Bayesian Analysis (2018) **13**, Number 3, pp. 721–748

Power-Expected-Posterior Priors for Generalized Linear Models

Dimitris Fouskakis[*], Ioannis Ntzoufras[†], and Konstantinos Perrakis[‡,†§]

Abstract. The power-expected-posterior (PEP) prior provides an objective, automatic, consistent and parsimonious model selection procedure. At the same time it resolves the conceptual and computational problems due to the use of imaginary data. Namely, (i) it dispenses with the need to select and average across all possible minimal imaginary samples, and (ii) it diminishes the effect that the imaginary data have upon the posterior distribution. These attributes allow for large sample approximations, when needed, in order to reduce the computational burden under more complex models. In this work we generalize the applicability of the PEP methodology, focusing on the framework of generalized linear models (GLMs), by introducing two new PEP definitions which are in effect applicable to any general model setting. Hyper-prior extensions for the power parameter that regulates the contribution of the imaginary data are introduced. We further study the validity of the predictive matching and of the model selection consistency, providing analytical proofs for the former and empirical evidence supporting the latter. For estimation of posterior model and inclusion probabilities we introduce a tuning-free Gibbs-based variable selection sampler. Several simulation scenarios and one real life example are considered in order to evaluate the performance of the proposed methods compared to other commonly used approaches based on mixtures of g-priors. Results indicate that the GLM-PEP priors are more effective in the identification of sparse and parsimonious model formulations.

Keywords: expected-posterior prior, g-prior, generalized linear models, hyper-g priors, imaginary data, objective Bayesian model selection, power-prior.

1 Introduction

1.1 Motivation

In this article, the variable selection problem in generalized linear models (GLMs) is analyzed from an objective and fully automatic Bayesian model choice perspective. The desire for an automatic Bayesian procedure is motivated by the appealing property of creating a method that can be easily implemented in complex models without the need

[*] Department of Mathematics, National Technical University of Athens, 15780 Athens, Greece, fouskakis@math.ntua.gr

[†] Department of Statistics, Athens University of Economics and Business, 10434 Athens, Greece, ntzoufras@aueb.gr

[‡] German Center for Neurodegenerative Diseases (DZNE), 53127 Bonn, Germany, konstantinos.perrakis@dzne.de

[§] This research has been co-financed in part by the European Union (European Social Fund-ESF) and by Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF)-Research Funding Program: Aristeia II/PEP-BVS.

https://doi.org/10.1214/17-BA1066

of specification of tuning parameters. Regarding the justification for the necessity of an objective model choice approach we can argue that in variable selection problems we are rarely confident about any given set of regressors as explanatory variables, which translates to little prior information about the regression coefficients. Therefore, we would like to consider default prior distributions, which in many cases are improper, thus leading to undetermined Bayes factors.

Intrinsic priors (Berger and Pericchi, 1996a,b) and expected-posterior (EP) priors (Pérez and Berger, 2002) can be considered as fully automatic, objective Bayesian methods for model comparison in regression models. They are developed through the utilization of the device of "training" or "imaginary" samples, respectively, of "minimal" size and therefore the resulting priors have a further advantage of being compatible across models; see Consonni and Veronese (2008). Intrinsic and EP priors have been proposed in many articles for variable selection in Gaussian linear models (see for example Casella and Moreno, 2006); however, to the best of our knowledge, there is only one study that proposes this methodology for GLMs, which is restricted to the case of the probit model (Leon-Novelo et al., 2012). We believe that this is due to the fact that derivation of such priors can be a very challenging task, especially under complex models, leading to computationally intensive solutions. Furthermore, by using minimal training samples, large sample approximations can not be applied in many cases.

Our contribution with this article is two-fold. First, we develop an automatic, objective Bayesian variable selection procedure for GLMs based on the EP prior methodology. In particular we consider the power-expected-posterior (PEP) prior of Fouskakis et al. (2015), that diminishes the effect that the imaginary data have upon the posterior distribution and therefore the need of using minimal training samples. Through this approach we can consider imaginary samples of sufficiently large size and therefore be able to apply, when needed, large sample approximations. Secondly, we introduce a simple tuning-free Gibbs-based variable selection sampler for estimating posterior model and variable inclusion probabilities.

1.2 Bayesian variable selection for generalized linear models

Despite the importance and popularity of GLMs, Bayesian variable selection techniques for non-Gaussian models are scarce in relation to the abundance of methods that are available for the normal linear model. This is mainly due to the analytical intractability which arises outside the context of the normal model. The relatively limited studies that focus on non-Gaussian models, mainly aim to overcome analytical intractability through the use of Laplace approximations and/or stochastic model search algorithms.

Chen and Ibrahim (2003) introduced a class of conjugate priors based on an initial prior prediction of the data (similar to the concept of imaginary data) associated with a scalar precision parameter. This approach essentially leads to a GLM analogue of the g prior (Zellner and Siow, 1980; Zellner, 1986) where the precision parameter has the role of g. However, the prior of Chen and Ibrahim (2003) is not analytically available for non-Gaussian GLMs and, therefore, Chen et al. (2008) proposed a Markov chain Monte Carlo (MCMC) based solution for this class of models. Ntzoufras et al.

(2003) used a unit-information g-prior (Kass and Wasserman, 1995) for variable selection and link determination in binomial models through reversible-jump MCMC sampling. Sabanés Bové and Held (2011) consider the asymptotic distribution of the prior of Chen and Ibrahim (2003), which results in the same g-prior form used in Ntzoufras et al. (2003), and further consider mixtures of g-priors along the lines of Liang et al. (2008). Computation of the marginal likelihood in Sabanés Bové and Held (2011) is handled through an integrated Laplace approximation, based on Gauss-Hermite quadrature, which allows variable selection through full enumeration for small/moderate model spaces or through MCMC model composition (MC^3) algorithms (Madigan and York, 1995) for spaces of large dimensionality. Other GLM variations of g-prior mixtures have an empirical Bayes (EB) flavor, using the observed or expected information matrix evaluated at the maximum-likelihood (ML) estimates as the prior variance-covariance matrix (Hansen and Yu, 2003; Wang and George, 2007; Li and Clyde, 2016). A computational benefit of the EB approach is that the integrated Laplace approximation can be expressed in closed form as a set of functions of the ML estimates. For large model spaces, where full enumeration is infeasible, Li and Clyde (2016) recommend using the Bayesian adaptive sampling algorithm. A relevant prior specification is the information-matrix prior of Gupta and Ibrahim (2009) which combines ideas from the g-prior and Jeffreys prior for GLMs (Ibrahim and Laud, 1991); however, in applications, Gupta and Ibrahim (2009) do not directly consider the problem of stochastic search over the entire model space. Finally, one application of Bayesian intrinsic variable selection for probit models via MCMC is presented in Leon-Novelo et al. (2012).

In this work we present an automatic, objective Bayesian variable selection procedure for GLMs based on the PEP methodology. The structure of the remainder of the paper is as follows. In Section 2 we provide an overview of the PEP prior formulation and discuss the applicability problems that arise in the case of non-Gaussian models. We proceed with two alternative definitions, which generalize the applicability of the PEP prior for GLMs. In Section 3 we introduce a Gibbs-based sampler suitable for variable selection and for single-model posterior inference. Section 4 presents an hierarchical extension of the methodology which involves assigning a hyper-prior to the power parameter that controls the contribution of the imaginary data. In Section 5 we examine the validity of certain desiderata proposed by Bayarri et al. (2012) and we proceed by presenting a general framework in Section 6 for all PEP priors under consideration. Illustrative examples and comparisons with other methods using both simulated and real life example are presented in Section 7. We conclude with a summary and a discussion of future research directions in Section 8.

2 PEP priors for generalized linear models

2.1 Model setting

We consider n realizations of a response variable Y accompanied by a set of potential predictors $X_1, X_2, \ldots, X_p$ which may characterize the response. To fix notation, let $\boldsymbol{\gamma} \in \{0,1\}^p$ index all 2^p subsets of predictors serving as a model indicator, where each element γ_j, for $j = 1, \ldots, p$, is an indicator of the inclusion of X_j in the structure of model $M_{\boldsymbol{\gamma}}$. Moreover, let $p_{\boldsymbol{\gamma}} = \sum_{j=1}^{p} \gamma_j$ denote the number of active covariates in model $M_{\boldsymbol{\gamma}}$. Within the GLM framework, the response Y follows a distribution which is a member of the exponential family. The sampling distribution of the response vector $\mathbf{y} = (y_1, \ldots, y_n)^T$ under model $M_{\boldsymbol{\gamma}}$ is given by

$$f_{\boldsymbol{\gamma}}(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi_{\boldsymbol{\gamma}}) = \exp\left(\sum_{i=1}^{n} \frac{y_i \vartheta_{\boldsymbol{\gamma}(i)} - b(\vartheta_{\boldsymbol{\gamma}(i)})}{a_i(\phi_{\boldsymbol{\gamma}})} + \sum_{i=1}^{n} c(y_i, \phi_{\boldsymbol{\gamma}}) \right). \tag{1}$$

The functions $a_i(\cdot)$, $b(\cdot)$ and $c(\cdot)$ determine the particular distribution of the exponential family. The parameter $\vartheta_{\boldsymbol{\gamma}(i)}$ is the canonical parameter which regulates the location of the distribution through the relationship $\vartheta_{\boldsymbol{\gamma}(i)} = \vartheta(\eta_{\boldsymbol{\gamma}(i)}) \equiv g \circ b'^{-1}(\eta_{\boldsymbol{\gamma}(i)})$, where $g(\cdot)$ is the link function connecting the mean of the response Y_i with the linear predictor $\eta_{\boldsymbol{\gamma}(i)} = \mathbf{X}_{\boldsymbol{\gamma}(i)} \boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $g \circ b'^{-1}(\eta_{\boldsymbol{\gamma}(i)})$ is the inverse function of $g \circ b'(\vartheta_{\boldsymbol{\gamma}(i)}) \equiv g(b'(\vartheta_{\boldsymbol{\gamma}(i)}))$. Commonly, a canonical ϑ function is used, so that $\vartheta_{\boldsymbol{\gamma}(i)} = \eta_{\boldsymbol{\gamma}(i)}$. We assume that a intercept term is included in all 2^p models under consideration, so $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is the $d_{\boldsymbol{\gamma}} \times 1$ vector of regression coefficients, where $d_{\boldsymbol{\gamma}} = p_{\boldsymbol{\gamma}} + 1$, and $\mathbf{X}_{\boldsymbol{\gamma}(i)}$ is the i–th row of the $n \times d_{\boldsymbol{\gamma}}$ design matrix $\mathbf{X}_{\boldsymbol{\gamma}}$ with a vector of 1's in the first column and the $\boldsymbol{\gamma}$–th subset of the $\boldsymbol{X}_j$'s in the remaining $p_{\boldsymbol{\gamma}}$ columns. The parameter $\phi_{\boldsymbol{\gamma}}$ controls the dispersion and the function $a_i(\cdot)$ is typically of the form $a_i(\phi_{\boldsymbol{\gamma}}) = \phi_{\boldsymbol{\gamma}}/w_i$, where w_i is a known fixed weight that may either vary or remain constant per observation. In addition, the nuisance parameter $\phi_{\boldsymbol{\gamma}}$ is commonly considered as a common parameter across models, therefore we assume throughout that $\phi_{\boldsymbol{\gamma}} \equiv \phi$ without loss of generality. Given the above formulation, we have that $\mathrm{E}(Y_i) = b'(\vartheta_{\boldsymbol{\gamma}(i)})$ and $\mathrm{Var}(Y_i) = b''(\vartheta_{\boldsymbol{\gamma}(i)}) a_i(\phi)$.

The GLM parameters $\boldsymbol{\theta}_{\boldsymbol{\gamma}} = (\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi)$ are divided into the predictor effects $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and the parameter ϕ which affects dispersion. In the following we work along the lines of Fouskakis and Ntzoufras (2016) considering the conditional PEP prior; i.e. we construct the PEP prior of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ conditional on ϕ.

2.2 An overview of the PEP prior

The PEP prior, initially formulated in Fouskakis et al. (2015) for the case of the normal linear model, fuses ideas from the power prior (Ibrahim and Chen, 2000) and the EP prior (Pérez and Berger, 2002). Let us first describe the EP prior approach. Consider that we have imaginary data $\mathbf{y}^* = (y_1^*, \ldots, y_{n^*}^*)^T$ coming from the prior-predictive distribution $m^*(\mathbf{y}^*)$ of a "suitable" *reference* model M^*. Then, given $\mathbf{y}^*$, for any model $M_{\boldsymbol{\gamma}}$ with sampling distribution $f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi)$ as defined in (1) and a default *baseline prior* of the form $\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi) = \pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\phi) \pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\phi)$, we have a corresponding *baseline posterior* distribution given by

$$\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi|\mathbf{y}^*) = \frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi) \pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\phi) \pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\phi)}{m_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^*)}, \tag{2}$$

where $m_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^*)$ is the normalizing constant of the baseline posterior distribution under model $M_{\boldsymbol{\gamma}}$. The EP prior for the parameters of model $M_{\boldsymbol{\gamma}}$ is then defined as the posterior distribution in (2), averaged over all possible imaginary samples, i.e.

$$\pi^{\mathrm{EP}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi) = \int \pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi|\mathbf{y}^*)\, m^*(\mathbf{y}^*)\mathrm{d}\mathbf{y}^* \,. \tag{3}$$

The reference model M^* is commonly considered to be the simplest model, i.e. the (null) intercept model in the regression framework. This selection makes the EP approach essentially equivalent to the arithmetic intrinsic Bayes factor of Berger and Pericchi (1996b).

A key issue in the implementation of the EP prior is the selection of the size n^* of the imaginary sample. In order to minimize the effect of the prior on posterior inference, the reasonable solution is to choose the smallest possible n^* for which the posterior (2) is proper. This leads to the concept of the so-called *minimal training sample*. When it comes to regression a problem arises with the design matrix as one has to choose appropriate covariate values for each minimal training sample. This requires calculating summaries of Bayes factors over all possible minimal training samples which further complicates the problem. Therefore, under the EP prior, computation of the Bayes factors require calculations over all possible configurations of the design matrix for each minimal training sample (Pérez, 1998) or, at least, calculations over an efficiently large number of random sub-samples of all possible configurations (Fouskakis and Ntzoufras, 2013). An alternative and simpler computational solution has been proposed by Casella and Moreno (2006) and Moreno and Girón (2008), however, this solution is only applicable under the normal linear regression model. Additionally, under this approach, it is not clear whether the resulting Bayes factors retain their intrinsic nature. Furthermore, the effect of the EP prior can become influential when the sample size is not much larger than the total number of predictors; see Fouskakis et al. (2015) for details. Finally, when n^* is small and (3) is hard to derive, large sample approximations cannot be applied.

The PEP prior resolves the problem of defining and averaging over minimal training samples and at the same time scales down the effect of the imaginary data on the posterior distribution. The core idea lies in substituting the likelihood function involved in the calculation of (3) by a powered-version of it, i.e. raising it to the power of $1/\delta$, similar to the power prior approach of Ibrahim and Chen (2000). Following Fouskakis and Ntzoufras (2016), the conditional PEP (PCEP) prior in the GLM setup, under the null-reference model M_0, is defined as follows

$$\pi^{\mathrm{PEP}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi|\delta) = \pi^{\mathrm{PEP}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\phi, \delta)\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\phi), \tag{4}$$

where

$$\begin{aligned}
\pi^{\mathrm{PEP}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\phi, \delta) &= \int \pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathbf{y}^*, \phi, \delta) m^{\mathrm{N}}_0(\mathbf{y}^*|\phi, \delta)\mathrm{d}\mathbf{y}^*, &(5)\\
\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathbf{y}^*, \phi, \delta) &= \frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi, \delta)\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}\,|\phi)}{m^{\mathrm{N}}_{\boldsymbol{\gamma}}(\mathbf{y}^*|\phi, \delta)}, &(6)\\
m^{\mathrm{N}}_{\boldsymbol{\gamma}}(\mathbf{y}^*|\phi, \delta) &= \int f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi, \delta)\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}\,|\phi)\mathrm{d}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, &(7)\\
f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi, \delta) &= \frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi)^{1/\delta}}{k_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi, \delta)}, &(8)
\end{aligned}$$

$$m_0^{\mathrm{N}}(\mathbf{y}^*|\phi,\delta) = \int f_0(\mathbf{y}^*|\beta_0,\phi,\delta)\pi_0^{\mathrm{N}}(\beta_0\,|\phi)\mathrm{d}\beta_0, \tag{9}$$

$$f_0(\mathbf{y}^*|\beta_0,\phi,\delta) = \frac{f_0(\mathbf{y}^*|\beta_0,\phi)^{1/\delta}}{k_0(\beta_0,\phi,\delta)}. \tag{10}$$

For the original PEP prior of Fouskakis et al. (2015), we consider the choice $k_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}},\phi,\delta) = \int f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}},\phi)^{1/\delta}d\mathbf{y}^*$ for all models $\boldsymbol{\gamma}\in\mathcal{M}$. Under this choice, the PEP prior of the intercept β_0 of the reference M_0 reduces to the baseline prior; i.e. $\pi_0^{\mathrm{PEP}}(\beta_0|\phi,\delta) = \pi_0^{\mathrm{N}}(\beta_0|\phi)$. The selection of $k_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}},\phi,\delta)$ and $k_0(\beta_0,\phi,\delta)$ is further discussed in Section 2.3.

Here the power parameter δ controls the weight that the imaginary data contribute to the "final" posterior distributions of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and ϕ. As noted in Fouskakis et al. (2015), the choice of $\delta = n^*$ leads to a minimally-informative prior with a unit-information interpretation (Kass and Wasserman, 1995) where the contribution of the imaginary data is down-weighted to account overall for one data point. Furthermore, by setting $n^* = n$ we avoid the complicated problem of sampling over numerous imaginary design sub-matrices, as in this case we have that $\mathbf{X}^*_{\boldsymbol{\gamma}} \equiv \mathbf{X}_{\boldsymbol{\gamma}}$. Under this framework, the unit-information property in combination with the empirical evidence presented in Fouskakis et al. (2015) suggest that the PEP prior is robust with respect to the specification of n^* and it also remains relatively non-informative even when the model dimensionality is close to the sample size.

Another advantage of setting $n^* = n$, which becomes more obvious in the GLM framework, is that one can now utilize large-sample approximations when needed for large n. For instance, consider the baseline posterior in (6), which can be expressed as

$$\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathbf{y}^*,\phi,\delta) \propto \exp\left(\sum_{i=1}^{n^*}\frac{y_i^*\vartheta_{\boldsymbol{\gamma}(i)} - b(\vartheta_{\boldsymbol{\gamma}(i)})}{\delta a_i(\phi)}\right)\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\phi). \tag{11}$$

This unnormalized distribution is recognized as the power prior for GLMs (Chen et al., 2000). Assuming a flat baseline prior for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, i.e. $\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\phi) \propto 1$, then, based on standard Bayesian asymptotic theory (Bernardo and Smith, 2000), for $n^* \to \infty$ the distribution in (11) converges to $\widehat{\pi}_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathbf{y}^*,\phi,\delta) \approx \mathrm{N}_{d_{\boldsymbol{\gamma}}}(\widehat{\boldsymbol{\beta}}^*_{\boldsymbol{\gamma}}, \delta\boldsymbol{J}^*_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}^*_{\boldsymbol{\gamma}})^{-1})$, where $\widehat{\boldsymbol{\beta}}^*_{\boldsymbol{\gamma}}$ is the ML estimate of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ for data $\mathbf{y}^*$ and design matrix $\mathbf{X}^*_{\boldsymbol{\gamma}}$, and $\boldsymbol{J}^*_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}^*_{\boldsymbol{\gamma}})$ is the observed information matrix evaluated at $\widehat{\boldsymbol{\beta}}^*_{\boldsymbol{\gamma}}$. Specifically, $\boldsymbol{J}^*_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}^*_{\boldsymbol{\gamma}}) = (\mathbf{X}^{*T}_{\boldsymbol{\gamma}}\mathbf{W}^*_{\boldsymbol{\gamma}}\mathbf{X}^*_{\boldsymbol{\gamma}})^{-1}$, with $\mathbf{W}^*_{\boldsymbol{\gamma}} = \mathrm{diag}(w^*_{\boldsymbol{\gamma}(i)})$, $w^*_{\boldsymbol{\gamma}(i)} = (\frac{\partial\mu_{\boldsymbol{\gamma}(i)}}{\partial\eta_{\boldsymbol{\gamma}(i)}})^2[a_i(\phi)b''(\vartheta_{\boldsymbol{\gamma}(i)})]^{-1}$ and $\mu_{\boldsymbol{\gamma}(i)} = b'(\vartheta_{\boldsymbol{\gamma}(i)})$. It is straightforward to see that the asymptotic distribution has a g-prior form according to the definitions for GLMs presented in Ntzoufras et al. (2003) and Sabanés Bové and Held (2011). The familiar zero-mean representation arises when the covariates are centered around their corresponding arithmetic mean and the imaginary response data are all the same, i.e. $\mathbf{y}^* = g^{-1}(0)\mathbf{1}_{n^*}$, where $\mathbf{1}_{n^*}$ is a vector of ones of size n^* since in this case we have that $\widehat{\boldsymbol{\beta}}^*_{\boldsymbol{\gamma}} = \mathbf{0}_{d_{\boldsymbol{\gamma}}}$; for details see Ntzoufras et al. (2003).

2.3 PEP prior extensions for GLMs via unnormalized power likelihoods

The sampling distribution of the imaginary data involved in the PEP prior via (6), (7) and (9) is a power version of the likelihood function. In the normal linear regression case Fouskakis et al. (2015) and Fouskakis and Ntzoufras (2016) naturally considered $k_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi, \delta) = \int f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \phi)^{1/\delta} \mathrm{d}\mathbf{y}^*$, i.e. the density normalized power likelihood

$$f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \phi, \delta) = \frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \phi)^{1/\delta}}{\int f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \phi)^{1/\delta} \mathrm{d}\mathbf{y}^*}, \tag{12}$$

which is also a normal distribution with variance inflated by a factor of δ. Similar results can be derived for specific distributions of the exponential family such as the Bernoulli, the exponential and the beta, where the normalized power likelihood is of the same distributional form. This property simplifies calculations when using the PEP methodology, especially for Gaussian models where the resulting posterior distribution and marginal likelihood are available in closed form. An application of the PEP prior using the normalized power likelihood for MCMC-based variable selection in binary logistic regression can be found in Perrakis et al. (2015).

However, this property does not hold for all members of the exponential family. For instance, for the binomial and Poisson regression models, the normalized power likelihoods are composed by products of discrete distributions that have no standard form. Although it is feasible to perform likelihood evaluations for each observation, the additional computational burden renders the implementation of the PEP prior methodology time-consuming and inefficient. One possible computational solution to the problem would be to utilize an exchange-rate algorithm for doubly-intractable distributions (Murray et al., 2006). However, this approach would further increase MCMC computational costs.

Here we pursue a more generic approach for the implementation of PEP methodology in GLMs by redefining the prior itself. Namely, we consider two adaptations of the PEP prior which, in principle, can be applied to any statistical model and, consequently, are applicable to all members of the exponential family. For the remainder of this paper, without loss of generality we restrict the scale parameter ϕ to be known, which is the case for the binomial, Poisson and normal with known error variance regression models. Moreover, in order to alleviate notation we remove ϕ from all conditional expressions in the following of the paper.

The core idea is to use the unnormalized power likelihood (8) and (10), i.e. set $k_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \delta) = k_0(\beta_0, \delta) = 1$, and normalize the baseline posterior density (11) resulting in

$$\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathbf{y}^*, \delta) = \frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta} \pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})}{\int f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta} \pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \mathrm{d}\boldsymbol{\beta}_{\boldsymbol{\gamma}}} \tag{13}$$

and accordingly for the reference model M_0. This is also the approach of Friel and Pettitt (2008, Eq. 4) in the definition of the power posteriors. Given this first step, we proceed by proposing two versions of the PEP prior which differentiate with respect to the definition of the prior predictive distribution used to average the baseline posterior in (13) across imaginary data sets. This prior predictive distribution can be alternatively viewed as a hyper-prior assigned to $\mathbf{y}^*$ (Fouskakis and Ntzoufras, 2016). More specifically we define the two PEP variants as follows.

Definition 1. The **concentrated-reference PEP prior** of model parameters $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is defined as the power posterior of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ in (13) "averaged" over all imaginary data coming from the prior predictive distribution of the reference model M_0 based on the actual likelihood, that is

$$\pi_{\boldsymbol{\gamma}}^{\mathrm{CR-PEP}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\delta) = \mathbb{E}_{\mathbf{y}^*}^{m_0^{\mathrm{N}}}\left[\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathbf{y}^*,\delta)\right] = \pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \int \frac{m_0^{\mathrm{N}}(\mathbf{y}^*)}{m_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^*|\delta)} f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta} \mathrm{d}\mathbf{y}^* \tag{14}$$

$$\text{with } m_0^{\mathrm{N}}(\mathbf{y}^*) = \int f_0(\mathbf{y}^*|\beta_0)\pi_0^{\mathrm{N}}(\beta_0)\mathrm{d}\beta_0 \tag{15}$$

$$\text{and } m_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^*|\delta) = \int f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta}\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})\mathrm{d}\boldsymbol{\beta}_{\boldsymbol{\gamma}}.$$

In order for the above prior to exist we need to consider for each model $M_{\boldsymbol{\gamma}}$ similar assumptions as in Pérez and Berger (2002), i.e.

$$0 < m_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^*|\delta) < \infty, \quad 0 < \int \frac{m_0^{\mathrm{N}}(\mathbf{y}^*)}{m_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^*|\delta)} f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta}\mathrm{d}\mathbf{y}^* < \infty. \tag{16}$$

In (14), m_0^{N} will not necessarily be proper, but still, by slightly abusing notation we define the concentrated-reference PEP prior as the expectation of $\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathbf{y}^*,\delta)$ with respect to m_0^{N}. Furthermore, impropriety of the baseline priors in (14) causes no indeterminacy of the resulting Bayes factors, since $\pi_{\boldsymbol{\gamma}}^{\mathrm{CR-PEP}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\delta)$ depends only on the normalizing constant of the baseline prior of the parameter of the null model. Finally, the concentrated-reference PEP prior for the parameter of the null model is no longer equal to the baseline prior $\pi_0^{\mathrm{N}}(\beta_0)$, since

$$\pi_0^{\mathrm{CR-PEP}}(\beta_0|\delta) = \pi_0^{\mathrm{N}}(\beta_0)\int \frac{m_0^{\mathrm{N}}(\mathbf{y}^*)}{m_0^{\mathrm{N}}(\mathbf{y}^*|\delta)} f_0(\mathbf{y}^*|\beta_0)^{1/\delta}\mathrm{d}\mathbf{y}^*. \tag{17}$$

Definition 2. The **diffuse-reference PEP prior** of model parameters $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ is defined as the power posterior of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ in (13) "averaged" over all imaginary data coming from the "normalized" prior predictive distribution of the reference model M_0 based on the unnormalized power likelihood, that is

$$\pi_{\boldsymbol{\gamma}}^{\mathrm{DR-PEP}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\delta) = \mathbb{E}_{\mathbf{y}^*|\delta}^{m_0^{\mathrm{Z}}}\left[\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathbf{y}^*,\delta)\right] = \pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \int \frac{m_0^{\mathrm{Z}}(\mathbf{y}^*|\delta)}{m_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^*|\delta)} f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta}\mathrm{d}\mathbf{y}^* \tag{18}$$

$$\text{with } m_0^{\mathrm{Z}}(\mathbf{y}^*|\delta) = \frac{m_0^{\mathrm{N}}(\mathbf{y}^*|\delta)}{\int m_0^{\mathrm{N}}(\mathbf{y}^*|\delta)\mathrm{d}\mathbf{y}^*} = \frac{\int f_0(\mathbf{y}^*|\beta_0)^{1/\delta}\pi_0^{\mathrm{N}}(\beta_0)\mathrm{d}\beta_0}{\int\int f_0(\mathbf{y}^*|\beta_0)^{1/\delta}\pi_0^{\mathrm{N}}(\beta_0)\mathrm{d}\beta_0\mathrm{d}\mathbf{y}^*}$$

$$\text{and } m_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^*|\delta) = \int f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta}\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})\mathrm{d}\boldsymbol{\beta}_{\boldsymbol{\gamma}}.$$

The conditions for the existence of the diffuse-reference PEP prior, for each model $M_{\boldsymbol{\gamma}}$, are similar to (16), i.e.

$$0 < m_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^*|\delta) < \infty, \quad 0 < \int \frac{m_0^{\mathrm{N}}(\mathbf{y}^*|\delta)}{m_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^*|\delta)} f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta} \mathrm{d}\mathbf{y}^* < \infty. \tag{19}$$

Again the definition of the diffuse-reference PEP prior as an expectation of $\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathbf{y}^*, \delta)$ with respect to m_0^{Z} is slightly abusive under improper baseline prior setups. The normalization of $m_0^{\mathrm{N}}(\mathbf{y}^*|\delta)$ is adopted in order to retain the "expected-posterior" interpretation under proper baseline prior setups. The induced normalizing constant

$$\mathcal{C}_0 = \int m_0^{\mathrm{N}}(\mathbf{y}^*|\delta)\mathrm{d}\mathbf{y}^* = \int \left\{ \int f_0(\mathbf{y}^*|\beta_0)^{1/\delta} \mathrm{d}\mathbf{y}^* \right\} \pi_0^{N}(\beta_0)\mathrm{d}\beta_0$$

exists under any proper baseline prior setup and has no effect on the posterior variable selection measures since it is common in all models under consideration. Additionally, impropriety of the baseline priors causes no indeterminacy of the resulting Bayes factors, since $\pi_{\boldsymbol{\gamma}}^{\mathrm{DR-PEP}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\delta)$ depends only on $\mathcal{C}_0$ which is common across all models. Note that the corresponding normalization is not needed for the concentrated-reference PEP since it will be equal to the normalizing constant of the prior and therefore equal to one for proper prior distributions. Finally, the diffuse-reference PEP prior for the parameter of the null model is no longer equal to the baseline prior, since

$$\pi_0^{\mathrm{DR-PEP}}(\beta_0|\delta) = \pi_0^{\mathrm{N}}(\beta_0) \frac{\int f_0(\mathbf{y}^*|\beta_0)^{1/\delta} \mathrm{d}\mathbf{y}^*}{\mathcal{C}_0} = \frac{\int f_0(\mathbf{y}^*|\beta_0)^{1/\delta} \pi_0^{\mathrm{N}}(\beta_0)\mathrm{d}\mathbf{y}^*}{\int\int f_0(\mathbf{y}^*|\beta_0)^{1/\delta} \pi_0^{\mathrm{N}}(\beta_0)\mathrm{d}\beta_0 \mathrm{d}\mathbf{y}^*}.$$

Definition 1 can be considered as a special case of Definition 2 since $m_0^{\mathrm{N}}(\mathbf{y}^*)$ is given by $m_0^{\mathrm{N}}(\mathbf{y}^*|\delta)$ with $\delta = 1$. Because the likelihood in (15) is not scaled down, it provides more information from the imaginary data resulting in a more concentrated (in relation to the alternative approach) predictive distribution. For this reason, this version is named *concentrated-reference* PEP (CR-PEP). The CR-PEP prior (14) is also given by

$$\pi_{\boldsymbol{\gamma}}^{\mathrm{CR-PEP}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\delta) \quad = \quad \pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \int\int \frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta} f_0(\mathbf{y}^*|\beta_0)}{m_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^*|\delta)} \pi_0^{\mathrm{N}}(\beta_0)\mathrm{d}\mathbf{y}^*\mathrm{d}\beta_0. \tag{20}$$

In Definition 2 the likelihood involved in $m_0^{\mathrm{N}}(\mathbf{y}^*|\delta)$ in (18) is raised to the power of $1/\delta$ and, therefore, the information incorporated in the prior predictive distribution becomes equal to n^*/δ points leading to a distribution which becomes increasingly diffuse as δ grows. Thus, this prior is coined as the *diffuse-reference* PEP (DR-PEP). Specifically, we have that

$$\pi_{\boldsymbol{\gamma}}^{\mathrm{DR-PEP}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\delta) = \mathcal{C}_0^{-1} \pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \int\int \frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta} f_0(\mathbf{y}^*|\beta_0)^{1/\delta}}{m_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^*|\delta)} \pi_0^{\mathrm{N}}(\beta_0)\mathrm{d}\mathbf{y}^*\mathrm{d}\beta_0. \tag{21}$$

In the normal regression case, the DR-PEP prior proposed here coincides with the conditional prior formulation of Fouskakis and Ntzoufras (2016), namely the PCEP prior. Assuming a Zellner's g-prior as baseline prior for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ with dispersion parameter $g = g_0$ and a reference baseline prior for the variance parameter $\pi(\sigma^2) \propto \sigma^{-2}$, then the DR-PEP is given by

$$\pi_{\boldsymbol{\gamma}}^{\mathrm{DR-PEP}}\big(\boldsymbol{\beta}_{\boldsymbol{\gamma}} | \, \sigma^2 \, \delta, \mathbf{X}_\ell\big) = \mathrm{N}_{d_{\boldsymbol{\gamma}}}\big(\mathbf{0}, \, \mathbf{V}_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}} \sigma^2\big), \tag{22}$$

where $w = g_0/(g_0 + \delta)$, $\mathbf{V}_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}} = \delta \left(\mathbf{X}_{\boldsymbol{\gamma}}^T \left[w^{-1}\mathbf{I}_n - (\delta\boldsymbol{\Lambda}_0 + w\mathbf{H}_\gamma)^{-1}\right] \mathbf{X}_{\boldsymbol{\gamma}}\right)^{-1}$ and $\boldsymbol{\Lambda}_0 = \delta^{-1}\left(\mathbf{I}_n - \frac{w}{n}\mathbf{1}_n\mathbf{1}_n^T\right)$, $\mathbf{H}_\gamma = \mathbf{X}_{\boldsymbol{\gamma}}(\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{X}_{\boldsymbol{\gamma}})^{-1}\mathbf{X}_{\boldsymbol{\gamma}}^T$. The CR-PEP prior has the same form as the DR-PEP in (22), differing only with respect to the variance-covariance matrix as in this case we have that $\boldsymbol{\Lambda}_0 = \mathbf{I}_n - \frac{g_0}{g_0+1}n^{-1}\mathbf{1}_n\mathbf{1}_n^T$. Both approaches lead to a consistent variable selection procedure for normal regression models; details are provided in Fouskakis et al. (2016).

Example. Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ be a random sample from the exponential distribution with mean λ and variance λ^2. Consider the hypothesis $H_0 : \lambda = \lambda_0$ versus $H_1 : \lambda \neq \lambda_0$. The baseline (reference) prior under H_1 is $\pi_1^{\mathrm{N}}(\lambda) \propto \lambda^{-1}$. Let $\boldsymbol{y}^* = (y_1^*, \ldots, y_{n^*})^T$ be a training (imaginary) sample of size n^*: $1 \leq n^* \leq n$. Under the null hypothesis as reference model M_0, the marginal likelihood under the baseline prior is $m_0^{\mathrm{N}}(\mathbf{y}^*|\delta) \propto \lambda_0^{-(n/\delta)}$ $\exp(-(\lambda_0\delta)^{-1}\sum_{i=1}^n y_i^*)$. For the CR-PEP we consider $m_0^{\mathrm{N}}(\mathbf{y}^*|\delta = 1)$, while for the DR-PEP we consider the density normalized version of $m_0^{\mathrm{N}}(\mathbf{y}^*|\delta)$, denoted by $m_0^{\mathrm{Z}}(\mathbf{y}^*|\delta)$; see Definition 2. The baseline posterior distribution of λ, under the baseline prior of H_1, for the CR/DR-PEP methods is $\lambda|\mathbf{y}^* \sim \text{Inv-Gamma}(n\delta^{-1}, \delta^{-1}\sum_{i=1}^n y_i^*)$, while for the original PEP (with the density normalized power likelihood) we have that $\lambda|\mathbf{y}^* \sim$ $\text{Inv-Gamma}(n, \delta^{-1}\sum_{i=1}^n y_i^*)$. The resulting PEP, CR-PEP and DR-PEP prior distributions are $\lambda/\lambda_0 \sim \mathrm{B}'(n^*, n^*)$, $\lambda/(\delta\lambda_0) \sim \mathrm{B}'(n^*, n^*\delta^{-1})$ and $\lambda/\lambda_0 \sim \mathrm{B}'(n^*, n^*\delta^{-1})$, respectively. Here $\mathrm{B}'(a, b)$ denotes the beta prime distribution with p.d.f. $f(x) = x^{a-1}(1 + x)^{-a-b}/B(a, b)$, where $a > 0$ and $b > 0$. Under this scenario the EP prior coincides with the original PEP prior. Table 1 presents the prior mean and variance, under the alternative hypothesis, for the different PEP formulations. For fixed values of δ, the variance of λ under the PEP and DR-PEP priors shrinks to zero as n^* grows. Therefore, for large n^*, the prior distributions degenerate to a point mass distribution on λ_0 with probability equal to one. Note that the mean and the variance of λ under the CR/DR-PEP priors are not defined for the default choice of $\delta = n^*$. For finite prior variances, the DR-PEP prior is more dispersed than the CR-PEP prior for any $\delta > 1$ since $\mathrm{Var}(\lambda \,|\, \text{DR-PEP}) = \delta^2 \, \mathrm{Var}(\lambda \,|\, \text{CR-PEP})$. Finally, when $\delta = an^* < n^*/2$, the prior variance of the CR-PEP converges to $a(1 - 2a)^{-1}(1 - a)^{-2}$ for large n^*, while the corresponding variance of the DR-PEP grows with the same rate as ${n^*}^2$.

2.4 Further prior specifications

To complete the model formulation we need to specify a baseline prior for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and also a prior distribution for $\boldsymbol{\gamma}$. In our setting we do not need to specify a prior for ϕ, which is considered known. For settings with random ϕ, common across all models, we propose working along the lines of Fouskakis and Ntzoufras (2016) using a flat prior on ϕ.

Standard options for the baseline prior of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ are either the flat prior $\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \propto 1$ or Jeffreys prior for GLMs (Ibrahim and Laud, 1991) which is of the form $\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \propto$

Prior	Mean ($n^* > \delta$)	Variance ($n^* > 2\delta$)
EP & PEP prior	$\frac{n^*}{n^*-1}\lambda_0$	$\frac{n^*(2n^*-1)}{(n^*-1)^2(n^*-2)}\lambda_0^2$
CR-PEP prior	$\frac{n^*}{n^*-\delta}\lambda_0$	$\frac{1}{n^*}\frac{1+\delta-\delta/n^*}{(1-\delta/n^*)^2(1-2\delta/n^*)}\lambda_0^2$
DR-PEP prior	$\frac{\delta n^*}{(n^*-\delta)}\lambda_0$	$\frac{1}{n^*}\frac{1+\delta-\delta/n^*}{(1-\delta/n^*)^2(1-2\delta/n^*)}\delta^2\lambda_0^2$

Table 1: Prior mean and variance, under the alternative hypothesis, for the exponential case for different PEP variations.

$|\mathbf{X}_{\boldsymbol{\gamma}}^T\mathbf{W}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})\mathbf{X}_{\boldsymbol{\gamma}}|^{1/2}$. For non-Gaussian GLMs, Jeffreys prior will depend on $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ through the matrix $\mathbf{W}_{\boldsymbol{\gamma}}(\cdot)$; see Section 2.2 for details. Note that Jeffreys prior for the parameter of the null model simplifies to $\pi_0^{\mathrm{N}}(\beta_0) \propto \mathrm{tr}(\mathbf{W}_0(\beta_0))^{1/2}$. Concerning $\boldsymbol{\gamma}$ the usual option is a product Bernoulli distribution where the prior inclusion probability of each predictor is equal to 0.5. This leads to a discrete uniform prior on model space, i.e. $\pi(\boldsymbol{\gamma}) = 2^{-p}$. An alternative choice better suited for moderate to large p, accounting for an appropriate multiplicity adjustment (Scott and Berger, 2010), is to use a hierarchical prior where the inclusion probability of each predictor is uniformly distributed so that $\pi(\boldsymbol{\gamma}) = (p+1)^{-1}\binom{p}{p_{\boldsymbol{\gamma}}}^{-1}$.

3 Posterior inference

3.1 Posterior distribution under the PEP prior

In normal linear regression models, the conditional PEP prior is a conjugate normal-inverse gamma distribution which leads to fast and efficients computations (Fouskakis and Ntzoufras, 2016). For non-Gaussian GLMs, the resulting PEP has no convenient conjugate formulation and therefore the integrals involved in the derivation of the corresponding posteriors are intractable. However, one can work with the hierarchical formulation, i.e. without marginalizing over the imaginary data, and use an MCMC algorithm in order to sample from the joint posterior distribution of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\mathbf{y}^*$.

For ease of exposition, for the remainder of this section we use the indicator ψ to distinguish between the CR-PEP prior ($\psi = 1$) and the DR-PEP prior ($\psi = \delta$) and simply use PEP to denote the joint posterior. Specifically, from (13), (14) and (18) we have the following hierarchical form

$$\pi_{\boldsymbol{\gamma}}^{\mathrm{PEP}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \mathbf{y}^*|\mathbf{y}, \delta) \propto f_{\boldsymbol{\gamma}}(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}})\frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta}\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})}{m_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^*|\delta)}m_0^{\mathrm{N}}(\mathbf{y}^*|\psi). \tag{23}$$

A computational problem arises in (23) related with the evaluation of the prior predictive distributions $m_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^*|\delta)$ and $m_0^{\mathrm{N}}(\mathbf{y}^*|\psi)$ which are not available in closed form. A solution can be obtained by using the Laplace approximation for both quantities. An empirical evaluation of the accuracy of the log-marginal likelihood is provided at the Appendix C, available at the supplementary material of this manuscript (Fouskakis et al., 2017).

Alternatively, a more accurate solution can be obtained by augmenting the parameter space further and include β_0 of M_0 in the joint posterior, thus avoiding to use an approximation of $m_0^{\mathrm{N}}(\mathbf{y}^*|\psi)$. Based on (20) and (21) the posterior in (23) is expanded as

$$\pi_{\boldsymbol{\gamma}}^{\mathrm{PEP}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \beta_0, \mathbf{y}^*|\mathbf{y}, \delta) \propto f_{\boldsymbol{\gamma}}(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta} \pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})}{m_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^*|\delta)} f_0(\mathbf{y}^*|\beta_0)^{1/\psi} \pi_0^{\mathrm{N}}(\beta_0), \tag{24}$$

which leaves us with the need of using only one Laplace approximation for $m_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^*|\delta)$. Sampling from (24) for a model $M_{\boldsymbol{\gamma}}$ is feasible using Metropolis-Hastings (M-H) within Gibbs sampling. Note that under flat baseline priors the posterior in (24) and the corresponding MCMC scheme are simplified. For variable selection, which is the topic of the next section, we further assign a prior on $\boldsymbol{\gamma}$, based on the options discussed in Section 2.4.

3.2 Gibbs variable selection under the PEP prior

The Gibbs variable selection (GVS; Dellaportas et al., 2002) method utilizes the vector of binary indicators $\boldsymbol{\gamma} \in \{0,1\}^p$ which partitions the regression vector $\boldsymbol{\beta}$ into $(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}})$, corresponding to those components of $\boldsymbol{\beta}$ that are included and excluded from the model, i.e. $\beta_j \in \boldsymbol{\beta}_{\boldsymbol{\gamma}}$ if $\gamma_j = 1$ and $\beta_j \in \boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}$ if $\gamma_j = 0$, for $j = 1, \ldots, p$. As the intercept term is always included, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}$ are of dimensionality $d_{\boldsymbol{\gamma}} = p_{\boldsymbol{\gamma}} + 1$ and $d_{\setminus\boldsymbol{\gamma}} = p - p_{\boldsymbol{\gamma}}$, respectively. The joint prior of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is specified as

$$\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta})\pi(\boldsymbol{\gamma}) = \pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}})\pi(\boldsymbol{\gamma}), \tag{25}$$

where $\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}})$ is just a *pseudo-prior* used to retain the dimensionality balance across different models. Suitable choices for the priors of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}$ have been discussed in Section 2.4, thus, we only need to specify the pseudo-prior and propose using $\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}) = \mathrm{N}_{d_{\setminus\boldsymbol{\gamma}}}(\boldsymbol{\beta}^*_{\setminus\boldsymbol{\gamma}}, \mathbf{I}_{d_{\setminus\boldsymbol{\gamma}}}\sigma^{*2}_{\setminus\boldsymbol{\gamma}})$, where $\boldsymbol{\beta}^*_{\setminus\boldsymbol{\gamma}}$ and $\sigma^*_{\setminus\boldsymbol{\gamma}}$ are the respective ML estimates and corresponding standard errors of $\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}$ from the full model using the actual data $\mathbf{y}$ and $\mathbf{I}_{d_{\setminus\boldsymbol{\gamma}}}$ is the $d_{\setminus\boldsymbol{\gamma}} \times d_{\setminus\boldsymbol{\gamma}}$ identity matrix. The full augmented posterior is

$$\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}, \beta_0, \mathbf{y}^*, \boldsymbol{\gamma}|\mathbf{y}, \delta) \propto f_{\boldsymbol{\gamma}}(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta} f_0(\mathbf{y}^*|\beta_0)^{1/\psi}}{m_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^*|\delta)} \pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}})\pi(\boldsymbol{\gamma})\pi_0^{\mathrm{N}}(\beta_0). \tag{26}$$

The proposed PEP-GVS sampling scheme is as follows. For starting values $\boldsymbol{\gamma}^{(0)}, \boldsymbol{\beta}^{(0)} = (\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(0)}, \boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}^{(0)}), \beta_0^{(0)}$, $\mathbf{y}^{*(0)}$ and iterations $t = 1, 2, \ldots, N$:

Step 1: Set current values $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t-1)}$, $\beta_0 = \beta_0^{(t-1)}$ $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(t-1)}$ and $\mathbf{y}^* = \mathbf{y}^{*(t-1)}$.
Step 2: For $j = 1, 2, \ldots, p$, sample $\gamma_j \sim \pi(\gamma_j|\boldsymbol{\beta}, \boldsymbol{\gamma}_{\setminus j}, \mathbf{y}^*, \mathbf{y}, \delta)$ for $\gamma_j \in \{0, 1\}$.
Step 3: Update $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}})$ based on the current configuration of $\boldsymbol{\gamma}$.
Step 4: Sample $\boldsymbol{\beta}_{\boldsymbol{\gamma}} \sim \pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\gamma, \mathbf{y}^*, \mathbf{y}, \delta)$ in a M-H step.
Step 5: Sample $\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}$ from the pseudo-prior.
Step 6: Sample $\beta_0 \sim \pi(\beta_0|\mathbf{y}^*, \psi) \propto f_0(\mathbf{y}^*|\beta_0)^{1/\psi}\pi_0^{\mathrm{N}}(\beta_0)$ in a M-H step.
Step 7: Sample $\mathbf{y}^* \sim \pi(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \beta_0, \boldsymbol{\gamma}, \delta, \psi) \propto \frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta} f_0(\mathbf{y}^*|\beta_0)^{1/\psi}}{m_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^*|\delta)}$ in a M-H step.
Step 8: Update $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}$, $\beta_0^{(t)} = \beta_0$ $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}$ and $\mathbf{y}^{*(t)} = \mathbf{y}^*$.

Implementation details and an analytic description of the algorithm are provided as supplementary material (Appendix B).

4 Hyper-δ extensions

The PEP prior for the normal regression model can be interpreted as a mixture of g-priors where the power parameter δ is equivalent to g and the mixing density is the prior predictive of the reference model (Fouskakis et al., 2015). Thus, under the PEP approach we assign a hyper-prior on the imaginary data $\mathbf{y}^*$, rather than to the variance multiplier, i.e. the power parameter δ. As discussed in Section 2.2, the same representation holds asymptotically in the GLM setting given a flat baseline prior. A natural extension of the PEP methodology arises by introducing an extra hierarchical level to the model formulation via the assignment of a hyper-prior on δ. Moving from a fixed (but reasonable) choice of δ to a stochastic version of this parameter is desirable as it simplifies prior specifications by letting the data to “speak” for δ leading, eventually, to a fully objective procedure.

The hyper-δ CR/DR-PEP priors can be approximately expressed as

$$\pi_{\boldsymbol{\gamma}}^{\mathrm{CR/DR-PEP}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \approx \int\int f_{\mathrm{N}_{d_{\boldsymbol{\gamma}}}}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}; \widehat{\boldsymbol{\beta}}^*_{\boldsymbol{\gamma}}, \delta \boldsymbol{J}^*_{\boldsymbol{\gamma}}\big(\widehat{\boldsymbol{\beta}}^*_{\boldsymbol{\gamma}}\big)^{-1}\right) m_0^{\mathrm{Z}}(\mathbf{y}^*|\psi)\pi(\delta)\mathrm{d}\mathbf{y}^*\mathrm{d}\delta \qquad (27)$$

under a baseline prior $\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \propto 1$, where $m_0^{\mathrm{Z}}(\mathbf{y}^*|\psi)$ is equal to $m_0^{\mathrm{N}}(\mathbf{y}^*)$ for $\psi = 1$ (CR-PEP) and equal to $m_0^{\mathrm{Z}}(\mathbf{y}^*|\delta)$ for $\psi = \delta$ (DR-PEP), $\widehat{\boldsymbol{\beta}}^*_{\boldsymbol{\gamma}}$ is the ML estimate given the imaginary data, $\boldsymbol{J}^*_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}^*_{\boldsymbol{\gamma}})$ is the observed information matrix evaluated at $\widehat{\boldsymbol{\beta}}^*_{\boldsymbol{\gamma}}$ and $f_{\mathrm{N}_{d_{\boldsymbol{\gamma}}}}(\cdot)$ denotes the $d_{\boldsymbol{\gamma}}$–dimensional multivariate normal distribution. Sensible options for $\pi(\delta)$ are the hyper-g analogues proposed in Liang et al. (2008). Specifically, we consider the hyper-δ prior $\pi(\delta) = \frac{\alpha-2}{2}(1+\delta)^{-\alpha/2}$, for $\alpha > 2$, $\delta > 0$, which corresponds to a $\mathrm{Beta}(1, \frac{\alpha}{2} - 1)$ distribution for the shrinkage factor $\frac{\delta}{1+\delta}$. Thinking in terms of shrinkage, Liang et al. (2008) propose setting $\alpha = 3$ in order to place most of the probability mass near 1 or $\alpha = 4$ which leads to a uniform prior. An alternative option is the hyper-δ/n prior given by $\pi(\delta) = \frac{\alpha-2}{2n}(1+\frac{\delta}{n})^{-\alpha/2}$, for $\alpha > 2$, $\delta > 0$. In principle, any other prior from the related literature can be incorporated in the PEP design; for instance, the inverse-gamma hyper-prior of Zellner and Siow (1980) or the recent g-prior mixtures proposed by Maruyama and George (2011) and Bayarri et al. (2012). Of course, when working outside the context of the normal linear model, the integration in (27) with respect to δ will not be tractable. Therefore, in order to incorporate the stochastic nature of δ we need to introduce one additional MCMC sampling step. In this case the augmented posterior is given by

$$\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}, \beta_0, \mathbf{y}^*, \boldsymbol{\gamma}, \delta|\mathbf{y}) \propto \pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}, \beta_0, \mathbf{y}^*, \boldsymbol{\gamma}|\mathbf{y}, \delta)\pi(\delta), \qquad (28)$$

where the first quantity in the right-hand side of (28) is given in (26). Details are provided in Appendix B of the supplementary material.

5 Desiderata for PEP priors in GLMs

5.1 Model selection consistency

With respect to model selection consistency (Bayarri et al., 2012), analytical proofs for the normal linear model are provided in Fouskakis et al. (2016). Here, we present empirical evidence suggesting that this criterion is also valid for non-Gaussian GLMs under the PEP priors. For further details and results, we defer to Section 7.2 where we illustrate, for several simulated scenarios with binomial and Poisson response models, that the posterior probability of the true model approaches one as the sample size increases.

5.2 Information consistency

The definition of information consistency is unclear under GLMs with known dispersion parameters. According to Li and Clyde (2016), for models with discrete responses and known variance (such as the Poisson and binomial models), information inconsistency, as defined by Bayarri et al. (2012), is not an issue since the likelihood is bounded even for saturated models.

5.3 Predictive matching

Under reasonable baseline assumptions, the CR/DR-PEP priors are satisfying the criteria of null and dimension predictive matching as defined in Bayarri et al. (2012). In order to illustrate this, we express the baseline prior of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ as a product of functions $\psi(\boldsymbol{\eta}_{\boldsymbol{\gamma}})$ and $\Psi_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\setminus 0,\boldsymbol{\gamma}})$, where $\boldsymbol{\eta}_{\boldsymbol{\gamma}} = (\eta_{\boldsymbol{\gamma}(i)}, \eta_{\boldsymbol{\gamma}(2)}, \ldots, \eta_{\boldsymbol{\gamma}(n)})^T$ is the linear predictor and $\boldsymbol{\beta}_{\setminus 0,\boldsymbol{\gamma}}$ is the vector of all elements of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ excluding the intercept $\beta_{0,\boldsymbol{\gamma}}$ of model $M_{\boldsymbol{\gamma}}$. Also we reintroduce ϕ covering the general case in which the nuisance parameter is under estimation. The statements are as follows.

Proposition 1. *Under a baseline prior $\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\phi) = \psi(\boldsymbol{\eta}_{\boldsymbol{\gamma}})\Psi_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\setminus 0,\boldsymbol{\gamma}})$ with $\delta = n^* = n$, the fixed δ PEP priors satisfy the null predictive matching criterion for samples of size one.*

Proof of Proposition 1 is provided in Appendix A.1. □

Proposition 2. *The hyper-δ and the hyper-δ/n DR-PEP priors with $n^* = n$ and baseline prior as in Proposition 1 satisfy the null predictive matching criterion for samples of size one.*

Proof of Proposition 2 is provided in Appendix A.2. □

Proposition 3. *The hyper-δ and the hyper-δ/n CR-PEP priors with $n^* = n$ and baseline prior as in Proposition 1 satisfy the null predictive matching criterion for samples of size one.*

Proof of Proposition 3 is provided in Appendix A.3. □

Proposition 4. *Under a baseline prior $\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\phi) = \psi(\boldsymbol{\eta}_{\boldsymbol{\gamma}})$ with $n^* = n$, the DR-PEP priors (fixed $\delta = n$, hyper-δ and hyper-δ/n) satisfy the dimension predictive matching criterion for samples of size $p_{\boldsymbol{\gamma}} + 1$.*

Proof of Proposition 4 is provided in Appendix A.4. □

Proposition 5. *The CR-PEP priors (fixed $\delta = n$, hyper-δ and hyper-δ/n) with $n^* = n$ and baseline prior as in Proposition 4 satisfy the dimension predictive matching criterion for samples of size $p_{\boldsymbol{\gamma}} + 1$.*

Proof of Proposition 5 can be obtained by using similar arguments as in the proof of Proposition 4. □

The baseline prior distributions discussed in Section 2.4 satisfy the requirements in Propositions 1 and 4; under a flat prior $\Psi(\boldsymbol{\beta}_{\setminus 0,\boldsymbol{\gamma}}) = 1$ and $\psi(\boldsymbol{\eta}_{\boldsymbol{\gamma}}) \propto 1$, while under Jeffreys prior $\Psi(\boldsymbol{\beta}_{\setminus 0,\boldsymbol{\gamma}}) = 1$ and $\psi(\boldsymbol{\eta}_{\boldsymbol{\gamma}}) \propto |\mathbf{X}^T_{\boldsymbol{\gamma}}\mathbf{W}_{\boldsymbol{\gamma}}(\boldsymbol{\eta}_{\boldsymbol{\gamma}})\mathbf{X}_{\boldsymbol{\gamma}}|^{1/2}$, where $\mathbf{W}_{\boldsymbol{\gamma}}(\boldsymbol{\eta}_{\boldsymbol{\gamma}}) = \mathrm{diag}(w_{\boldsymbol{\gamma}(i)})$, $w_{\boldsymbol{\gamma}(i)} = (\frac{\partial\mu(\eta_{\boldsymbol{\gamma},(i)})}{\partial\eta_{\boldsymbol{\gamma},(i)}})^2[a_i(\phi)b''(\vartheta(\eta_{\boldsymbol{\gamma},(i)}))]^{-1}$ and $\mu(\eta_{\boldsymbol{\gamma},(i)}) = b'(\vartheta(\eta_{\boldsymbol{\gamma},(i)}))$.

6 A general framework

In this section we present a synopsis for the various priors under consideration. This requires introducing a set of separate power parameters δ_0 and δ_1, which respectively relate to the marginal likelihood and the posterior distribution components. Under this setting we have the following general prior formulation $\pi^{\mathrm{G}}(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\omega}, \delta_0, \delta_1) = \pi^{\mathrm{G}}(\boldsymbol{\theta}_{\boldsymbol{\gamma}}|\boldsymbol{\omega}, \delta_0, \delta_1)\pi(\boldsymbol{\omega})\pi(\delta_0)\pi(\delta_1)$, where $\mathrm{G} \in \mathcal{P}$ with $\mathcal{P}$ being the set of PEP prior configurations considered in this paper, also including the EP prior. Here, $\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ corresponds to the model specific parameters, while $\boldsymbol{\omega}$ is a common nuisance parameter across all models. When $\boldsymbol{\omega}$ does not exist or is known, $\pi(\boldsymbol{\omega})$ should be omitted. Similarly, when δ_0 and/or δ_1 are fixed, $\pi(\delta_0)$ and/or $\pi(\delta_1)$ are omitted.

All priors in the set $\mathcal{P}$ are derived as follows:

$$\pi^{\mathrm{G}}(\boldsymbol{\theta}_{\boldsymbol{\gamma}}|\boldsymbol{\omega}, \delta_0, \delta_1) = \frac{\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\theta}_{\boldsymbol{\gamma}}|\boldsymbol{\omega})}{k_{\boldsymbol{\gamma}}(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\omega}, \delta_1)\mathcal{C}_0} \int \frac{m^{\mathrm{N}}_0(\mathbf{y}^*|\boldsymbol{\omega}, \delta_0)}{m^{\mathrm{N}}_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\omega}, \delta_1)} f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\omega})^{1/\delta_1} d\mathbf{y}^*, \tag{29}$$

where we have that $m^{\mathrm{N}}_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\omega}, \delta_1) = \int k_{\boldsymbol{\gamma}}(\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\omega}, \delta_1)^{-1} f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\omega})^{1/\delta_1}\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\theta}_{\boldsymbol{\gamma}}|\boldsymbol{\omega})d\boldsymbol{\theta}_{\boldsymbol{\gamma}}$ and $m^{\mathrm{N}}_0(\mathbf{y}^*|\boldsymbol{\omega}, \delta_0) = \int k_0(\boldsymbol{\theta}_0, \boldsymbol{\omega}, \delta_0)^{-1} f_0(\mathbf{y}^*|\boldsymbol{\theta}_0, \boldsymbol{\omega})^{1/\delta_0}\pi^{\mathrm{N}}_0(\boldsymbol{\theta}_0|\boldsymbol{\omega})d\boldsymbol{\theta}_0$. Each prior in the set $\mathcal{P}$ can be obtained from (29); details are provided in Table 2. In Table 3 we summarize issues and proposed solutions for all priors under consideration.

7 Illustrative examples

7.1 Methods

In this section we first present a simulation study for logistic and Poisson regression taking into account independent and correlated predictors. We proceed with a simulation

Prior (G)	$\boldsymbol{\theta}_\gamma$	$\boldsymbol{\omega}$	δ_0	δ_1	Hyper-prior $\pi(\delta)$	$k_0(\boldsymbol{\theta}_0, \boldsymbol{\omega}, \delta_0)$	$k_\gamma(\boldsymbol{\theta}_\gamma, \boldsymbol{\omega}, \delta_1)$	$\mathcal{C}_0$
EP	$\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi_{\boldsymbol{\gamma}}$	$\emptyset$	1	1		1	1	1
PEP	$\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi_{\boldsymbol{\gamma}}$	$\emptyset$	n^*	n^*		κ_0	κ_1	1
PCEP	$\boldsymbol{\beta}_{\boldsymbol{\gamma}}$	ϕ	n^*	n^*		κ_0	κ_1	1
CR-PEP	$\boldsymbol{\beta}_{\boldsymbol{\gamma}}$	ϕ	1	n^*		1	1	1
DR-PEP	$\boldsymbol{\beta}_{\boldsymbol{\gamma}}$	ϕ	n^*	n^*		1	1	c_0
CR-PEP hyper-δ	$\boldsymbol{\beta}_{\boldsymbol{\gamma}}$	ϕ	1	δ	$\frac{a-2}{2}(1+\delta)^{-a/2}$	1	1	1
DR-PEP hyper-δ	$\boldsymbol{\beta}_{\boldsymbol{\gamma}}$	ϕ	δ	δ	$\frac{a-2}{2}(1+\delta)^{-a/2}$	1	1	c_0
CR-PEP hyper-δ/n	$\boldsymbol{\beta}_{\boldsymbol{\gamma}}$	ϕ	1	δ	$\frac{a-2}{2n}(1+\frac{\delta}{n})^{-a/2}$	1	1	1
DR-PEP hyper-δ/n	$\boldsymbol{\beta}_{\boldsymbol{\gamma}}$	ϕ	δ	δ	$\frac{a-2}{2n}(1+\frac{\delta}{n})^{-a/2}$	1	1	c_0

$\kappa_0 = \int f_0(\mathbf{y}^*|\boldsymbol{\theta}_0, \boldsymbol{\omega})^{1/\delta_0} d\mathbf{y}^*$; $\kappa_1 = \int f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\theta}_{\boldsymbol{\gamma}}, \boldsymbol{\omega})^{1/\delta_1} d\mathbf{y}^*$; $c_0 = \int\int f_0(\mathbf{y}^*|\boldsymbol{\theta}_0, \boldsymbol{\omega})^{1/\delta_0} \pi_0^N(\boldsymbol{\theta}_0|\boldsymbol{\omega}) d\boldsymbol{\theta}_0 d\mathbf{y}^*$.

Table 2: Schematic presentation of all priors in $\mathcal{P}$.

Prior	**Issues**	**Solutions**
EP	– Selection of imaginary sample size n^* – Sub-sampling of $\boldsymbol{X}_{\boldsymbol{\gamma}}^*$ – Informative when using minimal training sample and p is close to n	– Issues are solved using PEP with $\delta = n^* = n$ and $\boldsymbol{X}_{\boldsymbol{\gamma}}^* = \boldsymbol{X}_{\boldsymbol{\gamma}}$
PEP	– Cumbersome normalized power likelihood in GLMs – Monte Carlo is needed for the computation of the marginal likelihood even in the normal linear model	– Use of unnormalized power likelihoods that lead to the CR/DR-PEP priors – Use PCEP that leads to a conjugate setup in the normal linear model
PCEP	– Not information consistent	– Use PEP which is information consistent
CR-PEP	– No clear definition of m_0^{N} under the unnormalized power likelihood – Selection of δ	– Use the original likelihood in m_0^{N} – Set $\delta = n^*$ to have unit information interpretation or consider random δ
DR-PEP	– No clear definition of m_0^{N} under the unnormalized power likelihood – Selection of δ	– Use the density normalized m_0^{Z} under the unnormalized power likelihood – Set $\delta = n^*$ to have unit information interpretation or consider random δ
CR/DR-PEP hyper-δ	– Demanding computation – Prior of δ is not centered to unit-information	– Use fixed-δ CR/DR-PEP versions – Use the hyper-δ/n prior
CR/DR-PEP hyper-δ/n	– Demanding computation	– Use fixed-δ CR/DR-PEP versions

Table 3: Issues and solutions of all priors in $\mathcal{P}$.

study for logistic models where the number of predictors is larger and the correlation structure is more complicated. The section concludes with a real data example for binary responses. In all illustrations we consider the CR/DR-PEP priors (introduced in Section 2.3) and their hyper-δ and hyper-δ/n extensions (discussed in Section 4) with parameter $\alpha = 3$. In all configurations $n^* = n$ and $\mathbf{X}_{\boldsymbol{\gamma}}^* = \mathbf{X}_{\boldsymbol{\gamma}}$, where the columns of the design matrix are centered around their sample means. For fixed δ, we consider the default unit-information approach, i.e. $\delta = n^*$. Jeffreys prior is used as baseline for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$; see Section 2.4. We compare the PEP variants with standard g-prior methods, using the GLM version of Sabanés Bové and Held (2011) for the parameter vector and a flat prior for the intercept. In particular, we consider the unit-information g-prior ($g = n$) and three mixtures of g-priors; the hyper-g and hyper-g/n priors with $\alpha = 3$ (Liang et al., 2008), and the beta hyper-prior proposed by Maruyama and George (2011). Henceforth, the latter will be referred to as MG hyper-g. These approaches are also implemented via GVS.

7.2 Simulation study 1

In this first example we consider logistic and Poisson simulations, presented in Hansen and Yu (2003) and Chen et al. (2008), respectively. Both cases have been also considered by Li and Clyde (2016). The number of predictors is $p = 5$ in the logistic model and $p = 3$ in the Poisson model, where each predictor is drawn from a standard normal distribution with pairwise correlations given by $\text{corr}(X_i, X_j) = r^{|i-j|},\ 1 \leq i < j \leq p$. We consider: (i) independent ($r = 0$) and (ii) correlated ($r = 0.75$) predictors. Four sparsity scenarios are assumed. For the logistic case we use the same sample size as in Hansen and Yu (2003), namely $n = 100$, but with lower effects resulting in smaller values of odds ratios. Specifically, $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)^T$ is set to $(0.1, 0, 0, 0, 0, 0)^T$ in the null scenario, $(0.1, 0.7, 0, 0, 0, 0)^T$ in the sparse scenario, $(0.1, 1.6, 0.8, -1.5, 0, 0)^T$ in the medium scenario and $(0.1, 1.75, 1.5, -1.1, -1.4, 0.5)^T$ in the full scenario. The resulting odds ratios are approximately 2, 2.5 and 3.5 for the sparse, medium and full models, respectively. For the Poisson simulation we consider $n = 100$ and the same regression coefficients as in Chen et al. (2008); i.e. $(\beta_0, \beta_1, \beta_2, \beta_3)^T$ equal to $(-0.3, 0, 0, 0)^T$ in the null scenario, $(-0.3, 0.3, 0, 0)^T$ in the sparse scenario, $(-0.3, 0.3, 0.2, 0)^T$ in the medium scenario and $(-0.3, 0.3, 0.2, -0.15)^T$ in the full scenario. Each simulation is repeated 100 times. Since p is small, we use a uniform prior on model space (Section 2.4).

Evaluation of model selection consistency of PEP methods

First we examine the behaviour of PEP methods for increasing sample size. Under the assumption of model selection consistency, we expect the posterior probability of the true model to approach the value of one as sample size increases. Indeed, all PEP methods under the sparse, medium and full scenarios confirm the consistency criterion as it is evident in Figures 1 and 2.

Comparison between different methods

Results based on the frequency of identifying the true data-generating model through the maximum a-posteriori (MAP) model for the logistic regression simulation are summarized in Table 4.

Comparison between PEP approaches versus the rest of the methods indicates the following:

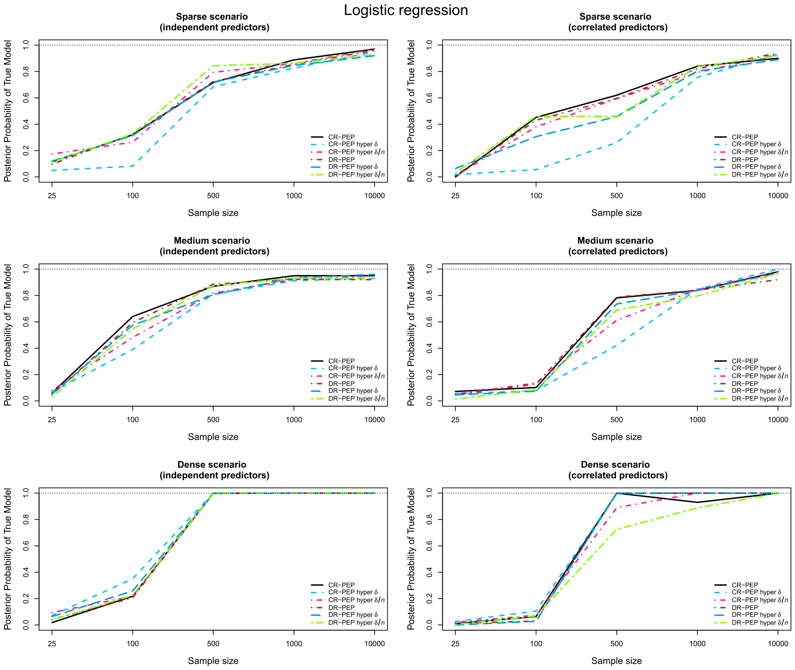

Figure 1: Posterior probabilities of the true model vs. sample size for the sparse, medium and dense logistic regression scenarios.

Scenario	**r**	**Prior distributions** g-prior	hyper g-prior	hyper g/n-prior	MG hyper g-prior	CR PEP	CR PEP hyper-δ	CR PEP hyper-δ/n	DR PEP	DR PEP hyper-δ	DR PEP hyper-δ/n
null	0.00	77	35	63	75	79	46	80	79	73	**82**
	0.75	91	52	81	88	**94**	60	82	93	91	92
sparse	0.00	67	57	63	67	**72**	58	68	**72**	**72**	**72**
	0.75	74	60	67	72	72	60	**76**	74	73	73
medium	0.00	83	82	**84**	**84**	83	**84**	81	83	**84**	**84**
	0.75	33	**38**	34	30	26	37	32	27	29	27
full	0.00	41	41	42	**43**	28	38	29	26	32	31
	0.75	14	15	**17**	14	8	12	10	8	10	8

Table 4: Number of times (over 100 replications) that the MAP model coincides with the true model in the logistic regressions of Simulation Study 1 (row-wise largest value in bold).

i) Overall the PEP procedures perform satisfactorily as in 5 out of the 8 simulated scenarios the "best" method for identifying the true model involves one of the PEP priors.

ii) The PEP procedures outperform all competing methods under the null and sparse scenarios.

iii) In the medium scenario, the PEP priors perform equally well to the rest of the methods in the case of independent predictors and slightly worse in the case of correlated predictors.

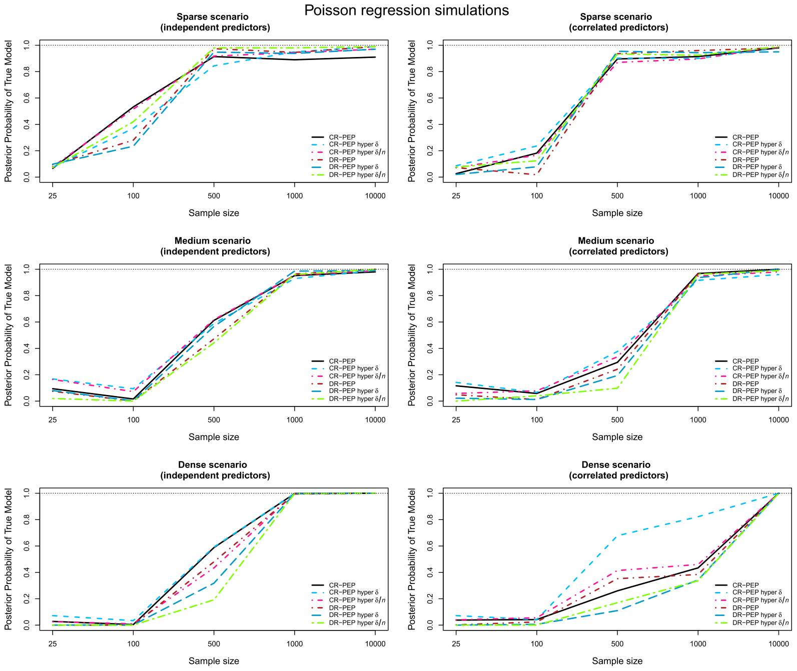

Figure 2: Posterior probabilities of the true model vs. sample size for the sparse, medium and dense Poisson regression scenarios.

iv) In the full model scenario, the g-prior based methods perform better than the PEP based approaches. This is no surprise since PEP priors support more parsimonious solutions.

The comparison between the CR-PEP and DR-PEP priors reveals no obvious differences between the two approaches for fixed $\delta = n$. Concerning the fixed δ approach versus the hyper-δ and δ/n extensions, we see that, under the DR-PEP approach, all results are more or less the same. However, this is not the case for the CR-PEP approach, where the hyper-δ version supports more complex models than the fixed-δ based method, while the results based on the hyper-δ/n prior are somewhere in the middle. Interestingly, a similar pattern is observed among the g-prior and the hyper-g, hyper-g/n priors. Boxplots of posterior inclusion probabilities (PIPs) are available in Appendix D.1 of the supplementary material.

Results from the Poisson simulations are presented in Table 5. Overall, conclusions are similar to the logistic case:

i) The PEP priors perform overall satisfactory; 6 out of the 8 best MAP success patterns are spotted by one of the PEP based methods.
ii) The PEP priors perform well under sparsity, i.e. under the null and sparse models.

		Prior distributions									
Scenario	**r**	g-prior	hyper g-prior	hyper g/n-prior	MG hyper g-prior	CR PEP	CR PEP hyper-δ	CR PEP hyper-δ/n	DR PEP	DR PEP hyper-δ	DR PEP hyper-δ/n
null	0.00	86	68	80	87	88	71	83	90	91	**94**
	0.75	91	68	90	94	95	75	91	95	**97**	95
sparse	0.00	75	74	74	75	76	68	**80**	73	68	69
	0.75	40	43	41	38	35	**44**	40	32	30	28
medium	0.00	29	43	37	36	27	**44**	30	28	25	20
	0.75	0	**5**	0	0	0	4	0	0	0	0
full	0.00	6	**23**	13	9	5	18	11	5	4	3
	0.75	0	0	1	0	0	**3**	0	0	0	0

Table 5: Number of times (over 100 replications) that the MAP model coincides with the true model in the Poisson regressions of Simulation Study 1 (row-wise largest value in bold).

iii) In the medium scenarios, the hyper-g and hyper-δ CR-PEP priors yield the best results; however, under correlated predictors the true model is rarely traced.

iv) In the full model with independent covariates, the rates are low; hyper-g has the highest rate but with the hyper-δ CR-PEP being close. In the correlated case all methods fail.

With respect to the various PEP prior distributions, the comparison in the Poisson case leads to the same findings as previously. Again, the most interesting finding is that inference under the DR-PEP prior is not affected by the choice of fixed versus random δ. On the contrary, this is not the case for the CR-PEP prior, where the hyper-δ extension systematically supports more complex models. Boxplots of PIPs are provided as supplementary material in Appendix D.2.

Finally, the accuracy of the log-marginal likelihood is assessed empirically and indicative results and comparisons are provided in Appendix C of the supplementary material of this manuscript.

7.3 Simulation study 2

Here we consider logistic simulations with $p = 10$ predictors and $n = 200$. The first five covariates are generated from a standard normal, while the remaining five are generated as $X_{ij} \sim N(0.3X_{i1} + 0.5X_{i2} + 0.7X_{i3} + 0.9X_{i4} + 1.1X_{i5}, 1)$, for $i = 1, \ldots, n$ and $j = 6, \ldots, 10$. Three models are assumed: $(\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5, \beta_6, \beta_7, \beta_8, \beta_9, \beta_{10})^T$ is equal to $(0.1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0)^T$ in a null design, to $(0.1, 0, 0, -0.9, 0, 0, 0, 1.2, 0, 0, 0.4)^T$ in a sparse design and to $(0.1, 0.6, 0, -0.9, 0, 1, 0.9, 1.2, -1.2, -0.5, 0)^T$ in a dense design. The odds ratios for the sparse and dense simulation models are approximately 2 and 3, respectively. Each simulation is repeated 100 times. We use the beta-binomial prior on model space; see Section 2.4. Figures 3, 4 and 5 present boxplots of PIPs under the null, sparse and dense scenarios, respectively.

Under the null scenario (Figure 3), all methods, except the hyper-g prior, exhibit strong shrinkage towards zero on the PIPs. The hyper-g prior leads to larger estimates that have higher variability. The hyper-δ CR-PEP prior also induces more variability.

Under the sparse scenario (Figure 4) there are no striking differences among methods.

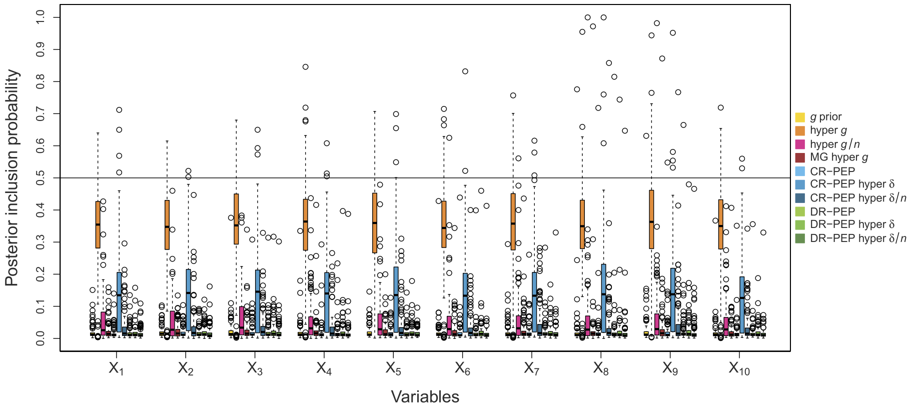

Figure 3: Posterior inclusion probabilities for Simulation Study 2 under the various priors from 100 repetitions of the null logistic simulation scenario.

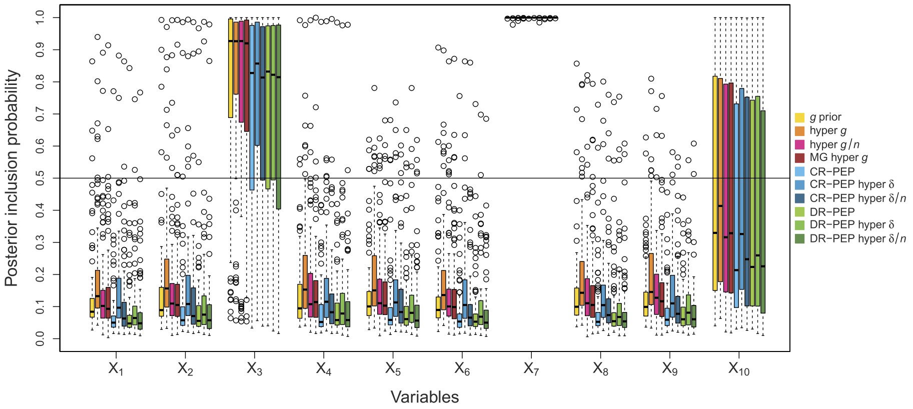

Figure 4: Posterior inclusion probabilities for Simulation Study 2 under the various priors from 100 repetitions of the sparse logistic simulation scenario where the true model is $X_3 + X_7 + X_{10}$.

All priors provide very strong support for the inclusion of X_7 and sufficient support for the inclusion of X_3. Moreover, all methods yield very wide PIP intervals for predictor X_{10}.

Finally, in the dense scenario (Figure 5) the fixed-δ PEP priors generally outperform other methods as they yield lower PIPs for the unimportant effects of X_2, X_4, X_{10}. The g-prior and the hyper DR-PEP extensions yield similar PIPs and generally perform well; however, these priors introduce some uncertainty concerning the inclusion of covariate X_4. The rest of the methods systematically support more complex models.

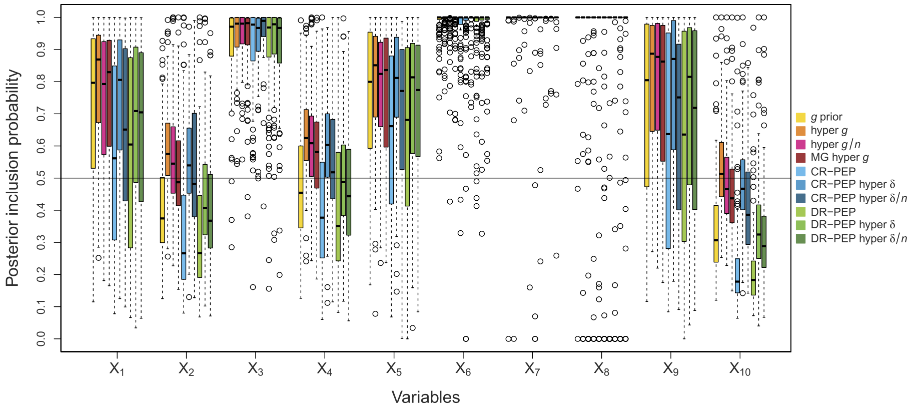

Figure 5: Posterior inclusion probabilities for Simulation Study 2 under the various priors from 100 repetitions of the dense logistic simulation scenario where the true model is $X_1 + X_3 + X_5 + X_6 + X_7 + X_8 + X_9$.

7.4 A real data example

Lastly, we consider the Pima Indians diabetes data set which has been analyzed in several studies (e.g. Holmes and Held, 2006; Sabanés Bové and Held, 2011). The data consist of $n = 532$ complete records on diabetes presence and $p = 7$ potential covariates; namely, number of pregnancies (X_1), plasma glucose concentration (X_2), diastolic blood pressure (X_3), triceps skin fold thickness (X_4), body mass index (X_5), diabetes pedigree function (X_6) and age (X_7). For each method we use 41000 iterations of the GVS algorithm, discarding the first 1000 as burn-in. The beta-binomial prior is used on model space (Section 2.4).

Table 6 shows the PIPs under the various methods. For comparison reasons we also include results from the Zellner and Siow (1980) inverse gamma (ZS-IG) prior, the hyper-g/n with $\alpha = 4$, and a non-informative inverse gamma (NI-IG) hyper-g prior with shape and scale equal to 10^{-3}. The PIPs we obtain via GVS are in agreement with the results presented in Sabanés Bové and Held (2011). For covariates X_1, X_2, X_5 and X_6, which seem to be highly influential, the results in Table 6 show no significant differences among methods. On the contrary, the PIPs of the "uncertain" covariates X_3, X_4 and X_7 vary substantially; specifically, the inclusion probabilities from the fixed-δ CR/DR-PEP priors, the hyper-δ/n DR-PEP prior and the g-prior are considerably lower than the inclusion probabilities resulting from the rest of the methods. In terms of the shrinkage factors $g/(g+1)$ and $\delta/(\delta+1)$, results show that the shrinkage effect is stronger when g or δ is fixed, which leads to a drastic reduction in the effects (and the PIPs) of low-influential covariates. On the other hand, the priors with random g or δ clearly result in higher PIPs. Among this category of priors, the hyper-δ/n DR-PEP is evidently the most parsimonious, as it yields PIPs which are actually quite close to those obtained from fixed δ PEP priors. Further results with comments can be found in Appendix E of the supplementary material.

Method	**Predictor**						
	X_1	X_2	X_3	X_4	X_5	X_6	X_7
ZS-IG hyper-g	0.961	1.000	0.252	0.250	0.998	0.994	0.530
NI-IG hyper-g	0.967	1.000	0.349	0.341	0.998	0.996	0.622
g-prior ($g = n$)	0.952	1.000	0.136	0.139	0.998	0.992	0.382
hyper-g ($\alpha = 3$)	0.970	1.000	0.397	0.379	0.998	0.996	0.669
hyper-g/n ($\alpha = 3$)	0.966	1.000	0.304	0.300	0.998	0.995	0.579
hyper-g/n ($\alpha = 4$)	0.965	1.000	0.307	0.299	0.997	0.995	0.582
MG hyper-g	0.958	1.000	0.262	0.259	0.998	0.994	0.548
CR-PEP	0.948	1.000	0.100	0.104	0.998	0.987	0.339
CR-PEP hyper-δ	0.964	1.000	0.296	0.291	0.998	0.995	0.602
CR-PEP hyper-δ/n	0.956	1.000	0.223	0.225	0.998	0.992	0.520
DR-PEP	0.948	1.000	0.102	0.104	0.997	0.988	0.324
DR-PEP hyper-δ	0.954	1.000	0.174	0.173	0.997	0.991	0.442
DR-PEP hyper-δ/n	0.951	1.000	0.125	0.120	0.998	0.987	0.346

Table 6: Posterior inclusion probabilities for the seven covariates of the Pima Indians diabetes data set.

Method	**MAP**	**False Neg. (%)**	**False Pos. (%)**	**MPM**	**False Neg. (%)**	**False Pos. (%)**
g-prior ($g = n$)	$\mathcal{M}_A$	10.8	16.5	$\mathcal{M}_A$	10.8	16.5
hyper-g ($\alpha = 3$)	$\mathcal{M}_A + X_3 + X_4 + X_7$	11.4	16.9	$\mathcal{M}_A + X_7$	11.1	16.8
hyper-g/n ($\alpha = 3$)	$\mathcal{M}_A$	11.0	16.6	$\mathcal{M}_A + X_7$	11.0	16.6
MG hyper-g	$\mathcal{M}_A$	10.9	16.6	$\mathcal{M}_A + X_7$	10.9	16.6
CR-PEP	$\mathcal{M}_A$	10.9	16.9	$\mathcal{M}_A$	10.9	16.9
CR-PEP hyper-δ	$\mathcal{M}_A$	10.9	17.0	$\mathcal{M}_A + X_7$	11.3	16.4
CR-PEP hyper-δ/n	$\mathcal{M}_A$	10.8	17.0	$\mathcal{M}_A + X_7$	11.0	16.6
DR-PEP	$\mathcal{M}_A$	10.9	16.8	$\mathcal{M}_A$	10.9	16.8
DR-PEP hyper-δ	$\mathcal{M}_A$	10.9	16.9	$\mathcal{M}_A$	10.9	16.9
DR-PEP hyper-δ/n	$\mathcal{M}_A$	10.9	16.8	$\mathcal{M}_A$	10.9	16.8

$\mathcal{M}_A : X_1 + X_2 + X_5 + X_6$

Table 7: Percentages of false negative and false positive detections for the Pima Indians diabetes data set under the MAP model and median probability model (MPM) for the various priors.

We conclude by examining the out-of-sample predictive accuracy under each prior. Table 7 summarizes the false positive and false negative prediction rates under the MAP and median probability models using a random split of the data into a half. Overall, we cannot find a dominant method in terms of predictive accuracy; we note however, that the most complex MAP model arises from the hyper-g prior which also results in the highest false negative prediction rates. In contrast, the unit-information g-prior, the CR-PEP prior with fixed δ, and the DR-PEP priors lead to the most parsimonious median probability model, which is comparable in terms of predictive performance with the model that further includes X_7, indicated as the median probability model by the rest of the methods.

8 Discussion

In this article we extended the PEP formulation to two new prior designs which significantly enhance the applicability of the proposed methodology. We focused on variable selection for GLMs, however, the CR/DR-PEP priors proposed here may in principle be used for any general model setting. The new approaches retain the desired features of the original PEP prior formulation by: i) resolving the problem of selecting and averaging across minimal imaginary samples, thus, also allowing for large-sample approximations, and ii) being minimally informative by scaling down the effect of the imaginary data on the posterior distribution. We further introduced hyper-prior distributions, analogues to the priors proposed in Liang et al. (2008), for the power parameter δ that controls the contribution of the imaginary data.

With respect to the criteria in Bayarri et al. (2012), we provided proofs for the null and dimensional predictive matching criteria for all priors under consideration. Regarding model selection consistency, proofs for the normal linear model are provided in Fouskakis et al. (2016). In this paper we illustrated empirically that this criterion appears to hold also for binomial and Poisson GLMs.

The empirical results suggest that the proposed PEP priors outperform mixtures of g-priors in terms of introducing larger shrinkage to the inclusion probabilities of non-influential or partially influential predictors, thus, leading to more parsimonious solutions with comparable predictive accuracy. When comparing PEP priors with fixed $\delta = n$ and random δ the results indicate that the former approach induces more stringent control in the inclusion of predictors. Therefore, fixed PEP priors support simpler models which is a desirable feature when the number of covariates is large. Concerning the choice between the CR and the DR prior setups, we conclude in favour to the use of the latter since it is rather robust with respect to the fixed vs. random specification of δ.

Future research aims to extend the PEP methodology to high-dimensional problems, including the small n–large p case, by incorporating shrinkage priors (e.g. ridge and LASSO procedures) into the PEP design. Another promising alternative is to embody the expectation-maximization variable selection approach of Ročková and George (2014) within the PEP prior.

Supplementary Material

Electronic Appendix of the “Power-Expected-Posterior Priors for Generalized Linear Models” (DOI: 10.1214/17-BA1066SUPP; .pdf).

References

Bayarri, M. J., Berger, J. O., Forte, A., and García-Donato, G. (2012). "Criteria for Bayesian model choice with application to variable selection." *The Annals of Statistics*, 40: 1550–1577. MR3015035. doi: https://doi.org/10.1214/12-AOS1013. 723, 733, 734, 744

Berger, J. O. and Pericchi, L. R. (1996a). "The intrinsic Bayes factor for linear models." in J. Bernardo, J. Berger, A. Dawid, and A. Smith, eds., *Bayesian Statistics*, Vol. 5, 25–44. Oxford University Press. MR1425398. 722

Berger, J. O. and Pericchi, L. R. (1996b). "The intrinsic Bayes factor for model selection and prediction." *Journal of the American Statistical Association*, 91: 109–122. MR1394065. doi: https://doi.org/10.2307/2291387. 722, 725

Bernardo, J. and Smith, A. (2000). *Bayesian Theory, 2nd edition*. Chichester, UK: Wiley. MR1274699. doi: https://doi.org/10.1002/9780470316870. 726

Casella, G. and Moreno, E. (2006). "Objective Bayesian variable selection." *Journal of the American Statistical Association*, 101: 157–167. MR2268035. doi: https://doi.org/10.1198/016214505000000646. 722, 725

Chen, M., Ibrahim, J. G., and Shao, Q.-M. (2000). "Power prior distributions for generalized linear models." *Journal of Statistical Planning and Inference*, 84: 121–137. MR1747500. doi: https://doi.org/10.1016/S0378-3758(99)00140-8. 726

Chen, M.-H., Huang, L., Ibrahim, J. G., and Kim, S. (2008). "Bayesian variable selection and computation for generalized linear models with conjugate priors." *Bayesian Analysis*, 3: 585–614. MR2434404. doi: https://doi.org/10.1214/08-BA323. 722, 737

Chen, M.-H. and Ibrahim, J. G. (2003). "Conjugate priors for generalized linear models." *Statistica Sinica*, 13: 461–476. MR1977737. 722, 723

Consonni, G. and Veronese, P. (2008). "Compatibility of prior specifications across linear models." *Statistical Science*, 23: 332–353. MR2483907. doi: https://doi.org/10.1214/08-STS258. 722

Dellaportas, P., Forster, J. J., and Ntzoufras, I. (2002). "On Bayesian model and variable selection using MCMC." *Statistics and Computing*, 12: 27–36. MR1877577. doi: https://doi.org/10.1023/A:1013164120801. 732

Fouskakis, D. and Ntzoufras, I. (2013). "Computation for intrinsic variable selection in normal regression models via expected-posterior prior." *Statistics and Computing*, 23: 491–499. MR3070406. doi: https://doi.org/10.1007/s11222-012-9325-9. 725

Fouskakis, D. and Ntzoufras, I. (2016). "Power-conditional-expected priors: Using g-priors with random imaginary data for variable selection." *Journal of Computational and Graphical Statistics*, 25: 647–664. MR3533631. doi: https://doi.org/10.1080/10618600.2015.1036996. 724, 725, 727, 728, 729, 730, 731

Fouskakis, D., Ntzoufras, I., and Draper, D. (2015). "Power-expected-posterior priors for variable selection in Gaussian linear models." *Bayesian Analysis*, 10: 75–107.

MR3420898. doi: https://doi.org/10.1214/14-BA887. 722, 724, 725, 726, 727, 733

Fouskakis, D., Ntzoufras, I., and Perrakis, K. (2016). "Variations of power-expected-posterior priors in normal regression models." arXiv:1609.06926v2. 730, 734, 744

Fouskakis, D., Ntzoufras, I., and Perrakis, K. (2017). "Electronic Appendix of the "Power-Expected-Posterior Priors for Generalized Linear Models"." *Bayesian Analysis*. doi: https://doi.org/10.1214/17-BA1066SUPP. 731

Friel, N. and Pettitt, A. N. (2008). "Marginal likelihood estimation via power posteriors." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 70: 589–607. MR2420416. doi: https://doi.org/10.1111/j.1467-9868.2007.00650.x. 727

Gupta, M. and Ibrahim, J. G. (2009). "An information matrix prior for Bayesian analysis in generalized linear models with high dimensional data." *Statistica Sinica*, 19: 1641–1663. MR2589202. 723

Hansen, M. and Yu, B. (2003). "Minimum description length model selection criteria for generalized linear models." *Lecture Notes-Monograph Series*, 6: 145–163. MR2004337. doi: https://doi.org/10.1214/lnms/1215091140. 723, 737

Holmes, C. C. and Held, L. (2006). "Bayesian auxiliary variable models for binary and multinomial regression." *Bayesian Analysis*, 145–168. MR2227368. doi: https://doi.org/10.1214/06-BA105. 742

Ibrahim, J. G. and Chen, M.-H. (2000). "Power prior distributions for regression models." *Statistical Science*, 15: 46–60. MR1842236. doi: https://doi.org/10.1214/ss/1009212673. 724, 725

Ibrahim, J. G. and Laud, P. W. (1991). "On Bayesian analysis of generalized linear models using Jeffreys's prior." *Journal of the American Statistical Association*, 86: 981–986. MR1146346. 723, 730

Kass, R. E. and Wasserman, L. (1995). "A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion." *Journal of the American Statistical Association*, 90: 928–934. MR1354008. 723, 726

Leon-Novelo, L., Moreno, E., and Casella, G. (2012). "Objective Bayes model selection in probit models." *Statistics in Medicine*, 31: 353–365. MR2879809. doi: https://doi.org/10.1002/sim.4406. 722, 723

Li, Y. and Clyde, M. A. (2016). "Mixtures of g-priors in generalized linear models." arXiv:1503.06913. MR3213874. doi: https://doi.org/10.1007/s11425-014-4815-1. 723, 734, 737

Liang, F., Paulo, R., Molina, G., Clyde, M. A., and Berger, J. O. (2008). "Mixtures of g-priors for Bayesian variable selection." *Journal of the American Statistical Association*, 103: 410–423. MR2420243. doi: https://doi.org/10.1198/016214507000001337. 723, 733, 737, 744

Madigan, D. and York, J. (1995). "Bayesian graphical models for discrete data." *International Statistical Review*, 63: 215–232. 723

Maruyama, Y. and George, E. I. (2011). "Fully Bayes factors with a generalized g-prior." *The Annals of Statistics*, 39: 2740–2765. MR2906885. doi: https://doi.org/10.1214/11-AOS917. 733, 737

Moreno, E. and Girón, F. J. (2008). "Comparison of Bayesian objective procedures for variable selection in linear regression." *Test*, 17: 472–490. MR2470092. doi: https://doi.org/10.1007/s11749-006-0039-1. 725

Murray, I., Ghahramani, Z., and MacKay, D. J. C. (2006). "MCMC for doubly-intractable distributions." in *Proceedings of the 22nd Annual Conference on Uncertainty in Artificial Intelligence*, (UAI-06), AUAI Press, 359–366. 727

Ntzoufras, I., Dellaportas, P., and Forster, J. J. (2003). "Bayesian variable and link determination for generalized linear models." *Journal of Statistical Planning and Inference*, 111: 165–180. MR1955879. doi: https://doi.org/10.1016/S0378-3758(02)00298-7. 722, 723, 726

Pérez, J. (1998). "*Development of Expected Posterior Prior Distribution for Model Comparisons*." Ph.D. thesis, Department of Statistics, Purdue University, USA. MR2699463. 725

Pérez, J. M. and Berger, J. O. (2002). "Expected-posterior prior distributions for model selection." *Biometrika*, 89: 491–511. MR1929158. doi: https://doi.org/10.1093/biomet/89.3.491. 722, 724, 728

Perrakis, K., Fouskakis, D., and Ntzoufras, I. (2015). "Bayesian Variable Selection for Generalized Linear Models Using the Power-Conditional-Expected-Posterior Prior." in S. Frühwirth-Schnatter, A. Bitto, G. Kastner, and A. Posekany, eds., *Bayesian Statistics from Methods to Models and Applications: Research from BAYSM 2014*, Vol. 126, 59–73. Springer Proceedings in Mathematics and Statistics. MR3374421. doi: https://doi.org/10.1007/978-3-319-16238-6_6. 727

Ročková, V. and George, E. I. (2014). "EMVS: The EM approach to Bayesian variable selection." *Journal of the American Statistical Association*, 109: 828–846. MR3223753. doi: https://doi.org/10.1080/01621459.2013.869223. 744

Sabanés Bové, D. and Held, L. (2011). "Hyper-g priors for generalized linear models." *Bayesian Analysis*, 6: 387–410. MR2843537. doi: https://doi.org/10.1214/ba/1339616469. 723, 726, 737, 742

Scott, J. G. and Berger, J. O. (2010). "Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem." *The Annals of Statistics*, 38: 2587–2619. MR2722450. doi: https://doi.org/10.1214/10-AOS792. 731

Wang, X. and George, E. I. (2007). "Adaptive Bayesian criteria in variable selection for generalized linear models." *Statistica Sinica*, 17: 667–690. MR2408684. 723

Zellner, A. (1986). "On Assessing Prior Distributions and Bayesian Regression Analysis Using G-Prior distributions." In Goel, P. and Zellner, A. (eds.), *Bayesian Inference* *and Decision Techniques: Essays in Honor of Bruno de Finetti*, 233–243. Amsterdam: North-Holland. MR0881437. 722

Zellner, A. and Siow, A. (1980). "Posterior Odds Ratios for Selected Regression Hypothesis (with discussion)." In J. M. Bernardo, M. H. DeGroot, D. V. Lindley and A. F. M. Smith, eds., *Bayesian Statistics*, Vol. 1, 585–606 & 618–647 (discussion). Oxford University Press. MR0862503. 722, 733, 742

Acknowledgments

The authors thank the Editor-in-Chief, the Associate Editor and the Referee for their comments on a previous version of our work that greatly strengthened the current article.