Power-Expected-Posterior Priors for Generalized Linear Models

Dimitris Fouskakis[1], Ioannis Ntzoufras[2] and Konstantinos Perrakis[2]

[1] *Department of Mathematics, National Technical University of Athens*

[2] *Department of Statistics, Athens University of Economics and Business*

Abstract

The power-expected-posterior (PEP) prior developed for variable selection in normal regression models provides an objective, automatic, consistent and parsimonious model selection procedure. At the same time it resolves the conceptual and computational problems which emerge owing 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 further considered. Under these approaches the resulting PEP prior can be asymptotically represented as a double mixture of g-priors. For estimation of posterior model and inclusion probabilities we introduce a tuning-free Gibbs-based variable selection sampler. Several simulation scenarios and one real data 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. Empirical results indicate that the GLM-PEP adaptations are more effective when the aim is parsimonious inference.

Keywords: expected-posterior prior, g/hyper-g priors, generalized linear models, 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 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, 1996*a*,*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 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 priors as well as EP priors have been proposed in many articles for use on the variable selection problem in the 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, Moreno and Casella, 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 (e.g Laplace 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, Ntzoufras and Draper (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. Therefore, 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

and hyper-g prior (Liang, Paulo, Molina, Clyde and Berger, 2008) distributions 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, Huang, Ibrahim and Kim (2008) proposed a Markov chain Monte Carlo (MCMC) based solution for this class of models. Ntzoufras, Dellaportas and Forster (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. 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 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 ML estimates, as the prior variance-covariance matrix (Hansen and Yu, 2003, Wang and George, 2007, Li and Clyde, 2015). 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 (2015) recommend using the Bayesian adaptive sampling algorithm (Clyde, Ghosh and Littman, 2011). 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); under a Gaussian likelihood the information-matrix prior becomes the standard g-prior, while for $g \to \infty$ it reduces to Jeffreys prior which is proper only for the case of the binomial model. 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).

As seen, at present most methods for GLMs are anchored to the g-prior approach (Zellner and Siow, 1980, Zellner, 1986) and therefore cannot be regarded as objective and fully automatic approaches in the sense that one cannot conduct an analysis starting with non-informative, flat priors. 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. Illustrative examples and comparisons with other methods using both simulated and real data sets are presented in Section 5. Section 6 concludes with a summary and a discussion of future research directions.

2 PEP priors for generalized linear models

We consider n realizations of a response variable Y accompanied by a set of predictors $X_1, X_2, ..., X_p$ which may potentially 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

$$f_{\boldsymbol{\gamma}}(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi_{\boldsymbol{\gamma}}) = \exp\left(\sum_{i=1}^{n} \frac{y_i \vartheta_{\boldsymbol{\gamma}(i)} - b(\vartheta_{\boldsymbol{\gamma}(i)})}{a_i(\phi_{\boldsymbol{\gamma}})} + \sum_{i=1}^{n} c(y_i, \phi_{\boldsymbol{\gamma}}) \right). \tag{2.1}$$

The functions $a(\cdot)$, $b(\cdot)$ 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 $\vartheta_{\boldsymbol{\gamma}(i)} = \vartheta\big(\eta_{\boldsymbol{\gamma}(i)}\big) \equiv g \circ b'^{-1}\big(\eta_{\boldsymbol{\gamma}(i)}\big)$, 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 $g \circ b'(\vartheta_{\boldsymbol{\gamma}(i)}) \equiv g\big(b'(\vartheta_{\boldsymbol{\gamma}(i)})\big)$. Commonly, a canonical ϑ function is used, so that $\vartheta_{\boldsymbol{\gamma}(i)} = \eta_{\boldsymbol{\gamma}(i)}$. We assume that an 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 $\alpha(\cdot)$ is typically of the form $\alpha_i(\phi_{\boldsymbol{\gamma}}) = \phi_{\boldsymbol{\gamma}}/w_i$, where the 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 $\mathrm{Var}(y_i) = b''(\vartheta_{\boldsymbol{\gamma}(i)})\alpha_i(\phi)$.

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.

2.1 An overview of the PEP prior

The PEP prior, initially formulated in Fouskakis et al. (2015) for the case of the normal linear model, creatively 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 (2.1) and a default *baseline-prior* of the form $\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi) = \pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\phi)\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\phi)$, we have a corresponding *baseline-posterior* distribution given by

$$\pi_{\gamma}^{\mathrm{N}}(\boldsymbol{\beta}_{\gamma}, \phi|\mathbf{y}^*) = \frac{f_{\gamma}(\mathbf{y}^*|\boldsymbol{\beta}_{\gamma}, \phi)\pi_{\gamma}^{\mathrm{N}}(\boldsymbol{\beta}_{\gamma}|\phi)\pi_{\gamma}^{\mathrm{N}}(\phi)}{m_{\gamma}^{\mathrm{N}}(\mathbf{y}^*)}. \tag{2.2}$$

The EP prior for the parameters of model M_{γ} is then defined as the posterior distribution in (2.2), averaged over all possible imaginary samples, i.e.

$$\pi_{\gamma}^{\mathrm{EP}}(\boldsymbol{\beta}_{\gamma}, \phi) = \int \pi_{\gamma}^{\mathrm{N}}(\boldsymbol{\beta}_{\gamma}, \phi|\mathbf{y}^*)\, m^*(\mathbf{y}^*)\mathrm{d}\mathbf{y}^*\,. \tag{2.3}$$

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 (1996*b*).

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 the posterior inference, the reasonable solution is to choose the smallest possible n^* for which the posterior is proper. This leads to the concept of the so-called *minimal training sample*, which however requires calculating the arithmetic mean (or other appropriate measures of centrality) of Bayes factors over all possible minimal training samples. In addition, when it comes to regression the same problem arises with the design matrix as one has to choose appropriate covariate values for each minimal training sample, and this further depends upon the choice of the reference model. A computational solution to deal with the aforementioned problems has been proposed in the literature (e.g. Casella and Moreno, 2006, Moreno and Girón, 2008), however, this solution is only applicable under the normal linear regression model and in addition 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 number of predictors; see Fouskakis et al. (2015) for details. Finally, when n^* is small and (2.3) 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 (2.2) 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 prior in the GLM setup, under the null-reference model M_0, is defined as follows

$$\pi_{\gamma}^{\mathrm{PEP}}(\boldsymbol{\beta}_{\gamma}, \phi|\delta) = \pi_{\gamma}^{\mathrm{PEP}}(\boldsymbol{\beta}_{\gamma}|\phi, \delta)\pi_{\gamma}^{\mathrm{N}}(\phi), \tag{2.4}$$

where

$$
\begin{aligned}
\pi_{\gamma}^{\mathrm{PEP}}(\boldsymbol{\beta}_{\gamma}|\phi,\delta) &= \int \pi_{\gamma}^{\mathrm{N}}(\boldsymbol{\beta}_{\gamma}|\mathbf{y}^{*},\phi,\delta) m_{0}^{\mathrm{N}}(\mathbf{y}^{*}|\phi,\delta)\mathrm{d}\mathbf{y}^{*}, &(2.5)\\
\pi_{\gamma}^{\mathrm{N}}(\boldsymbol{\beta}_{\gamma}|\mathbf{y}^{*},\phi,\delta) &= \frac{f_{\gamma}(\mathbf{y}^{*}|\boldsymbol{\beta}_{\gamma},\phi,\delta)\pi_{\gamma}^{\mathrm{N}}(\boldsymbol{\beta}_{\gamma}\,|\phi)}{m_{\gamma}^{\mathrm{N}}(\mathbf{y}^{*}|\phi,\delta)}, &(2.6)\\
m_{\gamma}^{\mathrm{N}}(\mathbf{y}^{*}|\phi,\delta) &= \int f_{\gamma}(\mathbf{y}^{*}|\boldsymbol{\beta}_{\gamma},\phi,\delta)\pi_{\gamma}^{\mathrm{N}}(\boldsymbol{\beta}_{\gamma}\,|\phi)\mathrm{d}\boldsymbol{\beta}_{\gamma}, &(2.7)\\
f_{\gamma}(\mathbf{y}^{*}|\boldsymbol{\beta}_{\gamma},\phi,\delta) &\propto f_{\gamma}(\mathbf{y}^{*}|\boldsymbol{\beta}_{\gamma},\phi)^{1/\delta}, &(2.8)\\
m_{0}^{\mathrm{N}}(\mathbf{y}^{*}|\phi,\delta) &= \int f_{0}(\mathbf{y}^{*}|\beta_{0},\phi,\delta)\pi_{0}^{\mathrm{N}}(\beta_{0}\,|\phi)\mathrm{d}\beta_{0}, &(2.9)\\
f_{0}(\mathbf{y}^{*}|\beta_{0},\phi,\delta) &\propto f_{0}(\mathbf{y}^{*}|\beta_{0},\phi)^{1/\delta}. &(2.10)
\end{aligned}
$$

Note that the PEP prior for the intercept term of M_0 essentially reduces to the baseline prior; i.e. $\pi_0^{\mathrm{PEP}}(\beta_0|\phi,\delta) = \pi_0^{\mathrm{N}}(\beta_0|\phi)$. Here the power-parameter δ controls the weight that the imaginary data contribute to the "final" posterior distributions of $\boldsymbol{\beta}_{\gamma}$ and ϕ. The default choice is to set it equal to the size of the imaginary sample, i.e. $\delta = n^*$. Under this approach the contribution of the imaginary data is downweighted to overall account for one data point, leading to a minimally-informative prior with a unit-information interpretation (Kass and Wasserman, 1995). 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 $\mathbf{X}_{\gamma}^{*} \equiv \mathbf{X}_{\gamma}$. As shown in Fouskakis et al. (2015) 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 instance, consider the baseline-posterior in (2.6), which can be expressed as

$$
\pi_{\gamma}^{\mathrm{N}}(\boldsymbol{\beta}_{\gamma}|\mathbf{y}^{*},\phi,\delta) \;\propto\; \exp\left(\sum_{i=1}^{n^*}\frac{y_i^*\vartheta_{\gamma(i)} - b(\vartheta_{\gamma(i)})}{\delta a_i(\phi)}\right)\pi_{\gamma}^{\mathrm{N}}(\boldsymbol{\beta}_{\gamma}|\phi). \qquad (2.11)
$$

This unnormalized distribution is recognized as the power-prior for GLMs (Chen, Ibrahim and Shao, 2000). If we assume a flat baseline prior for $\boldsymbol{\beta}_{\gamma}$, i.e. $\pi_{\gamma}^{\mathrm{N}}(\boldsymbol{\beta}_{\gamma}|\phi) \propto 1$, then, based on standard Bayesian asymptotic theory (Bernando and Smith, 2000), for $n^* \rightarrow \infty$ the distribution in (2.11) converges to

$$
\widehat{\pi}_{\gamma}^{\mathrm{N}}(\boldsymbol{\beta}_{\gamma}|\mathbf{y}^{*},\phi,\delta) \approx \mathrm{N}_{d_\gamma}\big(\widehat{\boldsymbol{\beta}}_{\gamma}^{*}, \delta\boldsymbol{J}_{\gamma}^{*}\big(\widehat{\boldsymbol{\beta}}_{\gamma}^{*}\big)^{-1}\big), \qquad (2.12)
$$

where $\widehat{\boldsymbol{\beta}}_{\gamma}^{*}$ is the MLE of $\boldsymbol{\beta}_{\gamma}$ for data $\mathbf{y}^*$ and design matrix $\mathbf{X}_{\gamma}^{*}$, and $\boldsymbol{J}_{\gamma}^{*}\big(\widehat{\boldsymbol{\beta}}_{\gamma}^{*}\big)$ is the observed information evaluated at $\widehat{\boldsymbol{\beta}}_{\gamma}^{*}$. Specifically, $\boldsymbol{J}_{\gamma}^{*} = \big(\mathbf{X}_{\gamma}^{*T}\mathbf{W}_{\gamma}^{*}\mathbf{X}_{\gamma}^{*}\big)^{-1}$, where $\mathbf{W}_{\gamma}^{*} = \mathrm{diag}(w_{\gamma(i)}^{*})$ with $w_{\gamma(i)}^{*} = \big(\frac{\partial\mu_{\gamma(i)}}{\partial\eta_{\gamma(i)}}\big)^2\big[a_i(\phi)b''(\vartheta_{\gamma(i)})\big]^{-1}$ and $\mu_{\gamma(i)} = b'(\vartheta_{\gamma(i)})$. It is straightforward to see that the asymptotic distribution in (2.12) has a g-prior form according to the definitions for GLMs presented in Ntzoufras et al. (2003) and Bové and Held (2011). The familiar zero-mean representation in (2.12) 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 $\widehat{\boldsymbol{\beta}}^*_\gamma = \mathbf{0}_{d_\gamma}$; for details see Ntzoufras et al. (2003).

2.2 PEP prior extensions for GLMs via unnormalized power-likelihoods

The sampling distribution of the imaginary data involved in the PEP prior via (2.6), (2.7) and (2.9) 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 the density normalized power-likelihood

$$f_\gamma(\mathbf{y}^*|\boldsymbol{\theta}_\gamma, \delta) = \frac{f_\gamma(\mathbf{y}^*|\boldsymbol{\theta}_\gamma)^{1/\delta}}{\int f_\gamma(\mathbf{y}^*|\boldsymbol{\theta}_\gamma)^{1/\delta} \mathrm{d}\mathbf{y}^*}, \tag{2.13}$$

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 distributions 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, Fouskakis and Ntzoufras (2015*a*).

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 identifiable 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, Ghahramani and MacKay, 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 fixed, which is the case for the binomial, Poisson and normal with known error variance regression models. Specifically, we assume that $\phi = 1$ and remove ϕ from all conditional expressions to alleviate notation.

The core idea is to use the unnormalized power likelihood (2.8) and (2.10) and normalize the baseline posterior density (2.11), i.e.

$$\pi_\gamma^{\mathrm{N}}(\boldsymbol{\beta}_\gamma|\mathbf{y}^*,\delta)=\frac{f_\gamma(\mathbf{y}^*|\boldsymbol{\beta}_\gamma)^{1/\delta}\pi_\gamma^{\mathrm{N}}(\boldsymbol{\beta}_\gamma)}{\int f_\gamma(\mathbf{y}^*|\boldsymbol{\beta}_\gamma)^{1/\delta}\pi_\gamma^{\mathrm{N}}(\boldsymbol{\beta}_\gamma)\mathrm{d}\boldsymbol{\beta}_\gamma}, \tag{2.14}$$

which is also the approach of Friel and Pettitt (2008, Eq.4) in the definition of the power posterior. 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 (2.14) across imaginary datasets. 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* $\boldsymbol{\beta}_\gamma$ *is defined as the power-posterior of* $\boldsymbol{\beta}_\gamma$ *in* (2.14) *"averaged" over all imaginary data coming from the prior predictive distribution of the reference model* M_0 *based on the actual likelihood, that is*

$$\pi_\gamma^{\mathrm{CR-PEP}}(\boldsymbol{\beta}_\gamma|\delta)=\mathbb{E}_{\mathbf{y}^*}^{m_0^{\mathrm{N}}}\left[\pi_\gamma^{\mathrm{N}}(\boldsymbol{\beta}_\gamma|\mathbf{y}^*,\delta)\right]=\pi_\gamma^{\mathrm{N}}(\boldsymbol{\beta}_\gamma)\int\frac{m_0^{\mathrm{N}}(\mathbf{y}^*)}{m_\gamma^{\mathrm{N}}(\mathbf{y}^*|\delta)}f_\gamma(\mathbf{y}^*|\boldsymbol{\beta}_\gamma)^{1/\delta}\mathrm{d}\mathbf{y}^* \tag{2.15}$$

$$\begin{aligned}\text{with } m_0^{\mathrm{N}}(\mathbf{y}^*) &= \int f_0(\mathbf{y}^*|\beta_0)\pi_0^{\mathrm{N}}(\beta_0)\mathrm{d}\beta_0 &(2.16)\\ \text{and } m_\gamma^{\mathrm{N}}(\mathbf{y}^*|\delta) &= \int f_\gamma(\mathbf{y}^*|\boldsymbol{\beta}_\gamma)^{1/\delta}\pi_\gamma^{\mathrm{N}}(\boldsymbol{\beta}_\gamma)\mathrm{d}\boldsymbol{\beta}_\gamma.\end{aligned}$$

In order for the above prior to exist, we need to consider, for each model M_γ, similar assumptions as in Pérez and Berger (2002), i.e.

$$0<m_\gamma^{\mathrm{N}}(\mathbf{y}^*|\delta)<\infty,\quad 0<\int\frac{m_0^{\mathrm{N}}(\mathbf{y}^*)}{m_\gamma^{\mathrm{N}}(\mathbf{y}^*|\delta)}f_\gamma(\mathbf{y}^*|\boldsymbol{\beta}_\gamma)^{1/\delta}\mathrm{d}\mathbf{y}^*<\infty. \tag{2.17}$$

In equation (2.15), m_0^{N} will not necessarily be proper, but still, by abusing slightly notation, we define the concentrated-reference PEP prior as the expectation of $\pi_\gamma^{\mathrm{N}}(\boldsymbol{\beta}_\gamma|\mathbf{y}^*,\delta)$ with respect to m_0^{N}. Furthermore, impropriety of the baseline priors in (2.15), causes no indeterminacy of the resulting Bayes factors, since $\pi_\gamma^{\mathrm{CR-PEP}}(\boldsymbol{\beta}_\gamma|\delta)$ 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

$$\pi_0^{\mathrm{CR-PEP}}(\beta_0|\delta)=\pi_0^{\mathrm{N}}(\beta_0)\int\frac{m_0^{\mathrm{N}}(\mathbf{y}^*)}{m_0^{\mathrm{N}}(\mathbf{y}^*|\delta)}f_0(\mathbf{y}^*|\beta_0)^{1/\delta}\mathrm{d}\mathbf{y}^*. \tag{2.18}$$

Definition 2 *The* **diffuse-reference PEP prior** *of model parameters* $\boldsymbol{\beta}_\gamma$ *is defined as the power-posterior of* $\boldsymbol{\beta}_\gamma$ *in* (2.14) *"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*

$$\pi_{\gamma}^{\mathrm{DR-PEP}}(\boldsymbol{\beta}_{\gamma}|\delta) = \mathbb{E}_{\mathbf{y}^*|\delta}^{m_0^{\mathrm{Z}}}\Big[\pi_{\gamma}^{\mathrm{N}}(\boldsymbol{\beta}_{\gamma}|\mathbf{y}^*,\delta)\Big] = \pi_{\gamma}^{\mathrm{N}}(\boldsymbol{\beta}_{\gamma}) \int \frac{m_0^{\mathrm{Z}}(\mathbf{y}^*|\delta)}{m_{\gamma}^{\mathrm{N}}(\mathbf{y}^*|\delta)} f_{\gamma}(\mathbf{y}^*|\boldsymbol{\beta}_{\gamma})^{1/\delta}\mathrm{d}\mathbf{y}^* \quad (2.19)$$

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

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

The conditions for the existence of the diffuse-reference PEP prior, for each model M_{γ}, are similar to (2.17), i.e.

$$0 < m_{\gamma}^{\mathrm{N}}(\mathbf{y}^*|\delta) < \infty, \quad 0 < \int \frac{m_0^{\mathrm{N}}(\mathbf{y}^*|\delta)}{m_{\gamma}^{\mathrm{N}}(\mathbf{y}^*|\delta)} f_{\gamma}(\mathbf{y}^*|\boldsymbol{\beta}_{\gamma})^{1/\delta}\mathrm{d}\mathbf{y}^* < \infty. \quad (2.20)$$

Again the definition of the diffuse-reference PEP prior as an expectation of $\pi_{\gamma}^{\mathrm{N}}(\boldsymbol{\beta}_{\gamma}|\mathbf{y}^*,\delta)$ 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 $\mathcal{C}_0 = \int m_0^{\mathrm{N}}(\mathbf{y}^*|\delta)\mathrm{d}\mathbf{y}^*$ 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 $\pi_{\gamma}^{\mathrm{DR-PEP}}(\boldsymbol{\beta}_{\gamma}|\delta)$ depends only on $\mathcal{C}_0$ which is common across all models. Finally, the diffuse-reference PEP prior for the parameter of the null model is no longer equal to the baseline prior, since

$$\pi_0^{\mathrm{DR-PEP}}(\beta_0|\delta) = \pi_0^{\mathrm{N}}(\beta_0)\frac{\int f_0(\mathbf{y}^*|\beta_0)^{1/\delta}\mathrm{d}\mathbf{y}^*}{\mathcal{C}_0} = \frac{\int f_0(\mathbf{y}^*|\beta_0)^{1/\delta}\pi_0^{\mathrm{N}}(\beta_0)\mathrm{d}\mathbf{y}^*}{\int\int f_0(\mathbf{y}^*|\beta_0)^{1/\delta}\pi_0^{\mathrm{N}}(\beta_0)\mathrm{d}\beta_0\mathrm{d}\mathbf{y}^*}. \quad (2.21)$$

Definition 1 is a special case of Definition 2 since $m_0^{\mathrm{N}}(\mathbf{y}^*)$ is a special case of $m_0^{\mathrm{N}}(\mathbf{y}^*|\delta)$ with $\delta = 1$. Because the likelihood in (2.16) 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 is also given by

$$\pi_{\gamma}^{\mathrm{CR-PEP}}(\boldsymbol{\beta}_{\gamma}|\delta) = \pi_{\gamma}^{\mathrm{N}}(\boldsymbol{\beta}_{\gamma}) \int\int \frac{f_{\gamma}(\mathbf{y}^*|\boldsymbol{\beta}_{\gamma})^{1/\delta} f_0(\mathbf{y}^*|\beta_0)}{m_{\gamma}^{\mathrm{N}}(\mathbf{y}^*|\delta)} \pi_0^{\mathrm{N}}(\beta_0)\mathrm{d}\mathbf{y}^*\mathrm{d}\beta_0. \quad (2.22)$$

In Definition 2 the likelihood involved in $m_0^{\mathrm{N}}(\mathbf{y}^*|\delta)$ in (2.19) 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 *diffuse-reference* PEP (DR-PEP). Specifically, we have that

$$\pi_{\gamma}^{\mathrm{DR-PEP}}(\boldsymbol{\beta}_{\gamma}|\delta) = \mathcal{C}_0^{-1}\pi_{\gamma}^{\mathrm{N}}(\boldsymbol{\beta}_{\gamma}) \int\int \frac{f_{\gamma}(\mathbf{y}^*|\boldsymbol{\beta}_{\gamma})^{1/\delta} f_0(\mathbf{y}^*|\beta_0)^{1/\delta}}{m_{\gamma}^{\mathrm{N}}(\mathbf{y}^*|\delta)} \pi_0^{\mathrm{N}}(\beta_0)\mathrm{d}\mathbf{y}^*\mathrm{d}\beta_0. \quad (2.23)$$

2.3 Further prior specifications

To complete the model formulation we need to specify a baseline prior for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, under each model $M_{\boldsymbol{\gamma}}$, and also a prior distribution on the model space $\mathcal{M} = \{0, 1\}^p$, as we are interested in variable selection rather than single-model inference in which case $\boldsymbol{\gamma} \in \mathcal{M}$ is fixed. In addition, we do not need to specify a prior for ϕ, which is considered fixed in our setting. For models with random (under estimation) ϕ, we propose working along the lines of Fouskakis and Ntzoufras (2016) and use a flat prior on ϕ; this will just add one additional step to the MCMC algorithm presented in Section 3.

Baseline prior distributions for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$

Common choices for the baseline prior of the regression vector $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, under each model $M_{\boldsymbol{\gamma}}$, are either the flat improper prior

$$\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \propto 1 \tag{2.24}$$

or Jeffreys' prior for GLMs (Ibrahim and Laud, 1991) which is of the form

$$\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \propto |\mathbf{X}_{\boldsymbol{\gamma}}^T \mathbf{W}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \mathbf{X}_{\boldsymbol{\gamma}}|^{1/2} \ . \tag{2.25}$$

For non-Gaussian GLMs Jeffreys' prior will depend on $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ through the GLM weight matrix $\mathbf{W}_{\boldsymbol{\gamma}}(\cdot)$; see Section 2.1 for details. Note that Jeffreys' prior for the parameter of the null model simplifies to $\pi_0^{\mathrm{N}}(\beta_0) \propto \mathrm{tr}\big(\mathbf{W}_0(\beta_0)\big)^{1/2}$.

Prior distributions on model space

A common non-informative option for $\boldsymbol{\gamma}$ is to use 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 the model space, i.e.

$$\pi(\boldsymbol{\gamma}) = 2^{-p}. \tag{2.26}$$

An alternative choice, better suited for moderate to large p, is to use the hierarchical prior design

$$\boldsymbol{\gamma}|\tau \sim \text{Bernoulli}(\tau) \text{ and } \tau \sim \text{Beta}(1, 1),$$

in order to account for an appropriate multiplicity adjustment (Scott and Berger, 2010). In this case the resulting prior is given by

$$\pi(\boldsymbol{\gamma}) = \frac{1}{p+1} \binom{p}{p_{\boldsymbol{\gamma}}}^{-1}. \tag{2.27}$$

3 Posterior Computation

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 there exist no convenient conjugate distributions and the integrals involved in the derivation of the CR/DR-PEP priors are intractable. However, one can work with the hierarchical model, i.e. without marginalizing over the imaginary data, and use an MCMC algorithm in order to sample from the joint posterior distribution of $\boldsymbol{\beta}_\gamma$ and $\mathbf{y}^*$.

For ease of exposition, for the remainder of this section we use indicator ψ to distinguish between the CR-PEP prior ($\psi = 1$) and the DR-PEP prior ($\psi = \delta$) and we simply use the general term "PEP" to denote the joint posterior. Specifically, from (2.14), (2.15) and (2.19) we have the following hierarchical form

$$\begin{aligned} \pi_{\boldsymbol{\gamma}}^{\text{PEP}}(\boldsymbol{\beta}_\gamma, \mathbf{y}^* | \mathbf{y}, \delta) &\propto f_{\boldsymbol{\gamma}}(\mathbf{y}|\boldsymbol{\beta}_\gamma)\pi_{\boldsymbol{\gamma}}^{\text{N}}(\boldsymbol{\beta}_\gamma|\mathbf{y}^*, \delta) m_0^{\text{N}}(\mathbf{y}^*|\psi) \\ &\propto f_{\boldsymbol{\gamma}}(\mathbf{y}|\boldsymbol{\beta}_\gamma)\frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_\gamma)^{1/\delta}\pi_{\boldsymbol{\gamma}}^{\text{N}}(\boldsymbol{\beta}_\gamma)}{m_{\boldsymbol{\gamma}}^{\text{N}}(\mathbf{y}^*|\delta)} m_0^{\text{N}}(\mathbf{y}^*|\psi), \end{aligned} \tag{3.1}$$

where $m_0^{\text{N}}(\mathbf{y}^*|1) \equiv m_0^{\text{N}}(\mathbf{y}^*)$. A further computational problem in (3.1) relates to the prior predictive distributions $m_{\boldsymbol{\gamma}}^{\text{N}}(\mathbf{y}^*|\delta)$ and $m_0^{\text{N}}(\mathbf{y}^*|\psi)$ which are not available in closed form. One solution is to use a Laplace approximation for both. Alternatively, a more accurate solution is to augment the parameter space further and include parameter β_0 of the reference model M_0 in the joint posterior, thus avoiding to approximate $m_0^{\text{N}}(\mathbf{y}^*|\psi)$. Based on (2.22) and (2.23) the posterior in (3.1) is expanded as

$$\pi_{\boldsymbol{\gamma}}^{\text{PEP}}(\boldsymbol{\beta}_\gamma, \beta_0, \mathbf{y}^*|\mathbf{y}, \delta) \propto f_{\boldsymbol{\gamma}}(\mathbf{y}|\boldsymbol{\beta}_\gamma)\frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_\gamma)^{1/\delta}\pi_{\boldsymbol{\gamma}}^{\text{N}}(\boldsymbol{\beta}_\gamma)}{m_{\boldsymbol{\gamma}}^{\text{N}}(\mathbf{y}^*|\delta)} f_0(\mathbf{y}^*|\beta_0)^{1/\psi}\pi_0^{\text{N}}(\beta_0), \tag{3.2}$$

which leaves us with the need of using only one Laplace approximation for $m_{\boldsymbol{\gamma}}^{\text{N}}(\mathbf{y}^*|\delta)$.

Sampling from (3.2) for a single model $M_{\boldsymbol{\gamma}}$, i.e. for a fixed configuration of $\boldsymbol{\gamma}$, is feasible using standard Metropolis-within-Gibbs algorithms. 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 2.3, and utilize the algorithm of Dellaportas, Forster and Ntzoufras (2002). Note that under flat baseline priors the posterior in (3.2) and the corresponding MCMC scheme are simplified. Finally, under a flat baseline prior one may also consider using the normal approximation in (2.12) for the entire fraction appearing in (3.2), instead of using a Laplace approximation for the prior predictive $m_{\boldsymbol{\gamma}}^{\text{N}}(\mathbf{y}^*|\delta)$.

3.1 Gibbs variable selection under the PEP prior

The Gibbs variable selection (GVS; Dellaportas et al., 2002) method is a stochastic search algorithm based on the vector of binary indicators $\boldsymbol{\gamma} \in \{0,1\}^p$ which represents which of the p covariates are included in a model. To formulate GVS we need to partition the regression vector $\boldsymbol{\beta}$ into $(\boldsymbol{\beta}_\gamma, \boldsymbol{\beta}_{\setminus\gamma})$, corresponding to those components of $\boldsymbol{\beta}$ that are included and excluded from the model, i.e. $\beta_j \in \boldsymbol{\beta}_\gamma$ if $\gamma_j = 1$ and $\beta_j \in \boldsymbol{\beta}_{\setminus\gamma}$ if $\gamma_j = 0$, for $j = 1, \ldots, p$. As we assume that the intercept term is included in all models under consideration, $\boldsymbol{\beta}_\gamma$ and $\boldsymbol{\beta}_{\setminus\gamma}$ are of dimensionality $d_{\boldsymbol{\gamma}} = p_{\boldsymbol{\gamma}} + 1$ and $d_{\setminus\boldsymbol{\gamma}} = p - p_{\boldsymbol{\gamma}}$, respectively.

Under the GVS setting the joint prior of $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ is specified as follows

$$\pi(\boldsymbol{\beta}, \boldsymbol{\gamma}) = \pi_{\boldsymbol{\gamma}}^{\text{N}}(\boldsymbol{\beta})\pi(\boldsymbol{\gamma}) = \pi_{\boldsymbol{\gamma}}^{\text{N}}(\boldsymbol{\beta}_\gamma)\pi_{\boldsymbol{\gamma}}^{\text{N}}(\boldsymbol{\beta}_{\setminus\gamma})\pi(\boldsymbol{\gamma}), \tag{3.3}$$

where the actual baseline prior choice involves only $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, since $\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}})$ is just a *pseudo-prior* used for balancing the dimensions between model spaces; see Dellaportas et al. (2002). Suitable choices for the priors of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and $\boldsymbol{\gamma}$ have been discussed in Section 2.3, thus, in order to complete the GVS setup, we only need to specify the pseudo-prior for the inactive part of the regression vector $\boldsymbol{\beta}$. In particular, we use a multivariate normal distribution of dimensionality $d_{\setminus\boldsymbol{\gamma}}$, with parameters specified by the ML estimates; namely,

$$\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}) = \mathrm{N}_{d_{\setminus\boldsymbol{\gamma}}}\left(\widehat{\boldsymbol{\beta}}_{\setminus\boldsymbol{\gamma}}, \mathbf{I}_{d_{\setminus\boldsymbol{\gamma}}}\widehat{\sigma}^2_{\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}}\right), \tag{3.4}$$

where $\widehat{\boldsymbol{\beta}}_{\setminus\boldsymbol{\gamma}}$ and $\widehat{\sigma}_{\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}}$ 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. Based on this formulation, the full augmented posterior used to build our MCMC has the following form

$$\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}, \beta_0, \mathbf{y}^*, \boldsymbol{\gamma}|\mathbf{y}, \delta) \propto f_{\boldsymbol{\gamma}}(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}})\frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta} f_0(\mathbf{y}^*|\beta_0)^{1/\psi}}{m^{\mathrm{N}}_{\boldsymbol{\gamma}}(\mathbf{y}^*|\delta)}\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}})\pi(\boldsymbol{\gamma})\pi^{\mathrm{N}}_0(\beta_0), \tag{3.5}$$

where, as a reminder, $\psi = 1$ in the CR-PEP setting and $\psi = \delta$ in the DR-PEP setting.

Then, the proposed PEP-GVS sampling scheme is the following:

Set starting values $\boldsymbol{\gamma}^{(0)}, \boldsymbol{\beta}^{(0)} = (\boldsymbol{\beta}^{(0)}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}^{(0)}_{\setminus\boldsymbol{\gamma}}), \beta_0^{(0)}$ and $\mathbf{y}^{*(0)}$.

For iterations $t = 1, 2, ..., N$:

Step 1: Set current values equal to $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t-1)}$, $\beta_0 = \beta_0^{(t-1)}$ $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(t-1)}$ and $\mathbf{y}^* = \mathbf{y}^{*(t-1)}$.

Step 2: For $j = 1, 2, ..., p$, sample $\gamma_j \sim \pi\big(\gamma_j\big|\boldsymbol{\beta}, \boldsymbol{\gamma}_{\setminus j}, \mathbf{y}^*, \mathbf{y}, \delta\big)$ which is a Bernoulli distribution.

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 the active effects $\boldsymbol{\beta}_{\boldsymbol{\gamma}} \sim \pi\big(\boldsymbol{\beta}_{\boldsymbol{\gamma}}\big|\boldsymbol{\gamma}, \mathbf{y}^*, \mathbf{y}, \delta\big)$ using a Metropolis-Hastings step.

Step 5: Sample the inactive effects $\boldsymbol{\beta}^{(t)}_{\setminus\boldsymbol{\gamma}}$ from the pseudo-prior in (3.4).

Step 6: Sample β_0 from $\pi(\beta_0|\mathbf{y}^*, \psi) \propto f_0(\mathbf{y}^*|\beta_0)^{1/\psi}\pi^{\mathrm{N}}_0(\beta_0)$ using a Metropolis-Hastings step.

Step 7: Sample $\mathbf{y}^*$ from

$$\pi(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \beta_0, \boldsymbol{\gamma}, \delta, \psi) \propto \frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta} f_0(\mathbf{y}^*|\beta_0)^{1/\psi}}{m^{\mathrm{N}}_{\boldsymbol{\gamma}}(\mathbf{y}^*|\delta)}$$

using a Metropolis-Hastings step.

Step 8: Update the parameter values at iteration t as $\boldsymbol{\beta}^{(t)} = \boldsymbol{\beta}$, $\beta_0^{(t)} = \beta_0$ $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}$ and $\mathbf{y}^{*(t)} = \mathbf{y}^*$.

Note that the generation of $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}$ (Steps 2 and 5) is straightforward since the corresponding conditional distributions are of known form. For the rest of parameters, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, β_0 and $\mathbf{y}^*$, we use Metropolis-Hastings (M-H) steps. Details are provided next.

3.2 Implementation details

Concerning the binary inclusion indicators γ_j, the conditional posterior distribution $\pi\big(\gamma_j|\boldsymbol{\beta}, \boldsymbol{\gamma}_{\setminus j}, \mathbf{y}^*, \mathbf{y}, \delta\big)$ is a Bernoulli distribution with success probability $O_j/(1+O_j)$ and

$$O_j \;=\; \frac{f_{\boldsymbol{\gamma}_{j_1}}\big(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}_{j_1}}\big)}{f_{\boldsymbol{\gamma}_{j_0}}\big(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}_{j_0}}\big)} \times \left[\frac{f_{\boldsymbol{\gamma}_{j_1}}\big(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}_{j_1}}\big)}{f_{\boldsymbol{\gamma}_{j_0}}\big(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}_{j_0}}\big)}\right]^{1/\delta} \times \frac{\pi^{\mathrm{N}}_{\boldsymbol{\gamma}_{j_1}}(\boldsymbol{\beta})}{\pi^{\mathrm{N}}_{\boldsymbol{\gamma}_{j_0}}(\boldsymbol{\beta})} \times \frac{m^{\mathrm{N}}_{\boldsymbol{\gamma}_{j_0}}(\mathbf{y}^*|\delta)}{m^{\mathrm{N}}_{\boldsymbol{\gamma}_{j_1}}(\mathbf{y}^*|\delta)} \times \frac{\pi(\boldsymbol{\gamma}_{j_1})}{\pi(\boldsymbol{\gamma}_{j_0})}, \quad (3.6)$$

where $\boldsymbol{\gamma}_{j_1} = (\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})$, $\boldsymbol{\gamma}_{j_0} = (\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})$ and $\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}) = \pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}})$ for $\boldsymbol{\gamma} \in \{\boldsymbol{\gamma}_{j_1}, \boldsymbol{\gamma}_{j_0}\}$. All of the quantities involved in (3.6) are available in closed form expressions except of the marginal likelihood $m^{\mathrm{N}}_{\boldsymbol{\gamma}}(\mathbf{y}^*|\delta)$. The latter is estimated through the following Laplace approximation

$$\widehat{m}^{\mathrm{N}}_{\boldsymbol{\gamma}}(\mathbf{y}^*|\delta) = (2\pi\delta)^{d_{\boldsymbol{\gamma}}/2}|\mathbf{X}^T_{\boldsymbol{\gamma}}\mathbf{W}_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}^*_{\boldsymbol{\gamma}})\mathbf{X}_{\boldsymbol{\gamma}}|^{-1/2} f_{\boldsymbol{\gamma}}\big(\mathbf{y}^*|\widehat{\boldsymbol{\beta}}^*_{\boldsymbol{\gamma}}\big)^{1/\delta}\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}\big(\widehat{\boldsymbol{\beta}}^*_{\boldsymbol{\gamma}}\big), \quad (3.7)$$

where $\widehat{\boldsymbol{\beta}}^*_{\boldsymbol{\gamma}}$ is the MLE for data $\mathbf{y}^*$ given the configuration of $\boldsymbol{\gamma}$, $\delta\Big[\mathbf{X}^T_{\boldsymbol{\gamma}}\mathbf{W}_{\boldsymbol{\gamma}}(\widehat{\boldsymbol{\beta}}^*_{\boldsymbol{\gamma}})\mathbf{X}_{\boldsymbol{\gamma}}\Big]^{-1}$ is equal to minus the inverse Hessian matrix evaluated at $\widehat{\boldsymbol{\beta}}^*_{\boldsymbol{\gamma}}$ and $\mathbf{W}_{\boldsymbol{\gamma}}$ is the $n \times n$ diagonal matrix containing the GLM weights. Under a Jeffreys baseline prior for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, the Laplace approximation simplifies to $\widehat{m}^{\mathrm{N}}_{\boldsymbol{\gamma}}(\mathbf{y}^*|\delta) = (2\pi\delta)^{d_{\boldsymbol{\gamma}}/2} f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\widehat{\boldsymbol{\beta}}^*_{\boldsymbol{\gamma}})^{1/\delta}$.

For the active effects $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ of model $M_{\boldsymbol{\gamma}}$ and the intercept term β_0 of the reference model M_0, we use independence sampler M-H steps. Specifically, for $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ we generate new candidate values as

$$\boldsymbol{\beta}'_{\boldsymbol{\gamma}} \sim q(\boldsymbol{\beta}'_{\boldsymbol{\gamma}}) \equiv \mathrm{N}_{d_{\boldsymbol{\gamma}}}\left(\widehat{\boldsymbol{\beta}}^{\mathrm{all}}_{\boldsymbol{\gamma}}, \widehat{\Sigma}_{\boldsymbol{\beta}^{\mathrm{all}}_{\boldsymbol{\gamma}}}\right),$$

where $\widehat{\boldsymbol{\beta}}^{\mathrm{all}}_{\boldsymbol{\gamma}}$ is the ML estimate from a weighted regression on $\mathbf{y}^{\mathrm{all}} = (\mathbf{y}, \mathbf{y}^*)^T$, using weights $\mathbf{w}^{\mathrm{all}} = (\mathbf{1}_n, \mathbf{1}_n\delta^{-1})^T$, and $\widehat{\Sigma}_{\boldsymbol{\beta}^{\mathrm{all}}_{\boldsymbol{\gamma}}}$ is the estimated variance-covariance matrix of $\widehat{\boldsymbol{\beta}}^{\mathrm{all}}_{\boldsymbol{\gamma}}$. The proposed move is accepted with probability

$$\alpha_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}} = \min\left[1, \frac{f_{\boldsymbol{\gamma}}(\mathbf{y}|\boldsymbol{\beta}'_{\boldsymbol{\gamma}})}{f_{\boldsymbol{\gamma}}(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}})}\left(\frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}'_{\boldsymbol{\gamma}})}{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})}\right)^{1/\delta}\frac{\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}'_{\boldsymbol{\gamma}})q(\boldsymbol{\beta}_{\boldsymbol{\gamma}})}{\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}})q(\boldsymbol{\beta}'_{\boldsymbol{\gamma}})}\right],$$

where $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ denotes the current value of the chain. The proposal distribution of β_0 is $q(\beta_0) = \mathrm{N}(\widehat{\beta}_0, \psi\widehat{\sigma}^2_{\beta_0})$ with $\widehat{\beta}_0$ and $\widehat{\sigma}_{\beta_0}$ being the respective ML estimate of β_0 and the standard error of $\widehat{\beta}_0$ from the null model with response data $\mathbf{y}^*$. The proposed move is accepted with the usual M-H transition probability where the likelihood of the reference model is raised to the power of $1/\psi$. Note that no specific fine tuning is required for the proposal distributions of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and β_0.

Finally, for the generation of the imaginary data we propose candidate values $\mathbf{y}^{*'}$ from a proposal distribution $q(\mathbf{y}^{*'})$ and accept the proposed move with probability

$$\alpha_{\mathbf{y}^*} = \min\left[1, \left(\frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^{*'}|\boldsymbol{\beta}_{\boldsymbol{\gamma}})}{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})}\right)^{1/\delta}\left(\frac{f_0(\mathbf{y}^{*'}|\beta_0)}{f_0(\mathbf{y}^*|\beta_0)}\right)^{1/\psi}\frac{\widehat{m}^{\mathrm{N}}_{\boldsymbol{\gamma}}(\mathbf{y}^*|\delta)}{\widehat{m}^{\mathrm{N}}_{\boldsymbol{\gamma}}(\mathbf{y}^{*'}|\delta)}\frac{q(\mathbf{y}^*)}{q(\mathbf{y}^{*'})}\right],$$

where the marginal likelihood estimates are obtained through (3.7) and $\mathbf{y}^*$ denotes the current value of the chain. The joint proposal density is formed by the product of independent distributions, i.e. $q(\mathbf{y}^*) = \prod_{i=1}^{n^*} q(y_i^*)$, where the proposal of each imaginary observation y_i^* is constructed by combining the two likelihood components of the PEP prior. Hence, for the logistic regression model we use

$$q(y_i^*) \equiv \text{Binomial}(N_i, \pi_i^*) \text{ with } \pi_i^* = \frac{\pi_0^{1/\psi} \pi_{\boldsymbol{\gamma}(i)}^{1/\delta}}{\pi_0^{1/\psi} \pi_{\boldsymbol{\gamma}(i)}^{1/\delta} + (1-\pi_0)^{1/\psi}(1-\pi_{\boldsymbol{\gamma}(i)})^{1/\delta}},$$

where $\pi_0 = (1+\exp(-\beta_0))^{-1}$, $\pi_{\boldsymbol{\gamma}(i)} = (1+\exp(-\mathbf{X}_{\boldsymbol{\gamma}(i)}\boldsymbol{\beta}_{\boldsymbol{\gamma}}))^{-1}$ and N_i denotes the number of trials of the observed data. Equivalently, for Poisson regression models we consider

$$q(y_i^*) \equiv \text{Poisson}\left(\lambda_0 \lambda_{\boldsymbol{\gamma}(i)}^{1/\delta}\right)$$

for the CR-PEP prior; where $\lambda_0 = \exp(\beta_0)$ and $\lambda_{\boldsymbol{\gamma}(i)} = \exp(\mathbf{X}_{\boldsymbol{\gamma}(i)}\boldsymbol{\beta}_{\boldsymbol{\gamma}})$. For the DR-PEP prior, the corresponding choice of a Poisson proposal with mean $(\lambda_0\lambda_{\boldsymbol{\gamma}})^{1/\delta}$ was not found to be efficient in practice. Therefore, we use instead a Poisson random-walk proposal with mean equal to the value of y_i^* at the current iteration.

A complete and thorough description of the PEP-GVS algorithm as implemented in this work is provided in algorithmic form at the electronic appendix of this paper.

4 Hyper-δ extensions

The initial 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 2.1, the same representation holds asymptotically in the GLM setting given a flat baseline prior. From this perspective, 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. Under this approach one can estimate the power-parameter instead of a-priori set it equal to a fixed predefined value. It should be noted, however, that when δ is not fixed at n^*, then PEP priors loose their unit-information interpretation.

We define the hyper-δ CR/DR-PEP priors as

$$\pi_{\boldsymbol{\gamma}}^{\text{CR/DR-PEP}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) = \int\int f_{\text{N}_{d_{\boldsymbol{\gamma}}}}\left(\boldsymbol{\beta}_{\boldsymbol{\gamma}}; \widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^*, \delta\big(\mathbf{X}_{\boldsymbol{\gamma}}^{*T}\mathbf{W}_{\boldsymbol{\gamma}}^*(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^*)\mathbf{X}_{\boldsymbol{\gamma}}^*\big)^{-1}\right) m_0^{\text{N}}(\mathbf{y}^*|\psi)\pi(\delta)\text{d}\mathbf{y}^*\text{d}\delta, \quad (4.1)$$

where $\psi = \{1, \delta\}$, $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^*$ is the ML estimate given the imaginary data, and $f_{\text{N}_{d_{\boldsymbol{\gamma}}}}(\cdot)$ denotes the $d_{\boldsymbol{\gamma}}$–dimensional multivariate normal distribution. Note that for ease of exposition in (4.1), and without loss of generality, we use the normal approximation given in (2.12) for the baseline posterior $\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}|\mathbf{y}^*, \delta)$. Sensible options for $\pi(\delta)$ are the hyper-g analogues proposed in Liang et al. (2008). Specifically, we consider the hyper-δ prior

$$\pi(\delta) = \frac{a-2}{2}(1+\delta)^{-a/2}, \tag{4.2}$$

which corresponds to a $\mathrm{Beta}\big(1, \frac{a}{2}-1\big)$ for the shrinkage factor $\frac{\delta}{1+\delta}$. Thinking in terms of shrinkage, Liang et al. (2008) propose setting $a = 3$ in order to place most of the probability mass near 1 or $a = 4$ which leads to a uniform prior. An alternative option is the hyper-δ/n prior given by

$$\pi(\delta) = \frac{a-2}{2n}\left(1+\frac{\delta}{n}\right)^{-a/2}. \tag{4.3}$$

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, Berger, Forte and García-Donato (2012).

Of course, when working outside the context of the normal linear model the integration in (4.1) 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

$$\pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}, \beta_0, \mathbf{y}^*, \boldsymbol{\gamma}, \delta|\mathbf{y}) \propto \pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}, \beta_0, \mathbf{y}^*, \boldsymbol{\gamma}|\mathbf{y}, \delta)\pi(\delta), \tag{4.4}$$

where the first quantity in the right-hand side of (4.4) is given in (3.5). The corresponding full conditionals we wish to sample from are

$$\pi^{\mathrm{CR-PEP}}(\delta|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \beta_0, \phi, \boldsymbol{\gamma}, \mathbf{y}^*, \mathbf{y}) \propto \frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta}\pi(\delta)}{m^{\mathrm{N}}_{\boldsymbol{\gamma}}(\mathbf{y}^*|\delta)}, \tag{4.5}$$

$$\pi^{\mathrm{DR-PEP}}(\delta|\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \beta_0, \phi, \boldsymbol{\gamma}, \mathbf{y}^*, \mathbf{y}) \propto \frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta} f_0(\mathbf{y}^*|\beta_0)^{1/\delta}\pi(\delta)}{m^{\mathrm{N}}_{\boldsymbol{\gamma}}(\mathbf{y}^*|\delta)}. \tag{4.6}$$

Looking at the above expressions, a subtle point is that δ is not directly linked to the actual data $\mathbf{y}$; however, it is linked indirectly via the posterior values of the parameters of models $M_{\boldsymbol{\gamma}}$ (for both approaches) and M_0 (for the DR-PEP prior). Sampling from (4.5) or (4.6) is achieved by adding one simple step (after Step 7) in the PEP-GVS algorithm described in Section 3.1. Specifically, we use a random walk M-H step where we propose a candidate value δ' from

$$q(\delta'|\delta) = \mathrm{Gamma}\left(\frac{\delta^2}{s^2_\delta}, \frac{\delta}{s^2_\delta}\right),$$

which has mean equal to the current value δ and variance s^2_δ. The latter is a tuning parameter which can be specified appropriately in order to have an acceptance rate between

0.2 and 0.5, as recommended by Roberts and Rosenthal (2001). The value of $s_\delta^2 = \delta$ proved to be efficient in the examples presented in Section 5. Given this proposal, the new candidate δ' is accepted with probability $\alpha_\delta = \min(1, A_\delta)$, with A_δ given by

$$A_\delta = \left(\frac{\delta}{\delta'}\right)^{d_\gamma/2} \left[\frac{f_\gamma(\mathbf{y}^*|\boldsymbol{\beta}_\gamma)}{f_\gamma(\mathbf{y}^*|\widehat{\boldsymbol{\beta}}^*_\gamma)}\right]^{\left\{\frac{1}{\delta'}-\frac{1}{\delta}\right\}} f_0\big(\mathbf{y}^*\big|\beta_0\big)^{\left\{\frac{1}{\psi'}-\frac{1}{\psi}\right\}} \frac{\pi(\delta')}{\pi(\delta)} \frac{q(\delta|\delta')}{q(\delta'|\delta)},$$

where $\psi' = \psi = 1$ for the CR-PEP prior and $\psi' = \delta'$, $\psi = \delta$ for the DR-PEP prior. This acceptance probability is derived from the marginal likelihood Laplace approximation presented in (3.7), keeping only the terms that include δ. The analytic description of the PEP-GVS algorithm found in the electronic appendix includes the additional sampling step discussed here.

5 Illustrative examples

In this section we present first a simulation study for logistic and Poisson regression taking into account independent and correlated predictors as well as different levels of sparsity for the true model. We proceed with a simulation 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-PEP and DR-PEP priors (introduced in Section 2.2) and their hyper-δ and hyper-δ/n extensions (presented in Section 4) with parameter $a = 3$, which is one of the main options proposed in Liang et al. (2008). For all PEP prior configurations we consider $n^* = n$ and $\mathbf{X}^*_\gamma = \mathbf{X}_\gamma$, where the columns of the design matrix are centered around their corresponding sample means. For fixed δ we consider the default unit-information approach, i.e. $\delta = n^*$. Jeffreys' prior for GLMs, given in (2.25), is used as baseline prior for $\boldsymbol{\beta}_\gamma$.

We compare the PEP variants with standard g-prior methods, using the GLM g-prior formulation of Bové and Held (2011) for the parameters of the predictor variables and a flat improper prior for the intercept term. In particular, we consider the unit-information g-prior (with $g = n$) and three mixtures of g-priors; namely, the hyper-g and hyper-g/n priors with $a = 3$ (Liang et al., 2008), and the beta mixture proposed by Maruyama and George (2011). Henceforth, the latter will be referred to as MG hyper-g. Note that for the MG hyper-g prior we only consider the part of the methodology that concerns the mixing density for g; we do not utilize the "generalized" g-prior design proposed in the same study because the application of this prior in the GLM framework is not straightforward. Stochastic model search under these approaches is also implemented via GVS sampling.

5.1 Simulation study 1

In this first example we consider two simulation scenarios for logistic and Poisson regression, presented in Hansen and Yu (2003) and Chen et al. (2008), respectively. Both of these scenarios are also considered by Li and Clyde (2015). 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

$$\text{corr}(X_i, X_j) = r^{|i-j|}, \ \ 1 \leq i < j \leq p.$$

Concerning the correlations between covariates we examine two cases: (i) independent predictors ($r = 0$) and (ii) correlated predictors ($r = 0.75$). In addition, four sparsity scenarios are assumed; the true data-generating models are summarized in Table 1. For the logistic case we use the same sample size as in Hansen and Yu (2003), namely $n = 100$, but with much lower effects in order to reflect more realistic values of odds ratios, thus, reducing the signal from the generated data. Given the coefficients in Table 1, the odds ratios are approximately 2, 2.5 and 3.5 for the sparse, medium and full models, respectively. For the Poisson simulation we use the same regression coefficients as in Chen et al. (2008), but with sample size equal to $n = 100$. Each simulation is repeated 100 times.

Scenario	**Logistic** ($n = 100$)						**Poisson** ($n = 100$)			
	β_0	β_1	β_2	β_3	β_4	β_5	β_0	β_1	β_2	β_3
null	0.1	0	0	0	0	0	-0.3	0	0	0
sparse	0.1	0.7	0	0	0	0	-0.3	0.3	0	0
medium	0.1	1.6	0.8	-1.5	0	0	-0.3	0.3	0.2	0
full	0.1	1.75	1.5	-1.1	-1.4	0.5	-0.3	0.3	0.2	-0.15

Table 1: Logistic and Poisson regression scenarios for Simulation Study 1 using independent ($r = 0$) and correlated predictors ($r = 0.75$).

As the number of predictors in both regression models is small we assign a uniform prior on model space as given in (2.26). 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 2. The comparison between the PEP prior approaches versus the rest of the methods indicates the following:

i) Overall the PEP based variable selection procedures perform well, since in 5 out of the 8 simulation scenarios the "best" prior for identifying the true model is at least one of the PEP priors.

ii) The PEP procedures perform better under the null and sparse simulation scenarios.

iii) Under the medium model scenario the PEP priors perform equally well to the other methods in the case of independent predictors and slightly worse in the case of correlated predictors.

iv) Under the full model scenario g-prior methods perform better than PEP priors. This is no surprise as PEP priors tend to support more parsimonious solutions in general.

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 2: Number of simulated samples (over 100 replications) that the MAP model coincides with the true model in the logistic case of Simulation Study 1 (row-wise largest value in bold).

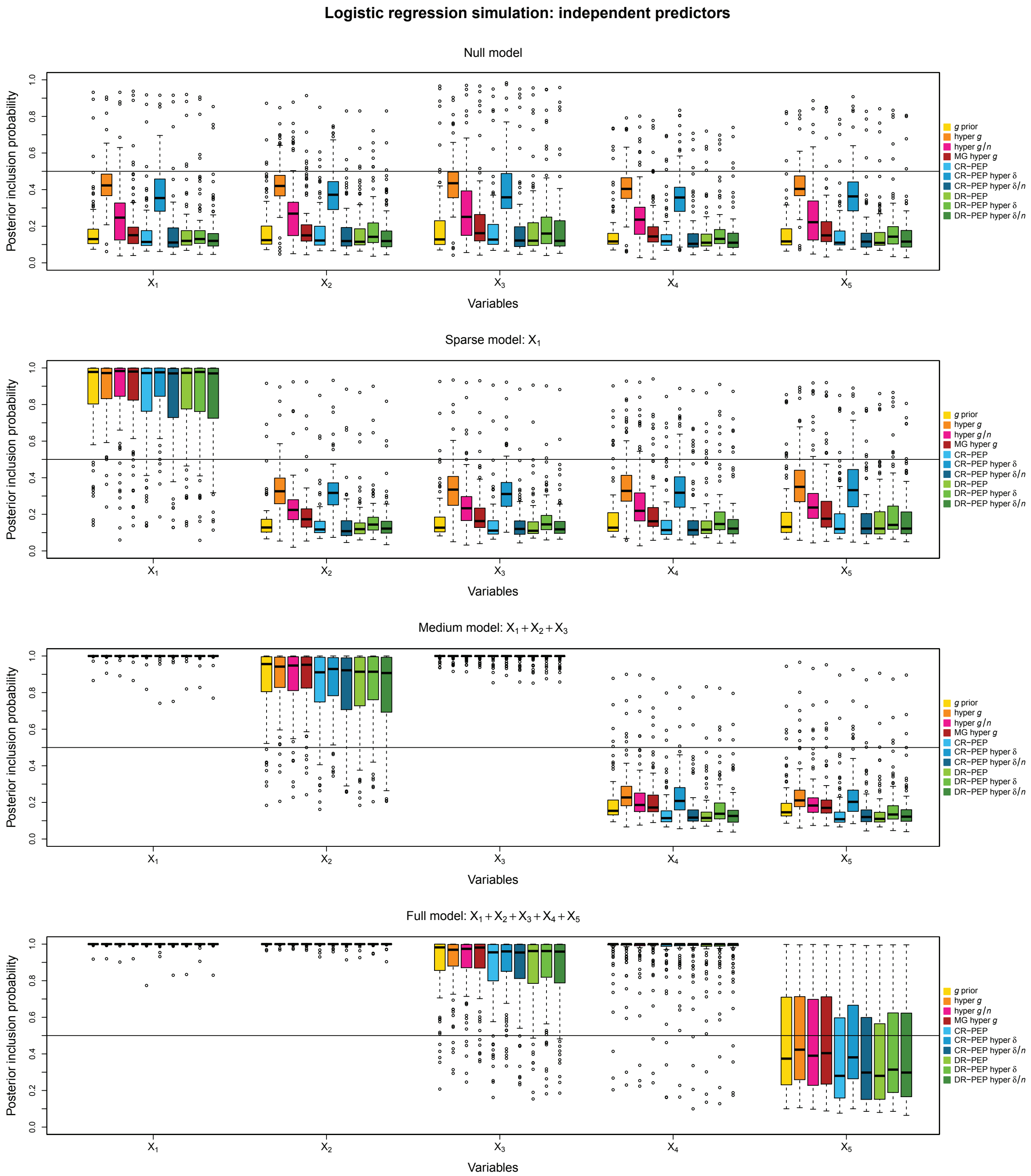


Figure 1: Posterior inclusion probabilities for Simulation Study 1 from 100 replicated samples of the null, sparse, medium and full logistic regression model scenarios with independent predictors ($r = 0$).

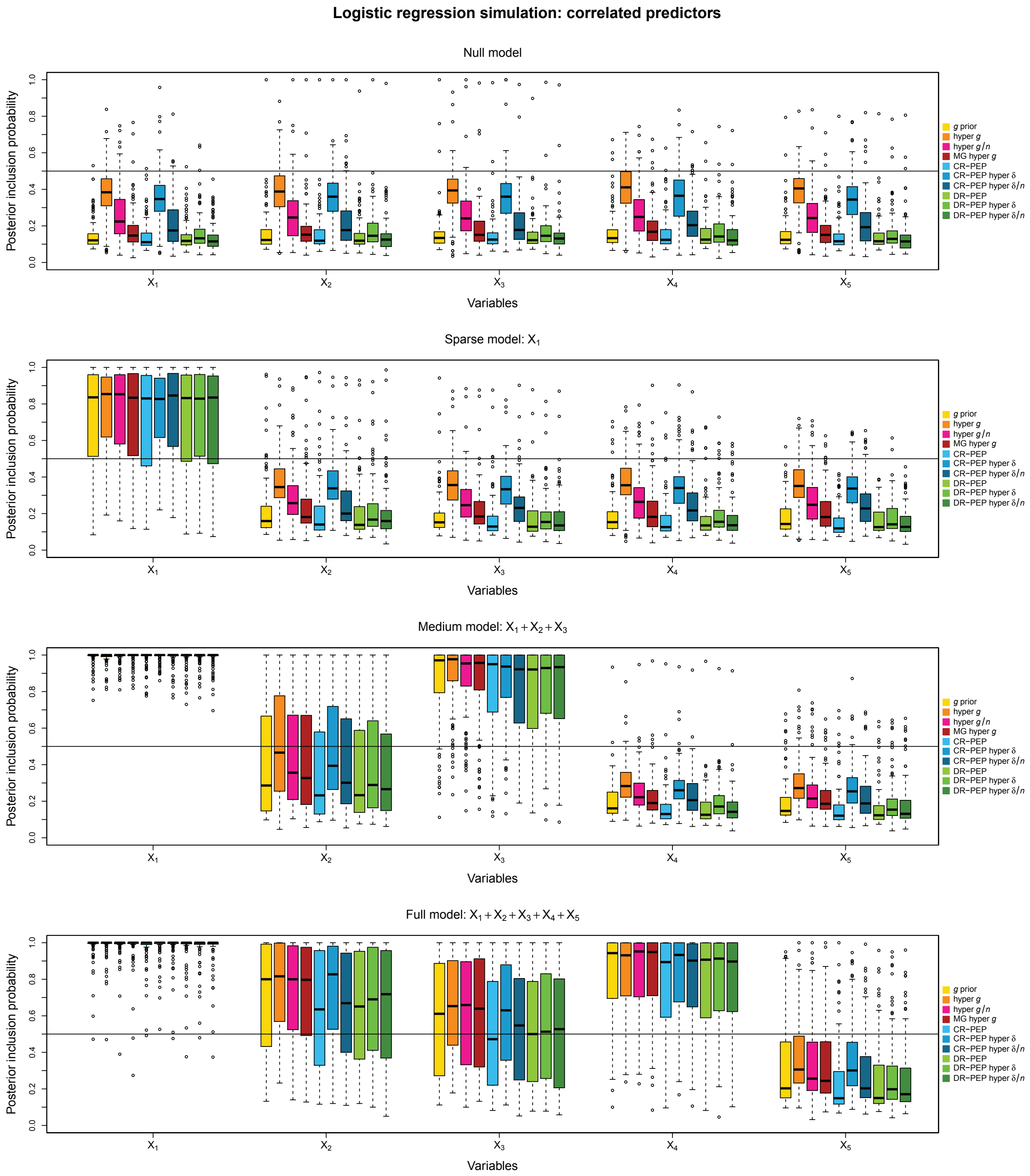


Figure 2: Posterior inclusion probabilities for Simulation Study 1 from 100 replicated samples of the null, sparse, medium and full logistic regression model scenarios with correlated predictors ($r = 0.75$).

With respect to the comparison between the CR-PEP and DR-PEP priors we find 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 the results are more or less the same in terms of MAP model success patterns. However, this is not the case under the CR-PEP approach as the hyper-δ prior seems to provide more support to complex models than the fixed-δ prior, while the hyper-δ/n prior is somewhere in the middle. Interestingly, a similar pattern is observed among the g-prior and the hyper-g, hyper-g/n priors. This pattern is identified more clearly by examining the resulting posterior inclusion probabilities; the corresponding boxplots under each method and simulation scenario are presented in Figure 1 for the case of independent predictors and in Figure 2 for the case of correlated predictors. As we can see the DR-PEP design is quite robust with respect to the choice between fixed versus random δ. Also, within the category of g-prior mixtures the MG hyper-g prior seems to have the strongest shrinkage effect.

The MAP-model results from the Poisson simulations are presented in Table 3. Boxplots of posterior inclusion probabilities under each method and simulation scenario are presented in Figure 3 for the case of independent predictors and in Figure 4 for the case of correlated predictors. Overall, conclusions similar to the logistic case can be drawn. Specifically, looking at the differences between the PEP priors and the various g-priors, we conclude to the following:

i) The PEP procedures perform overall satisfactory; in this example 6 out of the 8 best MAP success patterns are achieved by one of the PEP priors.

ii) The PEP procedures perform overall well under sparse conditions, i.e. when the true model is either the null model or the sparse model.

iii) For the model of medium complexity, the hyper-g and hyper-δ CR-PEP priors yield the best results; however, success rates under the model with correlated predictors are very low for all methods.

iv) For the full model with independent covariates, the MAP success rates under all methods are quite low; the hyper-g has the highest rate but with the hyper-δ CR-PEP prior being close and rather competitive. For the full model with correlated covariates, all methods fail; the hyper-δ CR-PEP prior has the highest success rate which is only 3%.

With respect to the various PEP prior distributions the comparison in the Poisson case leads to the same findings as in the logistic regression case. 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. To a lesser extend the same holds for the CR-PEP hyper-δ/n prior.

As a final remark, we note that all priors yield lower MAP success rates under the null scenario of the logistic simulations compared to the corresponding rates observed in the Poisson simulations. On the other hand, under the full scenario, the MAP success rates are higher in the logistic simulations. This can be attributed to the fact that the regression coefficients in the Poisson simulation are quite smaller in absolute value than the corresponding coefficients of the logistic formulation; see Table 1.

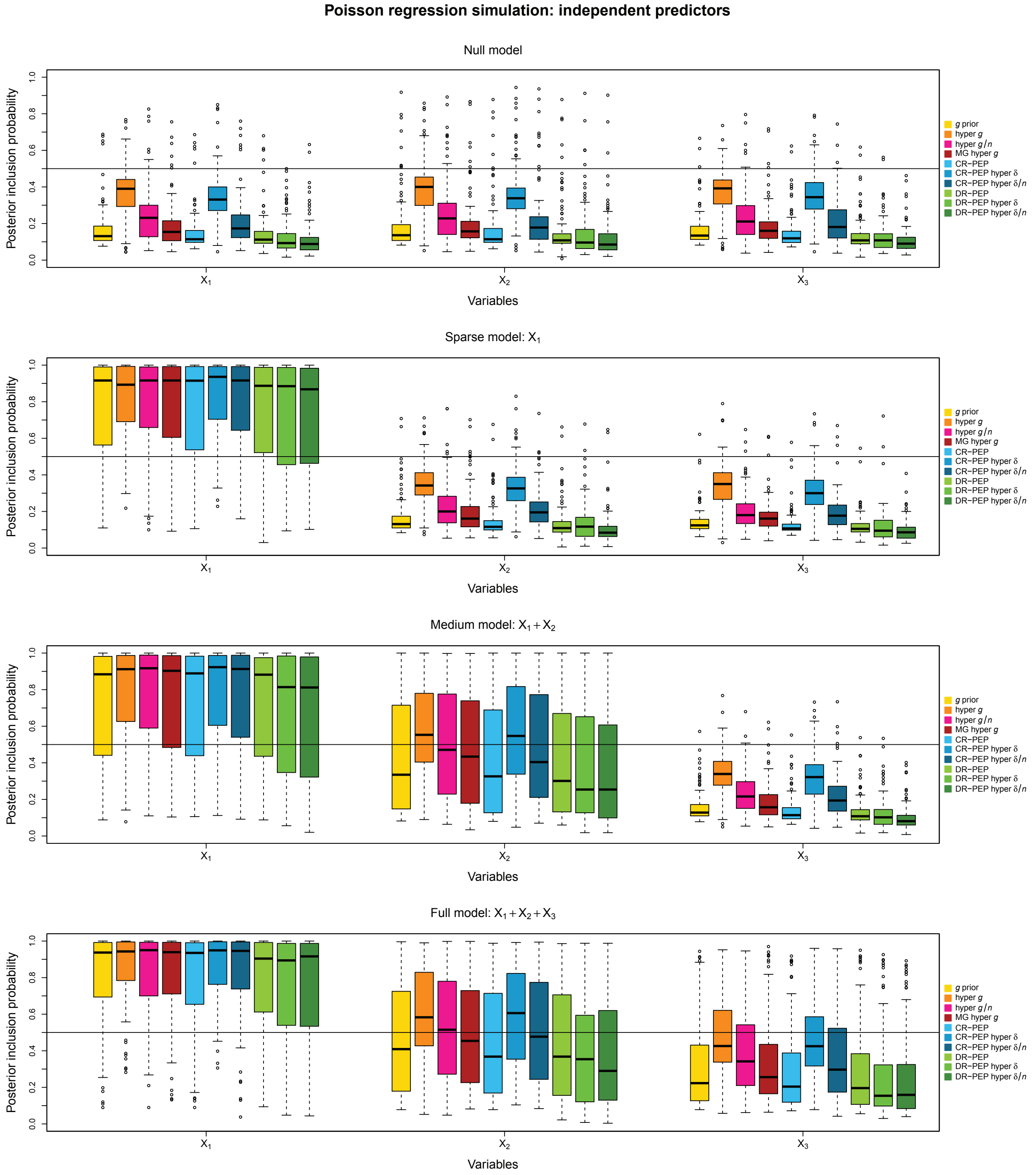


Figure 3: Posterior inclusion probabilities for Simulation Study 1 from 100 replicated samples of the null, sparse, medium and full Poisson model scenarios with independent predictors ($r = 0$).

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	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 3: Number of simulated samples (over 100 replications) that the MAP model coincides with the true model for the Poisson case in Simulation Study 1 (row-wise largest value in bold).

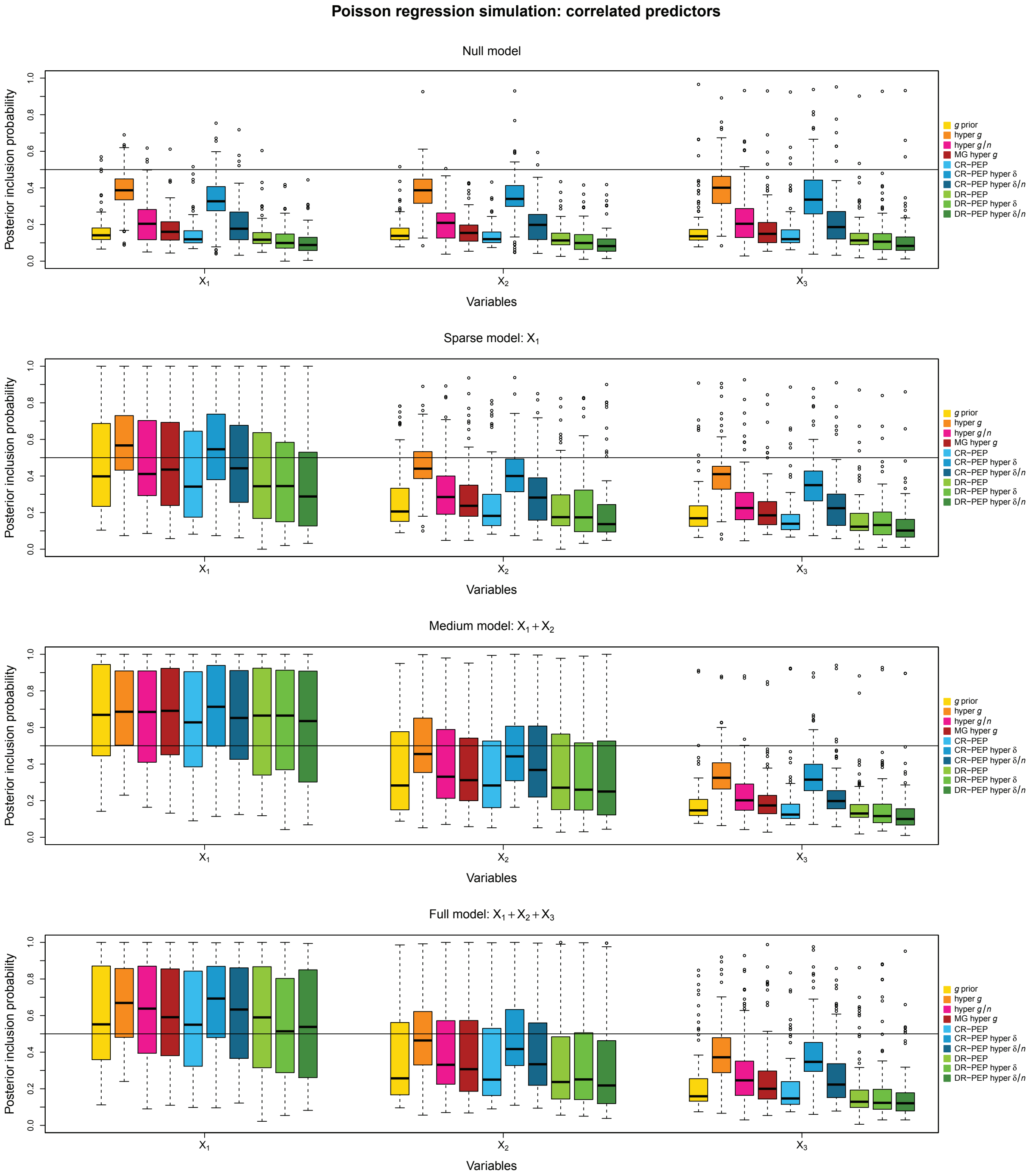


Figure 4: Posterior inclusion probabilities for Simulation Study 1 from 100 replicated samples of the null, sparse, medium and full Poisson model scenarios with correlated predictors ($r = 0.75$).

5.2 Simulation study 2

In this illustration we consider a more sophisticated scenario with $p = 10$ potential predictors (1024 models) and a more intriguing correlation structure. Similar to Nott and Kohn (2005), the first five covariates are generated from a standard normal distribution, while the remaining five covariates are generated from

$$X_{ij} = N(0.3X_{i1} + 0.5X_{i2} + 0.7X_{i3} + 0.9X_{i4} + 1.1X_{i5}, 1),$$

for $i = 1, \dots, n$ and $j = 6, \dots, 10$. We assume that sample size n is 200 and consider the three logistic regression data-generating models which are summarized in Table 4; the resulting odds ratios for the sparse and dense simulation models are approximately equal to 2 and 3, respectively. Each simulation is repeated 100 times.

Scenario	**Logistic** ($n = 200$)										
	β_0	β_1	β_2	β_3	β_4	β_5	β_6	β_7	β_8	β_9	β_{10}
null	0.1	0	0	0	0	0	0	0	0	0	0
sparse	0.1	0	0	-0.9	0	0	0	1.2	0	0	0.4
dense	0.1	0.6	0	-0.9	0	1	0.9	1.2	-1.2	-0.5	0

Table 4: Three logistic simulation scenarios for Simulation Study 2.

In this example we use the beta-binomial prior with a Beta(1,1) mixing distribution (Scott and Berger, 2010); see (2.27). The comparison that follows is based on the posterior inclusion probability of each covariate. Figures 5, 6 and 7 present boxplots of the posterior inclusion probabilities from the 100 simulated data sets for the null, sparse and dense simulation scenarios, respectively.

Under the null scenario, all priors exhibit good shrinkage effects except of the hyper-g prior which yields relatively large posterior inclusion probabilities with considerably higher variability. The hyper-δ CR-PEP prior also induces more variability, however, the resulting inclusion probabilities under this method are quite lower in comparison to those obtained from the hyper-g prior.

Under the sparse simulation scenario (true model: $X_1+X_7+X_{10}$), there are no striking differences among methods. All priors provide very strong support for the inclusion of X_7 and sufficient support for the inclusion of X_3, although the variability under PEP priors is larger for the latter variable. Also, all methods yield very wide posterior inclusion probability intervals for predictor X_{10}, thus leaving a lot of uncertainty concerning the inclusion of this variable. For the non-important variables we observe that the fixed-δ CR-PEP and the DR-PEP priors yield the lowest posterior inclusion probabilities.

Finally, in the dense simulation scenario (Figure 7), where the true model is $X_1+X_3+X_5+X_6+X_7+X_8+X_9$, the fixed-$\delta$ PEP priors generally outperform other methods in terms of providing low posterior inclusion probabilities for the insignificant covariates X_2, X_4 and X_{10}. The g-prior and the hyper DR-PEP extensions yield similar posterior inclusion probabilities and generally perform well, however, they introduce some uncertainty concerning the inclusion of covariate X_4. The rest of the methods systematically support

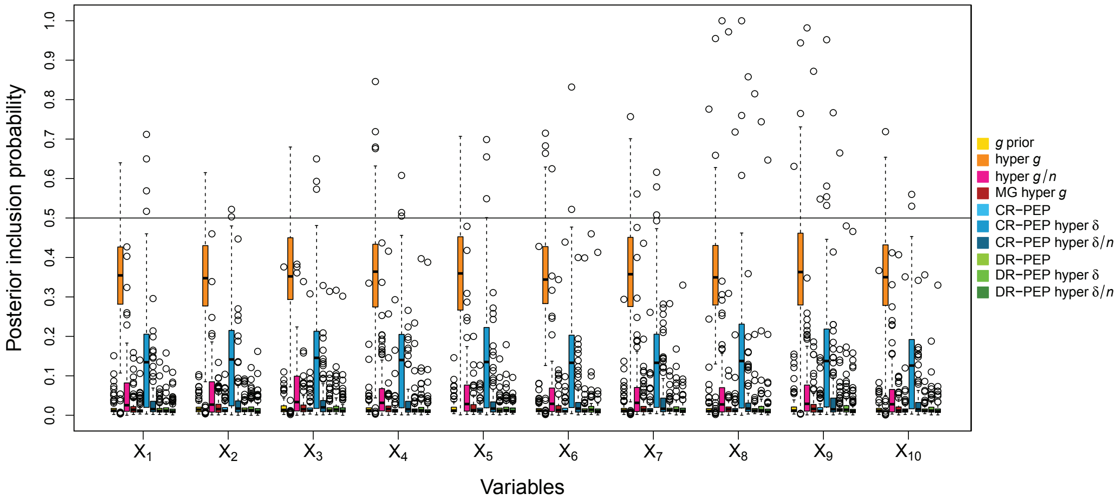


Figure 5: Posterior inclusion probabilities for Simulation Study 2 under the various priors from 100 repetitions of the null logistic simulation scenario.

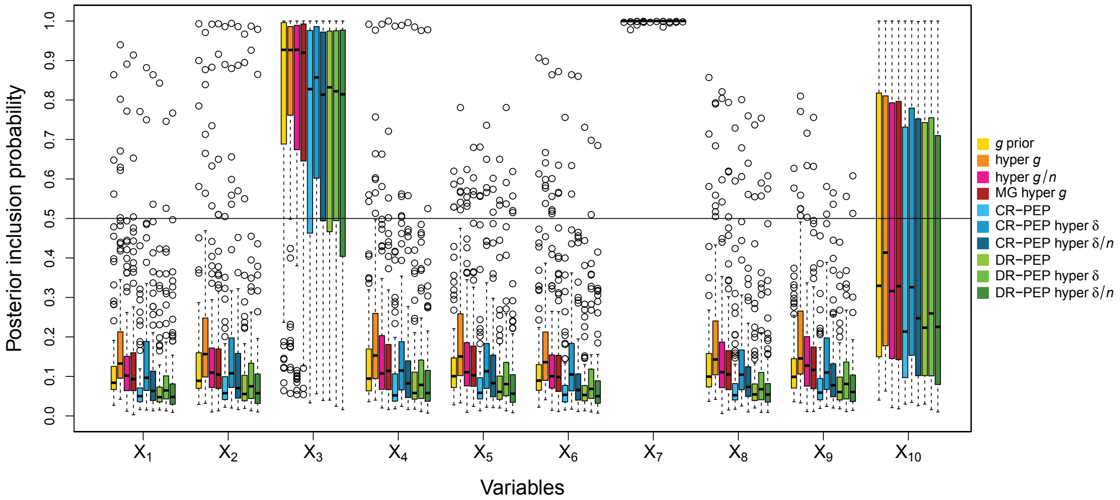


Figure 6: Posterior inclusion probabilities for Simulation Study 2 under the various priors from 100 repetitions of the sparse logistic simulation scenario where the true model is $X_3 + X_7 + X_{10}$.

more complex models as they provide elevated support for the inclusion of variables X_2 and X_4.

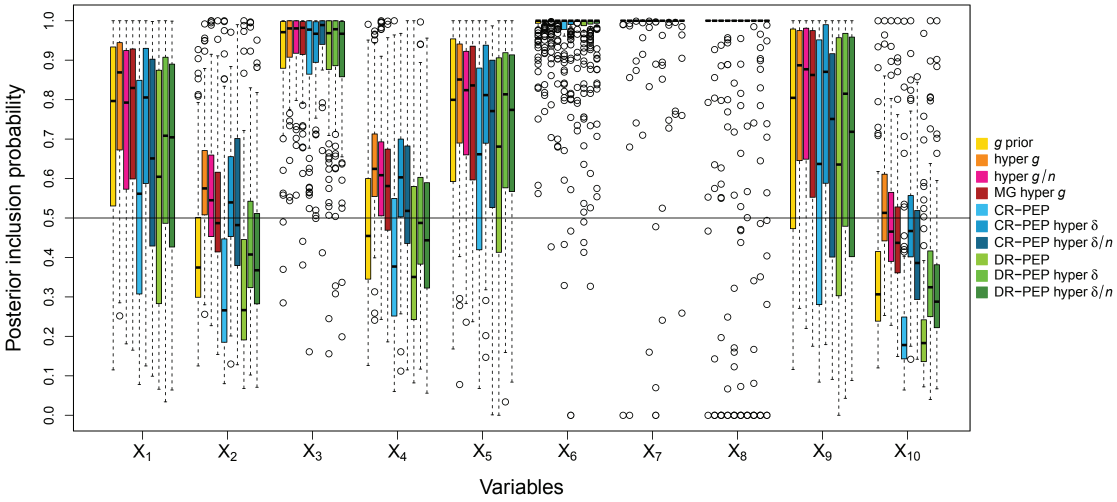


Figure 7: Posterior inclusion probabilities for Simulation Study 2 under the various priors from 100 repetitions of the dense logistic simulation scenario where the true model is $X_1 + X_3 + X_5 + X_6 + X_7 + X_8 + X_9$.

5.3 A real data example

In our last example we consider the Pima Indians diabetes data set (Ripley, 1996), which has been analyzed in several studies (e.g. Holmes and Held, 2006, Bové and Held, 2011). The data consist of $n = 532$ complete records on diabetes presence (present=1, not present=0) according to the WHO criteria for signs of diabetes. The presence of diabetes is associated with $p = 7$ potential covariates which are listed in Table 5.

For each method we used 41000 iterations of the GVS algorithm, discarding the first 1000 as burn-in period. We assigned a beta-binomial prior on model space (see Eq. 2.27). Table 6 shows the posterior inclusion probabilities of each covariate under the various methods. For comparison with the results presented in Bové and Held (2011), we also include in Table 6 the resulting posterior inclusion probabilities from the Zellner and Siow (1980) inverse gamma (ZS-IG) prior, the hyper-g/n with $a = 4$, and a non-informative inverse gamma (NI-IG) hyper-g prior with shape and scale equal to 10^{-3}. As seen, the posterior inclusion probabilities that we obtain from the GVS algorithm are in agreement with the results presented in Bové and Held (2011).

For the 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 posterior inclusion probabilities for 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

Covariate	Description
X_1	Number of pregnancies
X_2	Plasma glucose concentration (mg/dl)
X_3	Diastolic blood pressure (mm Hg)
X_4	Triceps skin fold thickness (mm)
X_5	Body mass index (kg/m^2)
X_6	Diabetes pedigree function
X_7	Age

Table 5: Potential predictors in the Pima Indians diabetes data set.

leads to a drastic reduction in the effects (and the inclusion probabilities) of low-influential covariates. On the other hand, the priors with random g or δ clearly result in higher posterior inclusion probabilities. Among this category of priors, the hyper-δ/n DR-PEP is evidently the most parsimonious, as it yields posterior inclusion probabilities which are actually quite close to those obtained from fixed δ PEP priors.

The uncertainty of the estimated posterior inclusion probabilities, for the standard methods considered in the previous examples, is depicted in Figure 8, where we present the corresponding boxplots produced by splitting the posterior samples into 40 batches of size 1000. As seen in Figure 8, introducing stochasticity to g and δ mainly affects the posterior inclusion probabilities of the "uncertain" covariates X_3, X_4 and X_7. For these variables the extra prior uncertainty induces higher posterior variability, as expected, and consequently larger Monte Carlo errors. Apart from that we observe once again the same patterns evident in the results of Sections 5.1 and 5.2. Among the category of g-prior

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 ($a = 3$)	0.970	1.000	0.397	0.379	0.998	0.996	0.669
hyper-g/n ($a = 3$)	0.966	1.000	0.304	0.300	0.998	0.995	0.579
hyper-g/n ($a = 4$)	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 data set.

mixtures, we see that in terms of shrinkage the MG hyper-g prior induces the strongest effect, followed by the hyper-g/n prior which has a stronger shrinkage effect in comparison to the hyper-g prior. Similarly, posterior inclusion probabilities under the hyper-δ/n PEP priors are lower than those resulting from the hyper-δ PEP priors. In addition, the DR design leads to a more stringent control for inclusion of variables in relation to the CR prior design.

Figure 9 shows convergence plots and the estimated posterior distribution of the shrinkage parameter $\delta/(1+\delta)$ under the four PEP hyper-prior approaches. The posterior histograms are indicative of the behavior of the shrinkage parameter. Comparison between the hyper-δ (Figure 9a) and the hyper-δ/n (Figure 9b) approaches shows that the posterior distribution of the shrinkage parameter under the latter priors is more concentrated to values close to one, thus, resulting to a stronger shrinkage effect. Also, the histograms in Figure 9a and 9b indicate that the posterior distributions of the shrinkage parameter under DR-PEP are more concentrated to values close to one in comparison to the corresponding posteriors under CR-PEP. Note that the shrinkage under the fixed-δ approaches is constant, equal to 0.998, which leads to considerably lower posterior inclusion probabilities as seen in Table 6 and Figure 8.

We conclude this example by examining the out-of-sample predictive accuracy of the

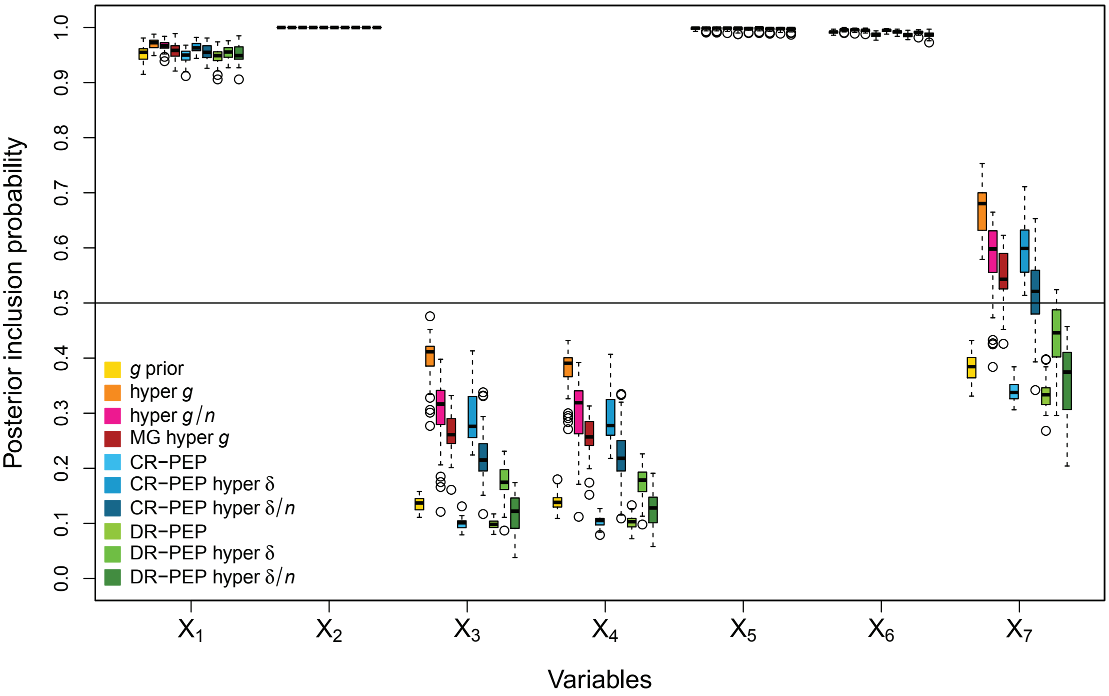
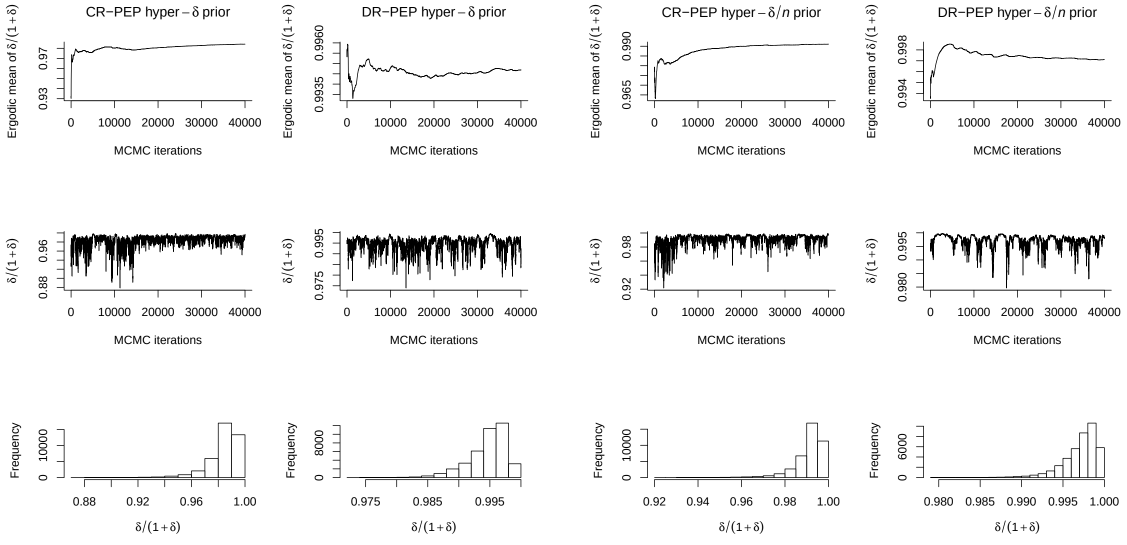


Figure 8: Boxplots of batched estimates of the posterior inclusion probabilities for the seven predictors in the Pima Indians dataset based on 40 batches of size 1000.

various prior designs. To this end we kept at random half of the data set in order to re-do the analysis based again on 41000 iterations of the GVS algorithm, discarding the first 1000 as the burn-in period. From these simulations we located the corresponding MAP and medium probability models under each prior, and subsequently used a single-model M-H algorithm in order to sample from the corresponding posterior distributions of the MAP and medium probability models. Then, based on posterior samples of size equal to 40000, we generated an equal number of binary predictions using the part of the design matrix corresponding to the test sample, and counted the number of false negative and false positive predictions under the corresponding MAP and medium probability models of each prior. The average percentages of false negatives and false positives are summarized in Table 7. Overall, we cannot say that there is dominating method in terms of predictive accuracy as the predictions are more or less the same across the prior designs. We may note however that the most complex MAP model arises from the hyper-g prior and actually results in the highest false negative prediction rate. Also, the unit-information g-prior, the CR-PEP with fixed δ, and the DR-PEP priors lead to a more a parsimonious medium probability model which has comparable predictive accuracy with the more complex model that includes covariate X_7, resulting from the other methods.



(a) Hyper-δ Priors

(b) Hyper-δ/n Priors

Figure 9: Ergodic mean plots, time-series plots and histograms of the shrinkage factor $\delta/(1+\delta)$ for the hyper-δ and hyper-δ/n PEP priors based on 40000 draws.

Method	**MAP model**	**False Neg. (%)**	**False Pos. (%)**
g-prior ($g = n$)	$X_1 + X_2 + X_5 + X_6$	10.8	16.5
hyper-g ($a = 3$)	$X_1 + X_2 + X_3 + X_4 + X_5 + X_6 + X_7$	11.4	16.9
hyper-g/n ($a = 3$)	$X_1 + X_2 + X_5 + X_6$	11.0	16.6
MG hyper-g	$X_1 + X_2 + X_5 + X_6$	10.9	16.6
CR-PEP	$X_1 + X_2 + X_5 + X_6$	10.9	16.9
CR-PEP hyper-δ	$X_1 + X_2 + X_5 + X_6$	10.9	17.0
CR-PEP hyper-δ/n	$X_1 + X_2 + X_5 + X_6$	10.8	17.0
DR-PEP	$X_1 + X_2 + X_5 + X_6$	10.9	16.8
DR-PEP hyper-δ	$X_1 + X_2 + X_5 + X_6$	10.9	16.9
DR-PEP hyper-δ/n	$X_1 + X_2 + X_5 + X_6$	10.9	16.8
Method	**Medium probability model**	**False Neg. (%)**	**False Pos. (%)**
g-prior ($g = n$)	same as MAP model		
hyper-g ($a = 3$)	$X_1 + X_2 + X_5 + X_6 + X_7$	11.1	16.8
hyper-g/n ($a = 3$)	$X_1 + X_2 + X_5 + X_6 + X_7$	11.0	16.6
MG hyper-g	$X_1 + X_2 + X_5 + X_6 + X_7$	10.9	16.6
CR-PEP	same as MAP model		
CR-PEP hyper-δ	$X_1 + X_2 + X_5 + X_6 + X_7$	11.3	16.4
CR-PEP hyper-δ/n	$X_1 + X_2 + X_5 + X_6 + X_7$	11.0	16.6
DR-PEP	same as MAP model		
DR-PEP hyper-δ	same as MAP model		
DR-PEP hyper-δ/n	same as MAP model		

Table 7: Percentages of false negative and false positive detections for the Pima Indian data set under the MAP model and medium probability model for the various priors.

6 Discussion

In this paper we presented an objective, automatic and compatible across competing models Bayesian procedure with applications to the variable selection problem in GLMs. Specifically we extended the PEP prior formulation through the use of unnormalized power-likelihoods and defined two new PEP priors, called CR-PEP and DR-PEP, which differentiate with respect to the definition of the prior predictive distribution of the reference model. Under the new definitions, the applicability of the PEP methodology is significantly enhanced. Although we focused on the variable selection problem in GLMs, the CR/DR-PEP priors proposed here may in principle be used for any general model setting. At the same time the new approaches retain the desired features of the original prior formulation; specifically, i) they resolve the problem of selecting and averaging across minimal training samples, thus, also allowing for large-sample approximations, and ii) they are minimally informative as they scale down the effect of the imaginary data on the posterior distribution. A further research direction that was pursued relates to the assignment of hyper-prior distributions to the power-parameter δ that controls the contribution of the imaginary data. Specifically, following the hyper-g and g/n priors proposed in Liang et al. (2008), we effectively introduced the hyper-δ and δ/n analogues.

The empirical results presented in this paper suggest that the proposed PEP priors outperform mixtures of g-priors in terms of introducing larger shrinkage to non-influential or to partially influential predictors, thus, leading to more parsimonious solutions with comparable predictive accuracy. With respect to the comparison between fixed δ vs. random δ PEP priors, the results indicate that the fixed unit-information approach induces more stringent control in the inclusion of predictors and therefore assigns more support to simpler models which is a desirable feature when having to select among a large number of potential predictors. Concerning the choice between the CR and the DR prior designs, we conclude in favouring the use of the DR-PEP as it seems that this prior is rather robust with respect to the fixed vs. random specification of the power-parameter, and this also translates to better shrinkage properties.

In the near future, we aim to investigate further the theoretical properties of the PEP extensions. So far we have proofs, which are available in an earlier technical report (Perrakis, Fouskakis and Ntzoufras, 2015*b*), that both extensions result in model selection consistency for the case of the Gaussian linear model. We intend to provide similar proofs within the formal GLM framework, possibly overcoming the problem of analytical intractability through Laplace-based approximations. In addition, we are currently working on extensions of the PEP methodology to high-dimensional problems, that include the small n–large p case, by incorporating shrinkage priors (e.g. ridge and LASSO procedures) into the PEP design. To this end, another promising alternative is to embody the expectation-maximization variable selection approach of Ročková and George (2014) within the PEP prior.

Acknowledgements/Funding

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.

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.

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

Berger, J. O. and Pericchi, L. R. (1996*b*), 'The intrinsic Bayes factor for model selection and prediction', *Journal of the American Statistical Association* **91**, 109–122.

Bové, D. S. and Held, L. (2011), 'Hyper-g priors for generalized linear models', *Bayesian Analysis* **6**, 387–410.

Casella, G. and Moreno, E. (2006), 'Objective Bayesian variable selection', *Journal of the American Statistical Association* **101**, 157–167.

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.

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

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.

Clyde, M. A., Ghosh, J. and Littman, M. L. (2011), 'Bayesian adaptive sampling for variable selection and model averaging', *Journal of Computational and Graphical Statistics* **20**, 80–101.

Consonni, G. and Veronese, P. (2008), 'Compatibility of prior specifications across linear models', *Statistical Science* **23**, 332–353.

Dellaportas, P., Forster, J. J. and Ntzoufras, I. (2002), 'On Bayesian model and variable selection using MCMC', *Statistics and Computing* **12**, 27–36.

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 (forthcoming); arXiv:1307.2449 [stat.CO]* .

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

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.

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.

Hansen, M. and Yu, B. (2003), 'Minimum description length model selection criteria for generalized linear models', *Lecture Notes-Monograph Series* **6**, 145–163.

Holmes, C. C. and Held, L. (2006), 'Bayesian auxiliary variable models for binary and multinomial regression', *Bayesian Analysis* pp. 145–168.

Ibrahim, J. G. and Chen, M.-H. (2000), 'Power prior distributions for regression models', *Statistical Science* **15**, 46–60.

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.

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.

Leon-Novelo, L., Moreno, E. and Casella, G. (2012), 'Objective Bayes model selection in probit models', *Statistics in Medicine* **31**, 353–365.

Li, Y. and Clyde, M. A. (2015), 'Mixtures of g-priors in generalized linear models', *arXiv:1503.06913v1 [stat.ME]* .

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.

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

Maruyama, Y. and George, E. I. (2011), 'Fully Bayes factors with a generalized g-prior', *The Annals of Statistics* **39**, 2740–2765.

Moreno, E. and Girón, F. J. (2008), 'Comparison of Bayesian objective procedures for variable selection in linear regression', *Test* **17**, 472–490.

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, pp. 359–366.

Nott, D. J. and Kohn, R. (2005), 'Adaptive sampling for Bayesian variable selection', *Biometrika* **92**, 747–763.

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.

Pérez, J. M. and Berger, J. O. (2002), 'Expected-posterior prior distributions for model selection', *Biometrika* **89**, 491–511.

Perrakis, K., Fouskakis, D. and Ntzoufras, I. (2015*a*), 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', Springer Proceedings in Mathematics and Statistics, pp. 59–73.

Perrakis, K., Fouskakis, D. and Ntzoufras, I. (2015*b*), 'Variations of the power-conditional-expected-posterior prior for Bayesian variable selection in generalized linear models', *technical report available at arXiv:1508.00793v2 [stat.ME]* .

Ripley, B. (1996), *Pattern Recognition and Neural Networks*, Cambridge University Press, Cambridge.

Roberts, G. O. and Rosenthal, J. S. (2001), 'Optimal scaling for various Metropolis-Hastings algorithms', *Statistical Science* **16**, 351–367.

Ročková, V. and George, E. I. (2014), 'EMVS: The EM Approach to Bayesian variable selection', *Journal of the American Statistical Association* **109**, 828–846.

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.

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

Zellner, A. (1986), On assessing prior distributions and Bayesian regression analysis using g-prior distributions, *in* P. Goel and A. Zellner, eds, '*Bayesian Inference and Decision Techniques: Essays in Honor of Bruno de Finetti*', North-Holland, Amsterdam, pp. 233–243.

Zellner, A. and Siow, A. (1980), Posterior odds ratios for selected regression hypothesis (with discussion), *in* J. Bernardo, M. DeGroot, D. Lindley and A. Smith, eds, '*Bayesian Statistics*, Vol. 1', Oxford University Press, pp. 585–606 & 618–647 (discussion).

Electronic Appendix for the the paper entitled “Power-Expected-Posterior Priors in Generalized Linear Models”’ by D.Fouskakis, I.Ntzoufras and K.Perrakis.

The PEP-GVS algorithm

Given the posterior distribution in Eq. 3.5, with $\psi = 1$ for the CR-PEP prior and $\psi = \delta$ for the DR-PEP prior, the PEP-GVS sampler proceeds as follows:

A. Set starting values $\boldsymbol{\gamma}^{(0)}, \boldsymbol{\beta}^{(0)} = (\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(0)}, \boldsymbol{\beta}_{\setminus\gamma}^{(0)}), \beta_0^{(0)}$ and $\mathbf{y}^{*(0)}$. For fixed δ set $\delta = n$, for random δ set starting starting value $\delta^{(0)}$.

B. For iterations $t = 1, 2, ..., N$:

Step 1: Sampling of $\gamma_j^{(t)}$, for $j = 1, 2, ..., p$, given the current state of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \boldsymbol{\beta}_{\setminus\gamma}, \gamma_{\setminus j}, \mathbf{y}^*$ and δ.

(a) Calculate the MLEs under $\boldsymbol{\gamma}_{j_1} = (\gamma_j = 1, \boldsymbol{\gamma}_{\setminus j})$, $\boldsymbol{\gamma}_{j_0} = (\gamma_j = 0, \boldsymbol{\gamma}_{\setminus j})$ and compute the Laplace approximations $\widehat{m}^{\mathrm{N}}_{\boldsymbol{\gamma}_{j_1}}(\mathbf{y}^*|\delta)$, $\widehat{m}^{\mathrm{N}}_{\boldsymbol{\gamma}_{j_0}}(\mathbf{y}^*|\delta)$ through Eq. 3.7.

(b) Evaluate the odds:

$$
\begin{aligned}
O_j \quad &= \quad \frac{f_{\boldsymbol{\gamma}_{j_1}}\left(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}_{j_1}}\right)}{f_{\boldsymbol{\gamma}_{j_0}}\left(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}_{j_0}}\right)} \left[\frac{f_{\boldsymbol{\gamma}_{j_1}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}_{j_1}})}{f_{\boldsymbol{\gamma}_{j_0}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}_{j_0}})}\right]^{1/\delta} \frac{\pi^{\mathrm{N}}_{\boldsymbol{\gamma}_{j_1}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}_{j_1}})}{\pi^{\mathrm{N}}_{\boldsymbol{\gamma}_{j_0}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}_{j_0}})} \frac{\pi^{\mathrm{N}}_{\boldsymbol{\gamma}_{j_1}}(\boldsymbol{\beta}_{\setminus\gamma_{j_1}})}{\pi^{\mathrm{N}}_{\boldsymbol{\gamma}_{j_0}}(\boldsymbol{\beta}_{\setminus\gamma_{j_0}})} \\
&\quad \times \frac{\widehat{m}^{\mathrm{N}}_{\boldsymbol{\gamma}_{j_0}}(\mathbf{y}^*|\delta)}{\widehat{m}^{\mathrm{N}}_{\boldsymbol{\gamma}_{j_1}}(\mathbf{y}^*|\delta)} \frac{\pi(\boldsymbol{\gamma}_{j_1})}{\pi(\boldsymbol{\gamma}_{j_0})}
\end{aligned}
$$

(c) Sample $\gamma_j^{'} \sim$ Bernoulli $\left(\frac{O_j}{1+O_j}\right)$ and set $\gamma_j^{(t)} = \gamma_j^{'}$ with probability equal to 1.

Step 2: Update $\boldsymbol{\beta}^{(t-1)} = (\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(t-1)}, \boldsymbol{\beta}_{\setminus\gamma}^{(t-1)})$ based on the current configuration of $\boldsymbol{\gamma}$.

Step 3: Sampling of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(t)}$ given the current state of $\boldsymbol{\gamma}, \mathbf{y}^*$ and δ.

(a) Generate $\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{'}$ from the proposal distribution $q(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{'}) = \mathrm{N}_{d_\gamma}(\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^{\mathrm{all}}, \widehat{\Sigma}_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{\mathrm{all}}})$, where $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^{\mathrm{all}}$ is the ML estimate from a weighted regression on $\mathbf{y}^{\mathrm{all}} = (\mathbf{y}, \mathbf{y}^*)^T$, using weights $\mathbf{w}^{\mathrm{all}} = (\mathbf{1}_n, \mathbf{1}_n\delta^{-1})^T$, and $\widehat{\Sigma}_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{\mathrm{all}}}$ is the estimated variance-covariance matrix of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{\gamma}}^{\mathrm{all}}$.

(b) Calculate the probability of move:

$$
\alpha_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}} = 1 \wedge \left[\frac{f_{\boldsymbol{\gamma}}(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{'})}{f_{\boldsymbol{\gamma}}(\mathbf{y}|\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(t-1)})} \left(\frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{'})}{f_{\boldsymbol{\gamma}}(\mathbf{y}^*|\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(t-1)})}\right)^{1/\delta} \frac{\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{'})}{\pi^{\mathrm{N}}_{\boldsymbol{\gamma}}(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(t-1)})} \frac{q(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(t-1)})}{q(\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{'})}\right].
$$

(c) Set $\boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(t)} = \begin{cases} \boldsymbol{\beta}_{\boldsymbol{\gamma}}^{'} & \text{with probability } \alpha_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}, \\ \boldsymbol{\beta}_{\boldsymbol{\gamma}}^{(t-1)} & \text{with probability } 1 - \alpha_{\boldsymbol{\beta}_{\boldsymbol{\gamma}}}. \end{cases}$

Step 4: Sampling of $\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}^{(t)}$ given the current state of $\boldsymbol{\gamma}$.

(a) Generate $\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}^{'}$ from the pseudo-prior $\pi_{\boldsymbol{\gamma}}^{\mathrm{N}}(\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}^{'}) = \mathrm{N}_{d_{\setminus\gamma}}\left(\widehat{\boldsymbol{\beta}}_{\setminus\boldsymbol{\gamma}}, \mathbf{I}_{d_{\setminus\gamma}}\widehat{\sigma}_{\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}}^2\right)$,where $\widehat{\boldsymbol{\beta}}_{\setminus\boldsymbol{\gamma}}$ and $\widehat{\sigma}_{\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}}$ are the respective MLEs and corresponding standard errors of $\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}$ from the full model given data $\mathbf{y}$.

(b) Set $\boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}^{(t)} = \boldsymbol{\beta}_{\setminus\boldsymbol{\gamma}}^{'}$ with probability equal to 1.

Step 5: Sampling of $\beta_0^{(t)}$ given the current state of $\mathbf{y}^*$ and δ.

(a) Generate $\beta_0^{'}$ from the proposal distribution $q(\beta_0^{'}) = \mathrm{N}(\widehat{\beta}_0, \psi\widehat{\sigma}_{\beta_0}^2)$, where $\widehat{\beta}_0$ and $\widehat{\sigma}_{\beta_0}$ are the respective MLE of β_0 and the standard error of $\widehat{\beta}_0$ from the null model given data $\mathbf{y}^*$.

(b) Calculate the probability of move:

$$\alpha_{\beta_0} = 1 \wedge \left[\frac{f_0(\mathbf{y}^*|\beta_0^{'})^{1/\psi}}{f_0(\mathbf{y}^*|\beta_0^{(t-1)})^{1/\psi}} \frac{\pi_0^{\mathrm{N}}(\beta_0^{'})}{\pi_0^{\mathrm{N}}(\beta_0^{(t-1)})} \frac{q(\beta_0^{(t-1)})}{q(\beta_0^{'})}\right].$$

(c) Set $\beta_0^{(t)} = \begin{cases} \beta_0^{'} & \text{with probability } \alpha_{\beta_0}, \\ \beta_0^{(t-1)} & \text{with probability } 1 - \alpha_{\beta_0}. \end{cases}$

Step 6: Sampling of $\mathbf{y}^{*(t)}$ given the current state of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, β_0, $\boldsymbol{\gamma}$ and δ.

(a) Generate $\mathbf{y}^{*'}$ from a proposal distribution $q(\mathbf{y}^{*'})$; see remark below for details about this proposal.

(b) Calculate the MLEs given $\mathbf{y}^{*(t-1)}$ and $\mathbf{y}^{*'}$ and compute the Laplace approximations $\widehat{m}_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^{*(t-1)}|\delta)$ and $\widehat{m}_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^{*'}|\delta)$ through Eq. 3.7.

(c) Calculate the probability of move:

$$\alpha_{\mathbf{y}^*} = 1 \wedge \left[\frac{f_{\boldsymbol{\gamma}}(\mathbf{y}^{*'}|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta}}{f_{\boldsymbol{\gamma}}(\mathbf{y}^{*(t-1)}|\boldsymbol{\beta}_{\boldsymbol{\gamma}})^{1/\delta}} \frac{f_0(\mathbf{y}^{*'}|\beta_0)^{1/\psi}}{f_0(\mathbf{y}^{*(t-1)}|\beta_0)^{1/\psi}} \frac{\widehat{m}_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^{*(t-1)}|\delta)}{\widehat{m}_{\boldsymbol{\gamma}}^{\mathrm{N}}(\mathbf{y}^{*'}|\delta)} \frac{q(\mathbf{y}^{*(t-1)})}{q(\mathbf{y}^{*'})}\right].$$

(d) Set $\mathbf{y}^{*(t)} = \begin{cases} \mathbf{y}^{*\prime} & \text{with probability } \alpha_{\mathbf{y}^*}, \\ \mathbf{y}^{*(t-1)} & \text{with probability } 1 - \alpha_{\mathbf{y}^*}. \end{cases}$

Step 7: Sampling of $\delta^{(t)}$ given the current state of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$, β_0 and $\boldsymbol{\gamma}$.

(a) If δ is fixed at n go to Step 1, else implement (b)-(e) of Step 7.

(b) Generate δ' from the proposal distribution $q(\delta'|\delta^{(t-1)}) = \mathrm{Gamma}(a, b)$ with $a = \delta^{(t-1)^2}/s_\delta^2$ and $b = \delta^{(t-1)}/s_\delta^2$.

(d) Calculate the probability of move:

$$\alpha_\delta = 1 \wedge \left[\left(\frac{\delta^{(t-1)}}{\delta'} \right)^{d_\gamma/2} \left(\frac{f_\gamma(\mathbf{y}^*|\boldsymbol{\beta}_\gamma)}{f_\gamma(\mathbf{y}^*|\widehat{\boldsymbol{\beta}}^*_\gamma)} \right)^{\left\{ \frac{1}{\delta'} - \frac{1}{\delta^{(t-1)}} \right\}} \right] \times$$
$$\times \left[f_0(\mathbf{y}^*|\beta_0)^{\left\{ \frac{1}{\psi'} - \frac{1}{\psi^{(t-1)}} \right\}} \frac{\pi(\delta')}{\pi(\delta^{(t-1)})} \frac{q(\delta^{(t-1)}|\delta')}{q(\delta'|\delta^{(t-1)})} \right],$$

where $\psi' = \psi^{(t-1)} = 1$ for the CR-PEP prior and $\psi' = \delta'$, $\psi^{(t-1)} = \delta^{(t-1)}$ for the DR-PEP prior.

(e) Set $\delta^{(t)} = \begin{cases} \delta' & \text{with probability } \alpha_\delta, \\ \delta^{(t-1)} & \text{with probability } 1 - \alpha_\delta. \end{cases}$

C. Repeat the steps in B until convergence.

Suggested proposals for Step 6: For $\mathbf{y}^*$ we recommend the following proposals depending on the likelihood of the model and on the PEP prior that is used:

i) For logistic regression a product binomial proposal distribution given by

$$q(\mathbf{y}^*) = \prod_{i=1}^{n^*} \text{Binomial}(N_i, \pi_i^*) \text{ with } \pi_i^* = \frac{\pi_0^{1/\psi} \pi_{\gamma(i)}^{1/\delta}}{\pi_0^{1/\psi} \pi_{\gamma(i)}^{1/\delta} + (1-\pi_0)^{1/\psi}(1-\pi_{\gamma(i)})^{1/\delta}},$$

where $\pi_0 = (1+\exp(-\beta_0))^{-1}$, $\pi_{\gamma(i)} = (1+\exp(-\mathbf{X}_{\gamma(i)}\boldsymbol{\beta}_\gamma))^{-1}$ and N_i denotes the number of trials of the observed data.

ii) For Poisson regression a product Poisson proposal distribution given by

$$q(\mathbf{y}^*) = \prod_{i=1}^{n^*} \text{Pois}(\lambda_i^*).$$

For the CR-PEP prior $\lambda_i^* = \lambda_0 \lambda_{\gamma(i)}^{1/\delta}$, where $\lambda_0 = \exp(\beta_0)$ and $\lambda_{\gamma(i)} = \exp(\mathbf{X}_{\gamma(i)}\boldsymbol{\beta}_\gamma)$. For the DR-PEP prior we utilize a random-walk proposal, i.e. $\lambda_i^* = y_i^{*(t-1)}$.