldsymbol{\mu}_0) \propto \frac{Z(1+\lambda, \boldsymbol{y} + \lambda\boldsymbol{m})}{Z(\lambda, \lambda\boldsymbol{m})}\,\pi_{\text{HPP}}(\boldsymbol{m}|\lambda_0, \boldsymbol{\mu}_0), \tag{3.14}$$

where $Z(a, \boldsymbol{c}) = \int \exp[a\{(\boldsymbol{c}'/a)\theta(\boldsymbol{X}\boldsymbol{\beta}) - b(\theta(\boldsymbol{X}\boldsymbol{\beta}))\}]$ is the normalizing constant of the CI prior (3.10), which has no general closed-form solution. For the i.i.d. case, it was shown in Section 2.2 that, as $\lambda \to \infty$, the posterior mean of m is between the prior prediction μ_0 and the MLE $\hat{\mu} = \bar{y}$.

When λ is treated as random and $\boldsymbol{m}$ is treated as fixed (e.g., the hyper-g prior of Sabanés Bové and Held (2011)), the posterior distribution of λ reflects how accurate the prior prediction is *overall*. For example, if many of the prior predictions are inaccurate and some are highly accurate, the posterior distribution of λ may be concentrated near 0, so that the prior has little influence on the posterior. Conversely, in the HPP, $\boldsymbol{m}$ is treated as random and λ is fixed, so that the posterior always borrows from the HPP. The HPP discounts

at the prediction level, so that it is possible for efficiency gains, such as lower mean squared error (MSE), to be made even when some of the prior predictions are inaccurate. In Section 6, we conduct a large-scale simulation study suggesting that, when the prior prediction is incompatible, the posterior under the HPP exhibits lower bias and MSE and better credible region (CR) coverage than selected priors.

For the normal linear model, Part 3 of Theorem 3.1 indicates that the posterior distribution of $\boldsymbol{m}$ under the HPP is multivariate normal with mean $E(\boldsymbol{m}|\boldsymbol{y}) = \boldsymbol{\Lambda}\boldsymbol{\mu}_0 + (\boldsymbol{I}_n - \boldsymbol{\Lambda})\hat{\boldsymbol{y}}$, where $\boldsymbol{\Lambda} = \left(\lambda_0 \boldsymbol{I}_n + \frac{\lambda}{1+\lambda}\boldsymbol{H}\right)^{-1} \lambda_0 \boldsymbol{I}_n$, $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$ is the orthogonal projection operator onto the space spanned by the columns of $\boldsymbol{X}$, and $\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$ is the predicted values of $\boldsymbol{y}$, where $\hat{\boldsymbol{\beta}}$ is the MLE of $\boldsymbol{\beta}$. Details of the derivation are given in Section 2.2 of the of the Supplementary material available at *Biostatistics* online. Hence, the posterior mean of $\boldsymbol{m}$ for the normal linear model is a convex combination of the prior prediction $\boldsymbol{\mu}_0$ and the predicted values $\hat{\boldsymbol{y}}$. Note that for larger values of λ (i.e., when more information from the prior is borrowed) and for fixed values of λ_0, the posterior mean of $\boldsymbol{m}$ depends more on the predicted values than the prior prediction. This fact and the relationship (3.14) illustrate the robustness of the HPP, namely, the posterior distribution of the regression coefficients is averaged over a distribution depending on the prior prediction $\boldsymbol{\mu}_0$, but highly modified by the data $\boldsymbol{y}$. Figures A.2 and A.3 of the Supplementary material available at *Biostatistics* online show that, for Poisson and logistic regression examples, the posterior mean of $\boldsymbol{m}$ was between the predicted values and $\boldsymbol{\mu}_0$.

3.3. *Computational development*

We now discuss how to obtain posterior samples under the HPP. We develop a Laplace approximation to the normalizing constant of the CI prior.

The joint posterior may be written as

$$p(\boldsymbol{\beta}, \boldsymbol{m}|\boldsymbol{y}, \lambda, \lambda_0, \mu_0) \propto \pi_{\text{CI}}\left(\boldsymbol{\beta} \,\middle|\, 1+\lambda, \frac{\boldsymbol{y} + \lambda \boldsymbol{m}}{1+\lambda}\right) \frac{\pi_{\text{HPP}}(\boldsymbol{m}|\lambda_0, \boldsymbol{\mu}_0)}{Z(\lambda, \lambda \boldsymbol{m})}. \tag{3.15}$$

While posterior inference in the i.i.d. case is analytically tractable, that for the regression setting is more complicated because the normalizing constant $Z(\lambda, \lambda\boldsymbol{m})$ in (3.15) does not have a closed form in general.

Taking a similar approach as Sabanés Bové and Held (2011), we may utilize an integrated Laplace approximation to estimate the normalizing constant. Note that $\pi_{\text{CI}}(\boldsymbol{\beta}|a, \boldsymbol{b}/a)$ is proportional to the likelihood of a GLM with response variable $\boldsymbol{b}$ and inverse dispersion parameter a with link function θ. Thus, we may utilize maximum likelihood methods to efficiently obtain, for each proposed value $\tilde{\boldsymbol{m}}$ of $\boldsymbol{m}$, $\hat{\boldsymbol{\beta}}_{\tilde{\boldsymbol{m}}}$, the value of $\boldsymbol{\beta}$ that maximizes $\pi_{\text{CI}}(\boldsymbol{\beta}|1+\lambda, (\boldsymbol{y}+\lambda\boldsymbol{m})/(1+\lambda))$, and $\mathcal{J}(\hat{\boldsymbol{\beta}}_{\tilde{\boldsymbol{m}}})$, the observed information matrix evaluated at the maximizer. A Laplace approximation to the normalizing constant $Z(\lambda, \lambda\boldsymbol{m})$ of the prior (3.10) is given by $\hat{Z}_{\text{L}}(\lambda, \lambda\boldsymbol{m}) \equiv (2\pi)^{1/2} \mid \lambda\mathcal{J}(\hat{\boldsymbol{\beta}}_{\boldsymbol{m}}) \mid^{-1/2} \exp\left[\lambda\left\{\boldsymbol{m}'\theta(\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{m}}) - \boldsymbol{J}'b(\theta(\boldsymbol{X}\hat{\boldsymbol{\beta}}_{\boldsymbol{m}}))\right\}\right]$. Because $\hat{Z}_{\text{L}}$ is a nonstochastic estimator of the normalizing constant, we may utilize Hamiltonian MCMC for posterior sampling, such as the highly efficient No-U-Turn Sampler of Hoffman and Gelman (2014) as implemented in the R package `rstan` (Stan Development Team, 2020).

4. Prior elicitation via summary statistics

In this section, we describe how the HPP may be utilized when we possess summary statistics from a previous study obtained, for example, from a publication. We compare and contrast our approach with the PP of Ibrahim and Chen (2000).

Suppose that we possess historical data with covariates, say, $D_0 = \{(y_{0i}, \boldsymbol{x}_{0i}), i = 1, \ldots, n_0\}$. The PP for GLMs is given by

$$\pi_{\text{PP}}(\boldsymbol{\beta}|a_0, D_0) \propto \exp\left[a_0 \left\{\boldsymbol{y}_0'\theta(\boldsymbol{X}_0\boldsymbol{\beta}) - \boldsymbol{J}'b(\theta(\boldsymbol{X}_0\boldsymbol{\beta}))\right\}\right]\pi_0(\boldsymbol{\beta}), \boldsymbol{\beta} \in \mathbb{R}^p, \tag{4.16}$$

where $a_0 \in (0, 1]$ is described above and π_0 here is an initial prior for $\boldsymbol{\beta}$, which we may take as $\pi_0(\boldsymbol{\beta}) \propto 1$. The PP in (4.16) is similar to CI prior with $\lambda = a_0$, but $\boldsymbol{y}_0$ is a n_0-dimensional vector while $\boldsymbol{m}$ in the CI prior is a n-dimensional vector and the covariates from the historical data set, $\boldsymbol{X}_0$, are utilized instead of those from the current data, $\boldsymbol{X}$. Note that the PP is quite restrictive in that it requires an IPD in order to be used.

We may, however, utilize the MLE from the historical data, $\hat{\boldsymbol{\beta}}_0$, to obtain a prior prediction for the CI prior as $\boldsymbol{\mu}_0 = g^{-1}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)$. However, $\hat{\boldsymbol{\beta}}_0$ is a statistic and each of its components has a variance, and there is no direct mechanism for implementing this uncertainty into the CI prior.

By contrast, using the HPP, we may elicit uncertainty surrounding the prior prediction $\boldsymbol{\mu}_0$ via the precision parameter λ_0 via the delta method. In particular,

$$\text{Var}\left\{g^{-1}\left(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_0\right)\right\} \approx \frac{\boldsymbol{x}_i'\text{Cov}(\hat{\boldsymbol{\beta}}_0)\boldsymbol{x}_i}{\left\{\dot{g} \circ g^{-1}(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_0)\right\}^2} \equiv \tau_{0i}, i = 1, \ldots, n. \tag{4.17}$$

Typically, previous studies report only the estimated standard errors, $\{\hat{\sigma}_{0j}, j = 1, \ldots, p\}$, of $\hat{\boldsymbol{\beta}}_0$ instead of its estimated covariance matrix. In such cases, we may substitute $\widehat{\boldsymbol{\Sigma}} = \text{diag}\{\hat{\sigma}_{0j}^2, j = 1, \ldots, p\}$ for $\text{Cov}(\hat{\boldsymbol{\beta}}_0)$ in (4.17), providing a reasonable approximation for τ_{0i} in (4.17), say, $\hat{\tau}_{0i}$ as $\hat{\tau}_{0i} = \frac{\sum_{j=1}^p x_{ij}^2\hat{\sigma}_{0j}^2}{\left\{\dot{g} \circ g^{-1}(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_0)\right\}^2}, i = 1, \ldots, n.$ Once an estimate $\hat{\tau}_{0i}$ for τ_{0i} is obtained, we may compute $\mu_{0i} = g^{-1}(\boldsymbol{x}_i'\hat{\boldsymbol{\beta}}_0)$ and find λ_{0i} such that $\text{Var}(m_i) = \hat{\tau}_{0i}$, $i = 1, \ldots, n$. For example, if the outcomes are binary then m_i is beta and hence $\tau_{0i} = \mu_{0i}(1 - \mu_{0i})/(\lambda_{0i} + 1)$ so that we may elicit $\lambda_{0i} = \frac{\mu_{0i}(1-\mu_{0i})}{\tau_{0i}} - 1$. Here, we allow each component of $\boldsymbol{m}$ to have its own precision. That is, we augment the hyperprior (3.11) to $\pi_{\text{HPP}}(\boldsymbol{m}|\boldsymbol{\lambda}_0, \boldsymbol{\mu}_0) \propto \prod_{i=1}^n \exp\left[\lambda_{0i}\left\{\dot{b}^{-1}(m_i)\mu_{0i} - (b \circ \dot{b}^{-1})(m_i)\right\}\right]\frac{1}{v(m_i)}$, where $\boldsymbol{\lambda}_0 = (\lambda_{01}, \ldots, \lambda_{0n})'$ is a n-dimensional vector of precision parameters. We stress that the hyperparameters $\boldsymbol{\lambda}_0$ and $\boldsymbol{\mu}_0$ in the example described are not elicited based on an individual's opinion. Rather, the hyperparameters are deterministic functions of summary statistics from a previous study.

Both the HPP and the PP induce an *a priori* correlation structure on the regression coefficients. However, with the PP, the correlation structure depends on the observed historical response variable $\boldsymbol{y}_0$ and design matrix $\boldsymbol{X}_0$. By contrast, the correlation structure in the HPP is based on the current design matrix $\boldsymbol{X}$, the prior prediction $\boldsymbol{\mu}_0 = g^{-1}(\boldsymbol{X}\hat{\boldsymbol{\beta}}_0)$, and the prior precision $\boldsymbol{\lambda}_0$. When one is in possession of an IPD, the PP may be used, which is less computationally demanding and has various desirable properties (see, e.g., Ibrahim *and others*, 2015a). However, the HPP may be used more generally, as it only requires a prior prediction for the mean response and associated uncertainty. These values may be obtained (i) through expert opinion (see, e.g., Section 3 of the Supplementary material available at *Biostatistics* online); (ii) using summary statistics from published studies; or (iii) using MLEs obtained from an IPD and the delta method to elicit a prior prediction and associated uncertainty for the mean.

When the parameter a_0 in the PP (4.16) is treated as fixed, the prior can be highly influential on the posterior density. For example, in Section 5, we generate a historical data set with a positive treatment effect and a current data set with a null treatment effect. For the incompatible setting, the posterior under the PP suggested that treatment was all but certain to be efficacious. By contrast, the posterior density of the treatment effect under the HPP suggested that treatment may be unbeneficial, which is the correct result based on how the current data were generated.

5. Data application with a previous study

In this section, we show results applying the proposed HPP against other priors for two generated historical data sets. The other selected priors are the CI prior, the PP of Ibrahim and Chen (2000), and a "Gaussian power prior" (GPP), which is an approximation to the asymptotic PP of Ibrahim *and others* (2015a) and is given by $\pi_{\text{GPP}}(\boldsymbol{\beta}|\boldsymbol{\mu}_{\boldsymbol{\beta}}, \boldsymbol{\Sigma}_{\boldsymbol{\beta}}, \lambda) \propto \left[(2\pi)^{-1/2} \mid \boldsymbol{\Sigma}_{\boldsymbol{\beta}} \mid^{-1/2} \exp\left\{-\frac{1}{2}(\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})' \boldsymbol{\Sigma}_{\boldsymbol{\beta}}^{-1} (\boldsymbol{\beta} - \boldsymbol{\mu}_{\boldsymbol{\beta}})\right\}\right]^{\lambda}$, where $\boldsymbol{\mu}_{\boldsymbol{\beta}}$ is the prior mean of $\boldsymbol{\beta}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ is the prior covariance matrix of $\boldsymbol{\beta}$. We assume for the GPP that only the ML estimates and associated standard errors are available from the previous study, and we elicit $\boldsymbol{\mu}_{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}}$ and $\boldsymbol{\Sigma}_{\boldsymbol{\beta}} = \text{diag}\{\hat{\sigma}_j, j = 1, \ldots, p\}$, where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_1, \ldots, \hat{\beta}_p)'$ is a vector of MLEs for the regression coefficients and $\hat{\sigma}_j$ is the standard error for the jth coefficient, $j = 1, \ldots, p$. The GPP is precisely equivalent to the asymptotic PP, except the off-diagonal elements of $\boldsymbol{\Sigma}_{\boldsymbol{\beta}}$ are set to 0. The reason for this is, in practical situations where a prior is being formulated on the basis of results in published studies, the full covariance matrix of the MLEs is not typically reported (typically, only the standard errors are reported).

The PP will use a full historical data set, while hyperparameter elicitation for the three remaining priors will utilize the maximum likelihood estimates (MLEs) and standard errors. For the HPP, $\boldsymbol{\lambda}_0$ will be elicited based on the standard errors as described in Section 4. For the GPP, the prior variance of the regression coefficients will be set equal to the squared standard errors. All analyses were performed using Hamiltonian Monte Carlo (HMC) via the `rstan` package.

5.1. *The generated data sets*

Following Ibrahim and Chen (2000), we use the ACTG036 study to generate data. The ACTG036 study was a clinical trial comparing AZT with a placebo in asymptomatic patients with hereditary coagulation disorders (hemophilia). The outcome variable was CD4 count, defined to be the number of CD4 cells (a type of white blood cell that destroys bacteria and fights off infection) per cubic millimeter of blood. Covariates included in the model were treatment ($x_{1i} = 1$ if subject i received AZT, 0 otherwise), race ($x_{2i} = 1$ if subject i was white, 0 otherwise), and age (x_{3i}), treated as a continuous variable. We assume a Poisson regression model with canonical link. The likelihood function for the current data is given by $L(\boldsymbol{\beta}|\boldsymbol{y}, \boldsymbol{X}) = \exp\left[\sum_{i=1}^{n} \left\{y_i \boldsymbol{x}_i' \boldsymbol{\beta} - \exp(\boldsymbol{x}_i' \boldsymbol{\beta})\right\}\right]$, where $n = 75$ is the sample size of the current data, y_i is the CD4 count for subject i, $\boldsymbol{x}_i = (1, x_{1i}, x_{2i}, x_{3i})'$, and $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$. Henceforth, we denote the current data by $D = \{(y_i, \boldsymbol{x}_i), i = 1, \ldots, n\}$. Historical data of size $n_0 = 50$ was generated from the same model, giving the data set $D_0 = \{(y_{0i}, \boldsymbol{x}_{0i}), i = 1, \ldots, n_0\}$. The quantities β_0, β_2, and β_3 were obtained directly from the MLEs of the ACTG036 data set. For the historical data, the treatment effect was set equal to the MLE of the ACTG036 study, that is, $\beta_1 \approx 0.048$. Two current data sets were generated: an "incompatible" current data set, where $\beta_1 = 0$, and a "compatible" data set, where $\beta_1 \approx 0.048$.

5.2. *Results*

Prior and posterior density plots for the treatment effect for the compatible and incompatible current data sets are depicted in Figure 1. The left panel depicts prior and posterior densities when the historical and current data are incompatible. The right panel shows the prior and posterior densities when the data sets are compatible. Of the four prior densities, the HPP yields the highest prior variance for the treatment effect β_1 (i.e., it is the least informative prior). When $\lambda \in \{0.75, 1.00\}$, the HPP is the only prior that places mass around the null, while the other three priors suggest that treatment is all but certain to be efficacious *a priori*. Since the CI prior, which is the most informative prior, is a special case of the HPP, we see that, for a fixed value of λ, the HPP was the most flexible prior to expressing uncertainty.

When the current data are incompatible with the historical data, i.e., when $\beta_1 = 0$ for the current data and $\beta_1 \approx 0.048$ for the historical data, the posterior distribution of β_1 utilizing the HPP is generally more

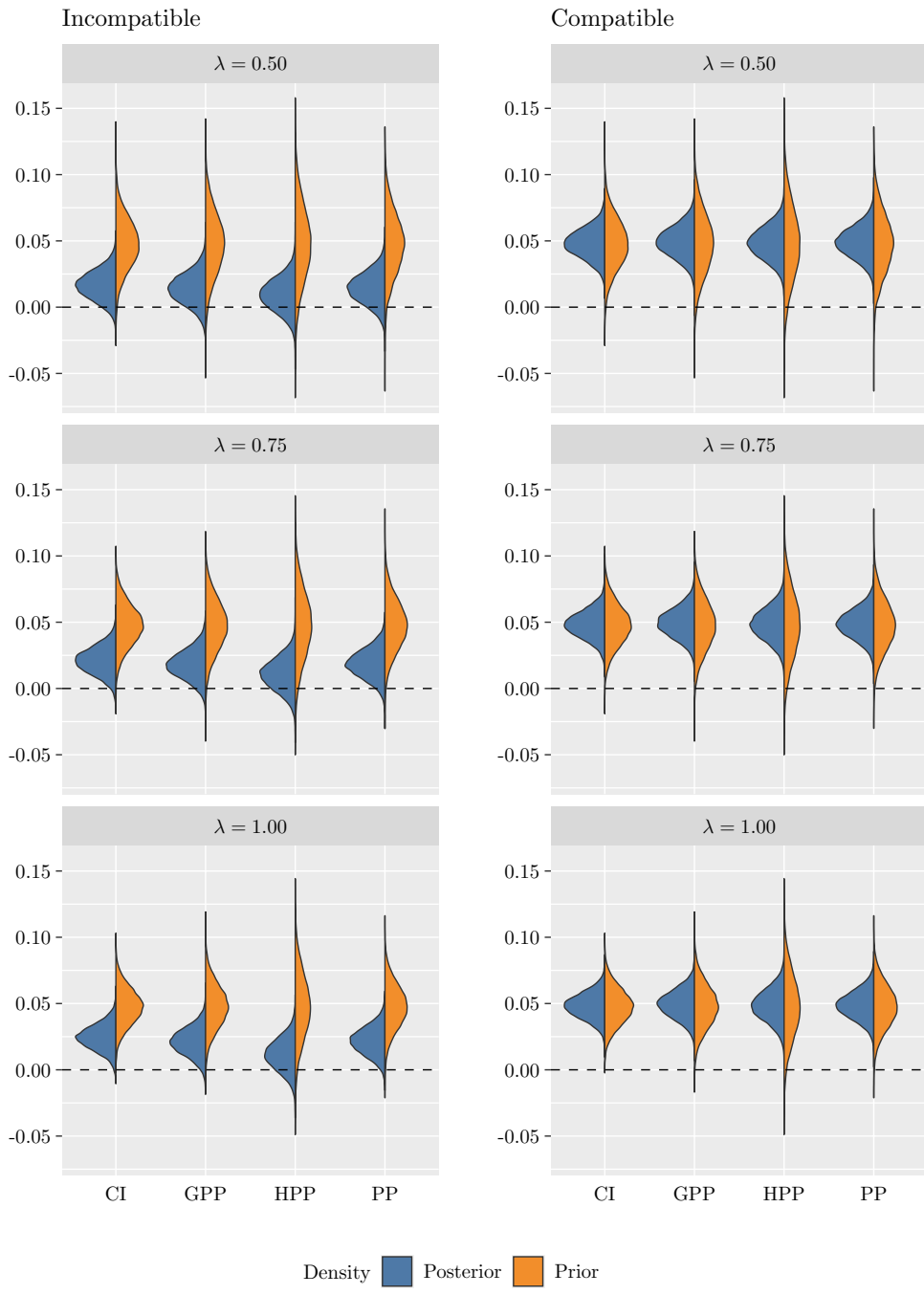

Fig. 1. Prior and posterior densities of the treatment effect for the generated data under various levels of borrowing (λ). The historical data suggest efficacy ($\beta_1 = 0.048$). The incompatible current data were generated from $\beta_1 = 0$, while that for the compatible was generated from $\beta_1 = 0.048$.

robust than its competitors (in the sense of placing the most mass on a nonpositive treatment effect). When enabling the prior to have as much influence on the posterior as the data, (i.e., when $\lambda = 1.00$), the posterior densities under the CI prior, GPP, and PP suggest treatment is all but certain to be efficacious. In contrast, the posterior density under the HPP suggests that treatment could be unbeneficial, which is the correct result.

When the two data sets are compatible, i.e., when $\beta_1 \approx 0.048$ for both the current and historical data sets, the four posterior densities are quite similar. Across the various levels of λ, the posterior densities from all four priors suggest that treatment is efficacious *a posteriori*. The posterior means across the four priors and all levels of λ are essentially the same, although the GPP yielded a somewhat higher posterior mean than the other priors. While increasing λ reduced the posterior variance markedly for the three competitor priors, the posterior under the HPP was virtually unchanged by the level of λ.

For this application, the effect of λ on the prior and the posterior densities of the treatment effect under the HPP is relatively marginal. When the data sets are incompatible, increasing the value of λ from 0.50 to 0.75 increases the posterior mean of the treatment effect by 14%. By contrast, the same change for the PP yields roughly a 29% increase, more than twice as much. This suggests that, when each component of the prior prediction for the HPP is given its own level of precision, the posterior density under the HPP is much less sensitive to increasing values of λ than the other priors.

The primary advantage of the HPP over the CI prior is that it allows practitioners to incorporate uncertainty in the prior guess in a flexible way. However, the HPP adds computational cost since the normalizing constant of the CI prior must be estimated. The GPP induces an *a priori* independent prior on the regression coefficients, which is not realistic practically. By contrast, the HPP induces a correlation structure *a priori* on the regression coefficients. The primary advantage of the PP over the HPP is that the PP is more computationally efficient. However, we have seen that the PP can be quite informative when λ is fixed, with high posterior probability that β_1 is positive in the incompatible data setting. By contrast, the posterior under the HPP suggests a nonzero chance that $\beta_1 \leq 0$ when the data sets are incompatible. Moreover, the PP is somewhat restrictive in that it requires access to a full historical data set.

6. Preposterior analysis

Although the data analysis example in Section 5 is insightful, it does not take into account Monte Carlo error in the generated data. Using the same historical data set for the prior, we generate $M = 5000$ current data sets from the Poisson regression model above. Let $\hat{\beta}_1$ denote the posterior mean of β_1. We compute bias as $\widehat{\text{bias}} = \frac{1}{M}\sum_{i=1}^{M}\left(\hat{\beta}_1^{(m)} - \beta_1^*\right)$, MSE as $\widehat{\text{MSE}} = \frac{1}{M}\sum_{i=1}^{M}\left(\hat{\beta}_1^{(m)} - \beta_1^*\right)^2$, where β_1^* is the truth, and 95% CR using the 2.5 and 97.5 percentiles of the posterior samples.

Results from the simulation exercise are presented in Table 1. When the data sets are compatible, bias is zero to three decimal places across all methods. The i.i.d. results in the highest MSE. This is intuitive since, in Section 5, it was discussed how the i.i.d. is the least informative prior. Note that the MSE under the CI prior is larger than the PP and the GPP when $n = 50$ but smaller when $n \geq 75$. This is due to the data proportionality issue, i.e., for the CI prior and HPP, the prior sample size is approximately λn whereas it is approximately λn_0 for the PP and the GPP. Note that the CR coverage across all four priors and simulation scenarios is well above 95%, which is to be expected from an informative prior that is compatible with the data. For compatible data sets, the posterior under the HPP has the lowest CR coverage by about 1–2 percentage points compared to the other priors.

However, when the data sets are incompatible, the HPP outperforms the other three priors with respect to bias, MSE, and CR coverage. The minimal 95% CR coverage probability is approximately 0.88. Conversely, the minimal CR coverage probability for the CI prior, PP, and GPP are 0.10, 0.45, and 0.47, respectively. These results agree with the data analysis results represented in Figure 1 in the main text,

where it is shown that the other three priors are quite informative, resulting in strong evidence of treatment efficacy despite the incompatibility. The relative MSE for the three competing priors is all larger than 1. The HPP is also the only prior for which the second decimal of bias is zero. To conclude, the HPP performs worse than the nonhierarchical priors when there is a little-to-no prior-data conflict but outperforms the other selected priors when the two data sets are incompatible.

Table 1. *Characteristics of the posterior under the HPP compared with other priors for compatible and noncompatible data sets*

			Compatible				Incompatible			
n	λ	Prior	Bias	Rel. MSE	Ave. CR width	CR coverage	Bias	Rel. MSE	Ave. CR width	CR coverage
50	0.50	HPP	−0.00032	1.00	0.061	0.973	0.00802	1.00	0.062	0.943
		CI	−0.00025	0.64	0.055	0.985	0.01630	1.46	0.055	0.836
		PP	−0.00026	0.61	0.054	0.986	0.01666	1.49	0.054	0.821
		GPP	−0.00025	0.62	0.054	0.985	0.01624	1.44	0.055	0.836
	0.75	HPP	−0.00031	1.00	0.060	0.974	0.00897	1.00	0.061	0.936
		CI	−0.00021	0.49	0.051	0.991	0.02086	1.91	0.051	0.677
		PP	−0.00023	0.47	0.050	0.990	0.02107	1.94	0.050	0.656
		GPP	−0.00022	0.48	0.050	0.991	0.02072	1.88	0.050	0.679
	1.00	HPP	−0.00030	1.00	0.060	0.975	0.00954	1.00	0.060	0.930
		CI	−0.00019	0.39	0.047	0.993	0.02427	2.35	0.048	0.474
		PP	−0.00020	0.37	0.046	0.995	0.02434	2.36	0.047	0.454
		GPP	−0.00020	0.38	0.047	0.994	0.02406	2.30	0.047	0.474
75	0.50	HPP	−0.00032	1.00	0.049	0.973	0.00786	1.00	0.050	0.925
		CI	−0.00026	0.64	0.044	0.985	0.01618	1.76	0.044	0.740
		PP	−0.00029	0.79	0.046	0.981	0.01241	1.32	0.047	0.851
		GPP	−0.00028	0.80	0.047	0.980	0.01198	1.27	0.047	0.861
	0.75	HPP	−0.00032	1.00	0.049	0.975	0.00882	1.00	0.049	0.915
		CI	−0.00022	0.49	0.041	0.991	0.02076	2.38	0.041	0.490
		PP	−0.00027	0.65	0.044	0.987	0.01636	1.70	0.044	0.729
		GPP	−0.00025	0.67	0.044	0.986	0.01593	1.64	0.044	0.748
	1.00	HPP	−0.00031	1.00	0.048	0.975	0.00940	1.00	0.049	0.909
		CI	−0.00020	0.39	0.038	0.995	0.02418	2.96	0.038	0.238
		PP	−0.00024	0.54	0.041	0.990	0.01947	2.10	0.042	0.561
		GPP	−0.00023	0.56	0.042	0.990	0.01908	2.03	0.042	0.589

(*Continued*)

Table 1. *Continued.*

			Compatible				Incompatible			
n	λ	Prior	Bias	Rel. MSE	Ave. CR width	CR coverage	Bias	Rel. MSE	Ave. CR width	CR coverage
100	0.50	HPP	−0.00022	1.00	0.042	0.966	0.00801	1.00	0.043	0.909
		CI	−0.00017	0.64	0.038	0.983	0.01629	1.98	0.038	0.642
		PP	−0.00022	0.90	0.041	0.970	0.01006	1.17	0.042	0.869
		GPP	−0.00021	0.92	0.042	0.970	0.00965	1.12	0.042	0.878
	0.75	HPP	−0.00021	1.00	0.042	0.969	0.00896	1.00	0.042	0.891
		CI	−0.00015	0.49	0.035	0.991	0.02085	2.71	0.035	0.324
		PP	−0.00020	0.78	0.039	0.978	0.01353	1.46	0.040	0.763
		GPP	−0.00019	0.80	0.040	0.979	0.01309	1.40	0.040	0.783
	1.00	HPP	−0.00021	1.00	0.042	0.971	0.00953	1.00	0.042	0.883
		CI	−0.00013	0.39	0.033	0.995	0.02426	3.36	0.033	0.098
		PP	−0.00019	0.68	0.038	0.985	0.01638	1.80	0.038	0.625
		GPP	−0.00017	0.69	0.038	0.983	0.01594	1.72	0.038	0.652

Rel. MSE = relative mean squared error; CR = 95% credible region.

We note that none of the selected priors were hierarchical priors. In Section 7 of the Supplementary material available at *Biostatistics* online, we conduct simulations to compare the HPP with the normalized asymptotic power prior (NAPP) (Ibrahim *and others*, 2015b) and the CP of Hobbs *and others* (2012). We find that the HPP and the NPP outperform the CP in all scenarios for the selected hyperparameters of the CP, although we acknowledge that a different set of hyperparameters could lead to a different conclusion. The NAPP had roughly half the MSE of the HPP when the data sets were compatible, but the MSE was at least 4.8 times higher than the HPP when data sets were incompatible. Furthermore, the HPP was the only prior that yielded at least 88% interval coverage through all simulation scenarios.

While the discussion thus far has been on analyzing a single parameter, we note that one of the strengths of the HPP is having a prior whose components are correlated *a priori*. To demonstrate this strength, in Section 3.2 of the Supplementary material available at *Biostatistics* online, we present simulation results predicting the mean response in a logistic regression example. The results indicate that the HPP can improve finite sample performance.

7. Discussion

We have developed the HPP for GLMs. The HPP is simple and intuitive (e.g., for logistic regression models, the hyperprior is simply a product of independent beta priors, which is a conjugate prior for each component of the response vector). However, we note that the development is flexible—any (proper) hyperprior with support on the mean of the responses may be utilized in practice. For example, for logistic regression models, we may elicit independent truncated normal priors for $\boldsymbol{m}$ over [0, 1] centered at $\boldsymbol{\mu}_0$.

The incorporation of uncertainty in the prior prediction of the mean of the responses has a natural practical application. As the prior prediction for the mean of the response typically comes in the form of historical data and/or expert opinion, there is uncertainty surrounding the elicited value. In Section 3 of the Supplementary material available at *Biostatistics* online, we show posterior quantities under the HPP and the CI prior to a logistic regression model using the data of Finney (1947) under the context where the hyperprior is formulated on the basis of expert opinion. We show that the posterior distribution under the HPP is less sensitive to the prior prediction than under the CI prior, particularly when the elicited prior prediction is incompatible with the observed data.

An area of future exploration is an HPP for time-to-event data. While the idea of eliciting a prior prediction for the mean survival time is straightforward, it is not clear how to handle the right-censored data in such a prior. In longitudinal data, the development of an HPP may be particularly useful since $\boldsymbol{m}$ could be elicited as a mean prediction for each individual. Finally, an important area of exploration is an HPP for overdispersed data. In Section 6 of the Supplementary material available at *Biostatistics* online, we conduct a Poisson analysis of negative binomial data, comparing results with frequentist approaches and the priors presented in Section 5, proposing a possible solution to handle overdispersion in count data with an HPP. We find that, in general, interval coverage is poor for all priors when analyzing negative binomial data as Poisson. However, the proposed method to handle overdispersion for the HPP outperformed other priors.

SOFTWARE AND DATA

Software for implementing all priors is available at github.com/ethan-alt/hpp. An analysis example using the Finney data is provided in Appendix 3 of the Supplementary material available at *Biostatistics* online.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

Conflict of Interest: None declared.

FUNDING

National Institute of Environmental Health Sciences (T32ES007018), in part.

REFERENCES

BEDRICK, E. J., CHRISTENSEN, R. AND JOHNSON, W. (1996). A new perspective on priors for generalized linear models. *Journal of the American Statistical Association* **91**, 1450–1460.

CHEN, M.-H. AND IBRAHIM, J. G. (2003). Conjugate priors for generalized linear models. *Statistica Sinica* **13**, 461–476.

DIACONIS, P. AND YLVISAKER, D. (1979). Conjugate priors for exponential families. *Annals of Statistics* **7**, 269–281.

EGIDI, L., PAULI, F. AND TORELLI, N. (2021). Avoiding prior–data conflict in regression models via mixture priors. *Can J Statistics* **50**, 491–510. https://doi.org/10.1002/cjs.11637

FINNEY, D. J. (1947). The estimation from individual records of the relationship between dose and quantal response. *Biometrika* **34**, 320–334.

HOBBS, B. P., SARGENT, D. J. AND CARLIN, B. P. (2012). Commensurate priors for incorporating historical information in clinical trials using general and generalized linear models. *Bayesian Analysis (Online)* **7**, 639–674.

HOFFMAN, M. D. AND GELMAN, A. (2014). The No-U-Turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research* **15**, 1593–1623.

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., GWON, Y. AND CHEN, F. (2015a). The power prior: theory and applications. *Statistics in Medicine* **34**, 3724–3749.

IBRAHIM, J. G., CHEN, M.-H., LAKSHMINARAYANAN, M., LIU, G. F. AND HEYSE, J. F. (2015b). Bayesian probability of success for clinical trials using historical data. *Statistics in Medicine* **34**, 249–264.

SABANÉS BOVÉ, D. AND HELD, L. (2011). Hyper-g priors for generalized linear models. *Bayesian Analysis* **6**, 387–410.

SCHMIDLI, H., GSTEIGER, S., ROYCHOUDHURY, S., O'HAGAN, A., SPIEGELHALTER, D. AND NEUENSCHWANDER, B. (2014). Robust meta-analytic-predictive priors in clinical trials with historical control information. *Biometrics* **70**, 1023–1032.

STAN DEVELOPMENT TEAM. (2020). *RStan: The R Interface to Stan.* R package version 2.21.2. http://mc-stan.org/.

[*Received October 13, 2021; revised May 3, 2022; accepted for publication June 9, 2022*]