Variations of Power-Expected-Posterior Priors in Normal Regression Models

Dimitris Fouskakis[1], Ioannis Ntzoufras[2] and Konstantinos Perrakis[3]

[1]*Department of Mathematics, National Technical University of Athens, Zografou Campus, Athens Greece.*

[2]*Computational and Bayesian Statistics Lab, Department of Statistics, Athens University of Economics and Business, Greece.*

[3]*Department of Mathematical Sciences, Durham University, UK.*

Abstract

The power-expected-posterior (PEP) prior is an objective prior for Gaussian linear models, which leads to consistent model selection inference, under the M-closed scenario, and tends to favor parsimonious models. Recently, two new forms of the PEP prior were proposed which generalize its applicability to a wider range of models. The properties of these two PEP variants within the context of the normal linear model are examined thoroughly, focusing on the prior dispersion and on the consistency of the induced model selection procedure. Results show that both PEP variants have larger variances than the unit-information g-prior and that they are M-closed consistent as the limiting behavior of the corresponding marginal likelihoods matches that of the BIC. The consistency under the M-open case, using three different model misspecification scenarios is further investigated.

1 Introduction

1.1 Prelude

In this paper focus is given on the Bayesian variable selection problem in normal linear regression models. For every model $M_\ell \in \mathcal{M}$ (the set of all models under consideration, using different combinations of the available explanatory variables) the sampling distribution $f_\ell(\cdot|\boldsymbol{\beta}_\ell, \sigma^2, \mathbf{X}_\ell)$ is specified by

$$(\boldsymbol{Y}|\mathbf{X}_\ell, \boldsymbol{\beta}_\ell, \sigma^2, M_\ell) \sim N_n(\mathbf{X}_\ell \boldsymbol{\beta}_\ell, \sigma^2 \mathbf{I}_n), \tag{1.1}$$

where $\boldsymbol{Y} = (Y_1, \ldots, Y_n)$ is a vector containing the responses for all subjects, $\mathbf{X}_\ell$ is an $n \times d_\ell$ design matrix containing the values of the explanatory variables in its columns

($d_\ell = p_\ell + 1$; i.e. p_ℓ covariates plus the intercept), $\mathbf{I}_n$ is the $n \times n$ identity matrix, $\boldsymbol{\beta}_\ell$ is a vector of length d_ℓ summarizing the effects of the covariates in model M_ℓ on the response $\boldsymbol{Y}$ and σ^2 is the common error variance for all models M_ℓ. Finally, by p we denote the total number of the explanatory variables under consideration.

The Bayesian approach requires, for each model $M_\ell \in \mathcal{M}$, specification of prior densities $\pi_\ell(\boldsymbol{\theta}_\ell)$ for the unknown model specific parameters $\boldsymbol{\theta}_\ell = (\boldsymbol{\beta}_\ell, \sigma^2)$ and also specification of prior model probabilities $\pi(M_\ell)$. Then using the observed data $\mathbf{y} = (y_1, \ldots, y_n)$, focus is usually given on the posterior probability of each model $M_\ell \in \mathcal{M}$, defined as

$$\pi(M_\ell|\mathbf{y}) = \frac{m_\ell(\mathbf{y}|\mathbf{X}_\ell)\pi(M_\ell)}{\sum_{M_k \in \mathcal{M}} m_k(\mathbf{y}|\mathbf{X}_k)\pi(M_k)} = \left(\sum_{M_k \in \mathcal{M}} PO_{k,\ell} \right)^{-1} = \left[\sum_{M_k \in \mathcal{M}} BF_{k,\ell} \frac{\pi(M_k)}{\pi(M_\ell)} \right]^{-1}, \tag{1.2}$$

where $PO_{k,\ell} = \frac{\pi(M_k|\mathbf{y})}{\pi(M_\ell|\mathbf{y})}$ is the *posterior model odds* and $BF_{k,\ell} = \frac{m_k(\mathbf{y}|\mathbf{X}_k)}{m_\ell(\mathbf{y}|\mathbf{X}_\ell)}$ is the *Bayes factor*, for comparing any two models M_k and M_ℓ from $\mathcal{M}$. In (1.2) the quantity $m_\ell(\mathbf{y}|\mathbf{X}_\ell)$ is called the *marginal likelihood* (or *prior predictive distribution*) of model M_ℓ and is given by

$$m_\ell(\mathbf{y}|\mathbf{X}_\ell) = \int f_\ell(\mathbf{y}|\boldsymbol{\theta}_\ell, \mathbf{X}_\ell)\pi_\ell(\boldsymbol{\theta}_\ell)d\boldsymbol{\theta}_\ell \ .$$

Regarding the prior on the model space, Scott & Berger (2010) argue that prior model probabilities should take into consideration multiplicity issues inherent in model comparisons. When applied to variable selection problems, this principle can be implemented by assuming that, conditionally on a random probability of inclusion ω, each predictor can enter a model independently, so that $\pi(M_\ell|\omega) = \omega^{p_\ell}(1-\omega)^{n-p_\ell}$. Next, a hyper-prior is assigned to ω; in particular if $\omega \sim Beta(a_\omega, b_\omega)$, the resulting prior becomes

$$\pi(M_\ell) = \frac{B(a_\omega + p_\ell, b_\omega + p - p_\ell)}{B(a_\omega, b_\omega)}, \tag{1.3}$$

which is commonly known as the beta-binomial prior on model space. The default choice $a_\omega = b_\omega = 1$ results in a uniform distribution for ω. Under this specification, (1.3) reduces to

$$\pi(M_\ell) = \frac{1}{p+1}\binom{p}{p_\ell}^{-1}, \tag{1.4}$$

which induces a uniform prior on model size:

$$\pi\big(\{M_\ell \in \mathcal{M} : p_\ell = d\}\big) = 1/(p+1) \text{ for } d = 0, 1, \ldots, p.$$

Regarding the prior on model-specific regression parameters, most of the times we are *a-priori* uncertain about the validity of any competing model. This justifies the need for an objective model-selection approach in which vague prior information is assumed, which in turn motivates the use of either diffuse-proper priors or "default" improper priors. However, this immediately leads to further well understood difficulties. Speciffically, under proper prior distributions with large variances (diffuse priors), the resulting Bayes factors

can be highly sensitive to the chosen prior variances (Bartlett 1957). On the other hand, the use of default improper priors is also problematic, since such priors are defined only up to a constant multiple and, therefore, the Bayes factor is itself a multiple of this arbitrary constant. A variety of methods have been suggested for overcoming this problem; for a review see for example Consonni et al. (2018) and Forte et al. (2018).

One of the main approaches used to construct prior distributions for objective Bayes methods is the concept of *imaginary observations*. The basic idea, whose origin can be traced back to the work of Good (2004) is to consider a thought experiment with an appropriate dataset that will be used to specify the normalizing constants involved in the Bayes factors when using improper (*baseline*) priors (Spiegelhalter & Smith 1982). As we discuss in detail next, Pérez & Berger (2002) defined the expected-posterior-prior using this idea, which motivated subsequently Fouskakis et al. (2015) to introduce the power-expected-posterior prior.

1.2 Motivation

Pérez & Berger (2002) developed priors for objective Bayesian model comparison, through utilization of the device of "imaginary training samples". The *expected-posterior prior* (EPP) for the parameter under a model is an expectation of the posterior distribution given imaginary observations $\mathbf{y}^*$ of size n^*. The expectation is taken with respect to a suitable probability measure of a *reference* model M_0, while the posterior distribution is computed *via* Bayes's theorem starting from a default *baseline* prior, which is typically improper. One of the advantages of using EPPs is that impropriety of baseline priors causes no indeterminacy in the computation of Bayes factors. However, EPPs crucially depend on the size of the training sample. A consequence of that in variable selection problems, is that imaginary design matrices X^* (under each competing model) must be included in the analysis, with the resulting prior (under each model) further depending upon this choice; for a detailed discussion on this issue see Fouskakis et al. (2015). The selection of a *minimal training sample*, of size n^*, has been proposed (see for example Berger & Pericchi (2004)), to make the information content of the prior as small as possible, which is an appealing idea. However, even under this setup the resulting prior can be influential when sample size n is not much larger than the total number of parameters under the full model; see Fouskakis et al. (2015).

The *power-expected-posterior* (PEP) prior, introduced by Fouskakis et al. (2015), is an objective prior which amalgamates ideas from the power-prior (Ibrahim & Chen 2000), the expected-posterior prior (Pérez & Berger 2002) and the unit-information-prior approach of Kass & Wasserman (1995) to simultaneously (a) produce a minimally-informative prior and (b) diminish the effect of training samples under the EPP methodology. For a quick overview of the method, under normal linear models, see Section 2.1. The main idea is to substitute the likelihood by a *density-normalized version of a power-likelihood* in EPP; see e.g. Fouskakis et al. (2015), Fouskakis & Ntzoufras (2016).

A limitation of the original PEP formulation is that the normalization of the power-likelihood does not always lead to distributions of known form, e.g. if the data come

from a binomial or a Poisson distribution. Fouskakis et al. (2018) tackled this problem by introducing two alternative versions of the PEP prior (named CR-PEP and DR-PER; see Section 2.2 for more details under normal linear models). To understand what entails each of the two alternative versions of PEP, Fouskakis et al. (2018) used a simple intuitive example, where the parameter of an exponential distribution was tested using the reference prior as baseline. The original PEP prior, under this example, has the undesired property of getting more informative as the sample size grows; this is not the case under the two alternative PEP definitions.

Although the original PEP methodology under the normal linear model is well defined and studied, the properties of the two new PEP variations remain unexplored in this case. That is due to the fact that these priors were defined under the broad generalized linear model (GLM) framework which made the theoretical study of their properties hard because of well-known intractabilities. Therefore, a thorough study under the normal linear setting is necessary in order to validate the new PEP versions and understand their relation to other standard methods. Specifically, we examine from a theoretical perspective the following:

- the relationship of the new PEP versions with the original approach;
- model consistency under correct model specification;
- the prior information-volume and its effect on parsimony and sparsity.

Model selection consistency is typically investigated under the assumption of correct model specification, the so-called *M-closed* setting. In general, the consistency property of proposed model selection procedures under model-misspecification has not been subjected to as much scrutiny. This latter setting is commonly referred to as *M-open*; see Rossell & Rubio (2018) for a relevant recent work on this topic. Although the M-open case is not our initial motivation for this work, we further take into account this setting and investigate the robustness of the proposed PEP procedures (and of other popular methods) in simulations under model-misspecification.

The remainder of the paper is structured as follows. In Section 2 we sketch out the formulation of the PEP prior starting from its original form and then proceeding to the two new variants. Section 3 focuses on certain properties of the PEP variants for the normal linear model; specifically: (i) prior dispersion, and (ii) model selection consistency in the M-closed setting. In Section 4 we present results from simulation studies guided by two overarching themes; firstly, the M-closed vs. M-open cases, and secondly, the setting where model dimensionality grows with sample size. We conclude with a discussion in Section 5.

2 Background Information

2.1 Power Expected Posterior Priors

Fouskakis et al. (2015) and Fouskakis & Ntzoufras (2016) studied in detail the PEP priors

under the variable selection problem in Gaussian regression models. In the former paper a PEP prior is introduced for both the model-specific regression coefficients and the error variance, while in the latter paper another version of PEP is studied, named PCEP, where we have a conditional prior for the regression coefficients given the error variance, which is treated as a common nuisance parameter. Here we focus on the later case, where all posterior quantities can be derived analytically.

Specifically, we denote by $\pi_\ell^{\mathrm{N}}(\boldsymbol{\beta}_\ell\,,\,\sigma^2) = \pi_\ell^{\mathrm{N}}(\boldsymbol{\beta}_\ell\,|\sigma^2)\pi^N(\sigma^2)$ the baseline prior of the parameters of model M_ℓ. We assume that in $\mathcal{M}$ there exists a model M_0, with parameters $\boldsymbol{\beta}_0$ and σ^2, sampling distribution $f_0(\cdot|\boldsymbol{\beta}_0,\sigma^2)$ and baseline prior $\pi_0^{\mathrm{N}}(\boldsymbol{\beta}_0\,,\,\sigma^2) = \pi_0^{\mathrm{N}}(\boldsymbol{\beta}_0\,|\sigma^2)\pi^N(\sigma^2)$, which is nested into each of the remaining models and we consider it as a reference model. This is the typical case in the variable selection problem, studied in this paper. Given then a set of imaginary data $\mathbf{y}^* = (y_1^*,\ldots,y_{n^*}^*)^T$ and a positive power-parameter δ, that is used to regulate, essentially, the contribution of the imaginary data on the "final" prior, we introduce the density-normalized power-likelihood, under model M_ℓ, given by

$$f_\ell(\mathbf{y}^*|\boldsymbol{\beta}_\ell,\sigma^2,\delta,\mathbf{X}_\ell^*) = \frac{f_\ell(\mathbf{y}^*|\boldsymbol{\beta}_\ell,\sigma^2,\mathbf{X}_\ell^*)^{1/\delta}}{\int f_\ell(\mathbf{y}^*|\boldsymbol{\beta}_\ell,\sigma^2,\mathbf{X}_\ell^*)^{1/\delta}\mathrm{d}\mathbf{y}^*}. \tag{2.1}$$

The above density-normalized power-likelihood is still a normal distribution with variance inflated by a factor of δ; in the above $\mathbf{X}_\ell^*$ denotes the imaginary design matrix under model M_ℓ. In a similar manner, under the reference model, the density-normalized power-likelihood takes the form of (2.1) but using now the likelihood $f_0(\mathbf{y}^*|\boldsymbol{\beta}_0,\sigma^2,\mathbf{X}_0^*)$ of M_0.

In order to apply the PEP methodology, the density-normalized power-likelihood (2.1) is used to evaluate, under the imaginary data and the baseline prior, the conditional prior predictive distribution $m_0^{\mathrm{N}}(\mathbf{y}^*|\sigma^2,\delta,\mathbf{X}_0^*)$ of model M_0 as well as the conditional posterior distribution of $\boldsymbol{\beta}_\ell$

$$\pi_\ell^{\mathrm{N}}(\boldsymbol{\beta}_\ell|\mathbf{y}^*,\sigma^2,\delta,\mathbf{X}_\ell^*) = \frac{f_\ell(\mathbf{y}^*|\boldsymbol{\beta}_\ell,\sigma^2,\delta,\mathbf{X}_\ell^*)\pi_\ell^{\mathrm{N}}(\boldsymbol{\beta}_\ell\,|\sigma^2,\mathbf{X}_\ell^*)}{m_\ell^{\mathrm{N}}(\mathbf{y}^*|\sigma^2,\delta,\mathbf{X}_\ell^*)}, \tag{2.2}$$

where

$$m_j^{\mathrm{N}}(\mathbf{y}^*|\sigma^2,\delta,\mathbf{X}_j^*) \;\;=\;\; \int f_j(\mathbf{y}^*|\boldsymbol{\beta}_j,\sigma^2,\delta,\mathbf{X}_j^*)\pi_j^{\mathrm{N}}(\boldsymbol{\beta}_j\,|\sigma^2,\mathbf{X}_j^*)\mathrm{d}\boldsymbol{\beta}_j, \tag{2.3}$$

is the conditional prior predictive distribution of model M_j for $j=\ell,0$.

Finally, the imposed prior for the parameters of any model M_ℓ has the following hierarchical structure

$$\pi_\ell^{\mathrm{PEP}}(\boldsymbol{\beta}_\ell,\sigma^2|\delta,\mathbf{X}_\ell^*) = \pi_\ell^{\mathrm{PEP}}(\boldsymbol{\beta}_\ell|\sigma^2,\delta,\mathbf{X}_\ell^*)\pi^{\mathrm{N}}(\sigma^2|\mathbf{X}_\ell^*), \tag{2.4}$$

with

$$\pi_\ell^{\mathrm{PEP}}(\boldsymbol{\beta}_\ell|\sigma^2,\delta,\mathbf{X}_\ell^*) \;\;=\;\; \int \pi_\ell^{\mathrm{N}}(\boldsymbol{\beta}_\ell|\mathbf{y}^*,\sigma^2,\delta,\mathbf{X}_\ell^*)m_0^{\mathrm{N}}(\mathbf{y}^*|\sigma^2,\delta,\mathbf{X}_0^*)\mathrm{d}\mathbf{y}^*. \tag{2.5}$$

The default choice for δ is to set it equal to n^*, i.e. the sample size of the imaginary data, so that the overall information of the imaginary data in the posterior is equal to one data point. Furthermore, setting $n^* = n$ and, consequently, the design matrix of the imaginary data $\mathbf{X}_\ell^* \equiv \mathbf{X}_\ell$ simplifies significantly the overwhelming computations required when considering all possible "minimal" training samples (Pérez & Berger 2002) while it also avoids the complicated issue (in some cases) of defining the size of the minimal training samples (Berger & Pericchi 2004). In addition, under the choice $n^* = n$, the PEP prior remains relatively non-informative even for models with dimension close to the sample size n, while the effect on the evaluation of each model is minimal since the resulting Bayes factors are robust over different values of n^*. Detailed information about the default specifications of the PEP prior is provided in Fouskakis et al. (2015). Finally, the null model (with no explanatory variables) is a standard choice for the reference model in regression problems; see, for example Pérez & Berger (2002).

In the rest of the paper, we briefly introduce the two PEP variants, under the normal linear model case, and we study and compare their properties by examining their dispersion and sparsity as well as the consistency of the induced model selection procedures.

2.2 PEP prior variants

Fouskakis et al. (2018) introduced two alternative definitions of the PEP prior under the generalized linear model case. The core idea is to use the unnormalized power-likelihood $f_\ell(\mathbf{y}^*|\boldsymbol{\beta}_\ell, \sigma^2, \mathbf{X}_\ell^*)^{1/\delta}$ and normalize the posterior density instead. This was also the approach followed by Ibrahim & Chen (2000) and Friel & Pettitt (2008, Eq.4). Under this view, the posterior distribution inside the integral of (2.5) is now derived as

$$\pi_\ell^{\mathrm{N},\mathcal{U}}(\boldsymbol{\beta}_\ell|\mathbf{y}^*, \sigma^2, \delta, \mathbf{X}_\ell^*) = \frac{f_\ell(\mathbf{y}^*|\boldsymbol{\beta}_\ell, \sigma^2, \mathbf{X}_\ell^*)^{1/\delta}\pi_\ell^{\mathrm{N}}(\boldsymbol{\beta}_\ell|\sigma^2, \mathbf{X}_\ell^*)}{\int f_\ell(\mathbf{y}^*|\boldsymbol{\beta}_\ell, \sigma^2, \mathbf{X}_\ell^*)^{1/\delta}\pi_\ell^{\mathrm{N}}(\boldsymbol{\beta}_\ell|\sigma^2, \mathbf{X}_\ell^*)\mathrm{d}\boldsymbol{\beta}_\ell}. \tag{2.6}$$

As a subsequent of using (2.6) in (2.5), two alternative PEP variants were introduced in Fouskakis et al. (2018): a concentrated and a (more) diffuse version of PEP. The difference lies in the second component of the PEP definition given by (2.5), that is, the predictive distribution $m_0^{\mathrm{N}}(\mathbf{y}^*|\sigma^2, \delta, \mathbf{X}_0^*)$ of M_0 which is used to average the posterior distribution (2.6).

In the concentrated-reference PEP (CR-PEP) we consider the usual prior predictive distribution $m_0^{\mathrm{N}}(\mathbf{y}^*|\sigma^2, \mathbf{X}_0^*)$ of M_0 given by (2.3) for $\ell = 0$ and $\delta = 1$, that is,

$$m_0^{\mathrm{N},\mathcal{CR}}(\mathbf{y}^*|\sigma^2, \delta, \mathbf{X}_0^*) = \int f_0(\mathbf{y}^*|\beta_0, \sigma^2, \mathbf{X}_0^*)\pi_0^{\mathrm{N}}(\beta_0|\sigma^2, \mathbf{X}_0^*)\mathrm{d}\beta_0. \tag{2.7}$$

This approach adjusts the posterior distribution to account for n^*/δ data points but this (adjusted/power) posterior is averaged by using data from the actual predictive distribution of M_0 using data of size n^*.

On the other hand, for the construction of the diffuse-reference PEP (DR-PEP), we consider the prior predictive distribution of M_0 based on the unnormalized power-likelihood. This makes the approach more diffuse than CR-PEP in the sense that also

the predictive distribution results from a sample of size n^*/δ. Hence, in DR-PEP the predictive distribution used in (2.5) is given by

$$m_0^{\mathrm{N},\mathcal{D}R}(\mathbf{y}^*|\sigma^2,\delta,\mathbf{X}_0^*)\propto\int f_0(\mathbf{y}^*|\beta_0,\sigma^2,\mathbf{X}_0^*)^{1/\delta}\pi_0^{\mathrm{N}}(\beta_0|\sigma^2,\mathbf{X}_0^*)\mathrm{d}\beta_0 \tag{2.8}$$

where the above quantity is normalized in order $m_0^{\mathrm{N},\mathcal{D}R}(\mathbf{y}^*|\sigma^2,\delta,\mathbf{X}_0^*)$ to be a probability density function in terms of $\mathbf{y}^*$.

Hence the definition of the two versions of PEP priors can be summarized as:

$$\pi_\ell^{\mathrm{VR-PEP}}(\boldsymbol{\beta}_\ell|\sigma^2,\delta,\mathbf{X}_\ell^*) \;=\; \int\pi_\ell^{\mathrm{N},\mathcal{U}}(\boldsymbol{\beta}_\ell|\mathbf{y}^*,\sigma^2,\delta,\mathbf{X}_\ell^*)m_0^{\mathrm{N},\mathcal{V}R}(\mathbf{y}^*|\sigma^2,\delta,\mathbf{X}_0^*)\mathrm{d}\mathbf{y}^* \tag{2.9}$$

with $\mathrm{VR}\in\{\mathrm{CR},\mathrm{DR}\}$, $\pi_\ell^{\mathrm{N},\mathcal{U}}(\boldsymbol{\beta}_\ell|\mathbf{y}^*,\sigma^2,\delta,\mathbf{X}_\ell^*)$ defined in (2.6), $m_0^{\mathrm{N},\mathcal{D}R}(\mathbf{y}^*|\sigma^2,\delta,\mathbf{X}_0^*)$ given by (2.8) and $m_0^{\mathrm{N},\mathcal{C}R}(\mathbf{y}^*|\sigma^2,\delta,\mathbf{X}_0^*)=m_0^{\mathrm{N}}(\mathbf{y}^*|\sigma^2,\mathbf{X}_0^*)$ (see Eq. 2.7).

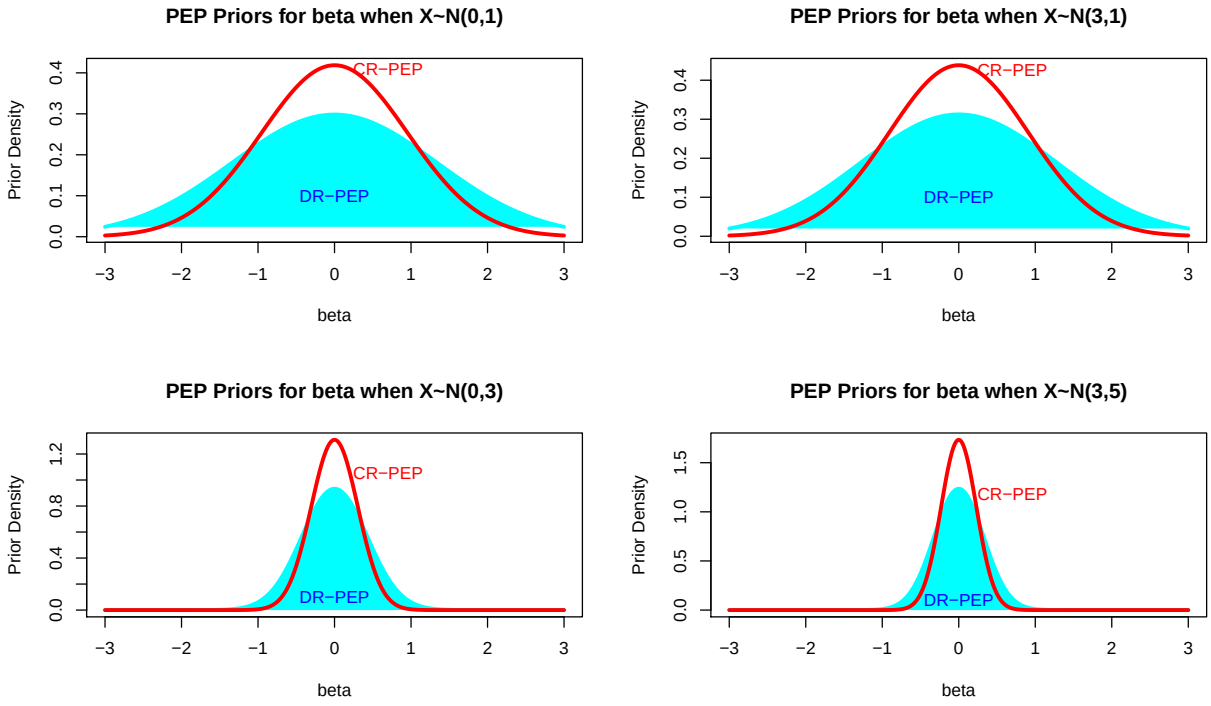

Figure 1: The marginal DR-PEP and CR-PEP priors, conditional on $\sigma=1$, using a simple normal linear regression model.

The above priors will be well defined under similar assumptions as in Pérez & Berger (2002). Furthermore, impropriety of the baseline priors does not cause indeterminacy to the resulting Bayes factors, since $\pi_\ell^{\mathrm{CR-PEP}}(\boldsymbol{\beta}_\ell|\sigma^2,\delta,\mathbf{X}_\ell^*)$ depends only on the normalizing constant of the baseline prior of the parameter of the null model and in $\pi_\ell^{\mathrm{DR-PEP}}(\boldsymbol{\beta}_\ell|\sigma^2,\delta,\mathbf{X}_\ell^*)$ the normalizing constants cancel out.

In order to visualize the two PEP priors in (2.9), we consider a simple normal linear regression case. Plots are given for the two marginal versions of PEP priors on the coefficient of the covariate X = X*. We use $\sigma = 1$, $\delta = n^* = 100$ and the values of the explanatory variable are drawn from four different normal distributions. From Figure 1 we see, in all cases, that both priors are centred at zero and the DR-PEP is more dispersed. Both priors are less dispersed as the variance of X increases.

3 Properties of PEP variants in normal regression

In this section we examine the properties of the DR-PEP and CR-PEP priors and compare them to the corresponding properties of the original PEP prior. We work within the conjugate framework considered in Fouskakis & Ntzoufras (2016); specifically, we use as baseline priors a Zellner's g-prior for $\boldsymbol{\beta}_\ell$ conditional on σ^2 and a reference prior for σ^2, that is

$$\pi_\ell^{\mathrm{N}}(\boldsymbol{\beta}_\ell|\sigma^2, \mathbf{X}_\ell^*) = f_{N_{d_\ell}}\big(\boldsymbol{\beta}_\ell; \mathbf{0}, g_0(\mathbf{X}_\ell^{*T}\mathbf{X}_\ell^*)^{-1}\sigma^2\big) \text{ and } \pi^{\mathrm{N}}(\sigma^2) \propto \sigma^{-2},$$

where d_ℓ is the dimension of $\boldsymbol{\beta}_\ell$ and $f_{N_k}\big(\cdot\,;\, \boldsymbol{\mu},\, \boldsymbol{\Sigma}\big)$ is the density function of the k–dimensional multivariate normal distribution with mean vector $\boldsymbol{\mu}$ and variance-covariance matrix $\boldsymbol{\Sigma}$. In the following we use the default values for the hyperparameters discussed in Section 2.1, namely $\delta = n$, $n^* = n$, $\mathbf{X}_\ell^* = \mathbf{X}_\ell$. In addition, following Fouskakis & Ntzoufras (2016) we set $g_0 = n^2$; this way, the overall contribution of the PEP prior to the posterior will be equal to $(1 + 1/n)$ data points, corresponding to one point contributed from the power-likelihood part and $1/n$ from the baseline g-prior. As a reference model M_0 we consider the simplest nested model under consideration.

3.1 Power-posterior component in PEP variants

Under both approaches and for any given model M_ℓ, it is straightforward to show that the unnormalized likelihood is given by

$$\begin{aligned} f_\ell\big(\mathbf{y}^* \,\big|\, \boldsymbol{\beta}_\ell, \sigma^2, \mathbf{X}_\ell\big)^{1/\delta} &= f_{N_n}\big(\mathbf{y}^*;\, \mathbf{X}_\ell\boldsymbol{\beta}_\ell,\, \sigma^2\mathbf{I}_n\big)^{1/\delta} \\ &= \delta^{\frac{n}{2}}(2\pi\sigma^2)^{\frac{n(\delta-1)}{2\delta}} f_{N_n}(\mathbf{y}^*; \mathbf{X}_\ell\boldsymbol{\beta}_\ell, \sigma^2\delta\mathbf{I}_n). \end{aligned} \tag{3.1}$$

Therefore, for both DR-PEP and CR-PEP priors, the posterior distribution (2.6), conditional on the imaginary data, is given by

$$\begin{aligned} \pi_\ell^{\mathrm{N},\mathcal{U}}\big(\boldsymbol{\beta}_\ell \,|\, \mathbf{y}^*, \sigma^2, \delta, \mathbf{X}_\ell\big) &\propto f_\ell\big(\mathbf{y}^* \,\big|\, \boldsymbol{\beta}_\ell, \sigma^2, \mathbf{X}_\ell\big)^{1/\delta} \pi_\ell^{\mathrm{N}}\big(\boldsymbol{\beta}_\ell \,|\, \sigma^2, \mathbf{X}_\ell\big), \\ &\propto f_{N_n}\big(\mathbf{y}^*;\, \mathbf{X}_\ell\boldsymbol{\beta}_\ell,\, \mathbf{I}_n\delta\sigma^2\,\big) f_{N_{d_\ell}}\Big(\boldsymbol{\beta}_\ell;\, \mathbf{0},\, g_0(\mathbf{X}_\ell^T\mathbf{X}_\ell)^{-1}\sigma^2\Big) \\ &= f_{N_{d_\ell}}\big(\boldsymbol{\beta}_\ell;\, w\widehat{\boldsymbol{\beta}}_\ell^*,\, w\delta(\mathbf{X}_\ell^T\mathbf{X}_\ell)^{-1}\sigma^2\big), \end{aligned} \tag{3.2}$$

where $w = g_0/(g_0 + \delta)$ and $\widehat{\boldsymbol{\beta}}_\ell^*$ is the maximum likelihood estimate based on the imaginary response $\mathbf{y}^*$. Thus, the posterior distribution involved in (2.9) is identical to the corre-

sponding posterior under the original conditional PEP prior; see Equation 3 in Fouskakis & Ntzoufras (2016).

3.2 Prior distributions and dispersion

In this section we examine the volume of the variance covariance matrix of the two new versions of PEP priors and we compare them with the one under the g-prior. This is important due to the connection of the volume of the variance with the dimensionality penalty induced in the Bayes factor for each pairwise model comparison; see for example Dellaportas et al. (2012). In short, the largest this volume is the highest the imposed penalty gets resulting, in its extreme form, to Lindley's paradox.

3.2.1 Diffuse-reference PEP prior

For the DR-PEP setup, the prior predictive distribution of the imaginary data under model M_ℓ is given by

$$m_\ell^{\mathrm{N},\mathcal{DR}}(\mathbf{y}^*|\sigma^2,\delta,\mathbf{X}_\ell)=\frac{m_\ell^{\mathrm{U},\mathcal{DR}}(\mathbf{y}^*|\sigma^2,\delta,\mathbf{X}_\ell)}{\int m_\ell^{\mathrm{U},\mathcal{DR}}(\mathbf{y}^*|\sigma^2,\delta,\mathbf{X}_\ell)d\mathbf{y}^*},$$

where $m_\ell^{\mathrm{U},\mathcal{DR}}(\mathbf{y}^*|\sigma^2,\delta,\mathbf{X}_\ell)$ is the normalizing constant of the power-posterior in (2.6) which is derived as follows

$$\begin{aligned} m_\ell^{\mathrm{U}}(\mathbf{y}^*|\sigma^2,\delta,\mathbf{X}_\ell) &= \int f_\ell(\mathbf{y}^*|\boldsymbol{\beta}_\ell,\sigma^2,\mathbf{X}_\ell)^{1/\delta}\pi_\ell^{\mathrm{N}}(\boldsymbol{\beta}_\ell|\sigma^2,\mathbf{X}_\ell)\mathrm{d}\boldsymbol{\beta}_\ell \\ &= \delta^{\frac{n}{2}}(2\pi\sigma^2)^{\frac{n(\delta-1)}{2\delta}}\times \\ &\quad\times\int f_{N_n}(\mathbf{y}^*;\mathbf{X}_\ell\boldsymbol{\beta}_\ell,\sigma^2\delta\mathbf{I}_n)f_{N_{d_\ell}}(\boldsymbol{\beta}_\ell;\mathbf{0},g_0(\mathbf{X}_\ell^T\mathbf{X}_\ell)^{-1}\sigma^2)\mathrm{d}\boldsymbol{\beta}_\ell \\ &= \delta^{\frac{n}{2}}(2\pi\sigma^2)^{\frac{n(\delta-1)}{2\delta}}f_{N_n}(\mathbf{y}^*;\mathbf{0},\boldsymbol{\Lambda}_\ell^{-1}\sigma^2), \end{aligned} \tag{3.3}$$

with

$$\boldsymbol{\Lambda}_\ell^{-1}=\delta\mathbf{I}_n+g_0\mathbf{X}_\ell(\mathbf{X}_\ell^T\mathbf{X}_\ell)^{-1}\mathbf{X}_\ell^T \text{ and } \boldsymbol{\Lambda}_\ell=\delta^{-1}\Big(\mathbf{I}_n-w\mathbf{X}_\ell(\mathbf{X}_\ell^T\mathbf{X}_\ell)^{-1}\mathbf{X}_\ell^T\Big).$$

From the previous equations, it immediately follows that

$$m_0^{\mathrm{N},\mathcal{DR}}(\mathbf{y}^*|\sigma^2,\delta,\mathbf{X}_0) = f_{\mathrm{N}_n}\big(\mathbf{y}^*;\,\mathbf{0},\,\boldsymbol{\Lambda}_0^{-1}\sigma^2\big) \tag{3.4}$$

with

$$\boldsymbol{\Lambda}_0^{-1}=\delta\mathbf{I}_n+\tfrac{g_0}{n}\mathbf{1}_n\mathbf{1}_n^T \text{ and } \boldsymbol{\Lambda}_0=\delta^{-1}\left(\mathbf{I}_n-\tfrac{w}{n}\mathbf{1}_n\mathbf{1}_n^T\right). \tag{3.5}$$

Thus, both components of the DR-PEP prior, that is the power-posterior in (3.2) and the prior predictive in (3.4), are exactly the same as the corresponding components

of the conditional PEP approach of Fouskakis & Ntzoufras (2016), where the density-normalized likelihood in (2.1) was used. Hence, for Gaussian linear models the DR-PEP prior coincides to the original version of the conditional PEP and is given by

$$\begin{aligned} \pi_{\ell}^{\text{DR-PEP}}\left(\boldsymbol{\beta}_{\ell} | \, \sigma^2 \, \delta, \mathbf{X}_{\ell}\right) &= f_{N_{d_{\ell}}}\left(\boldsymbol{\beta}_{\ell}; \, \mathbf{0}, \, \mathbf{V}_{\ell}\sigma^2\right), \\ \mathbf{V}_{\ell} &= \delta \left(\mathbf{X}_{\ell}^T \left[w^{-1}\mathbf{I}_n - (\delta\boldsymbol{\Lambda}_0 + w\mathbf{H}_{\ell})^{-1}\right] \mathbf{X}_{\ell}\right)^{-1} \end{aligned} \tag{3.6}$$

with $\mathbf{H}_{\ell} = \mathbf{X}_{\ell}(\mathbf{X}_{\ell}^T\mathbf{X}_{\ell})^{-1}\mathbf{X}_{\ell}^T$ and $\boldsymbol{\Lambda}_0$ given in (3.5).

The volume of dispersion of the DR-PEP prior is given by the determinant of the covariance matrix $\mathbf{V}_{\ell}$ and equals

$$\left|\mathbf{V}_{\ell}\right| = \xi \times \left|\mathbf{X}_{\ell}^T\mathbf{X}_{\ell}\right|^{-1} \text{ with } \xi = \left\{\delta w(w+1)\right\}^{d_{\ell}-d_0} g^{d_0}. \tag{3.7}$$

For the default values $\delta = n$ and $g_0 = n^2$, the variance multiplier ξ appearing in (3.7) is equal to

$$\xi = n^{2d_{\ell}} \left[\frac{2n+1}{(n+1)^2}\right]^{d_{\ell}-d_0} > n^{d_{\ell}}, \tag{3.8}$$

where on the right hand side of the inequality we have the corresponding variance multiplier of Zellner's unit-information g-prior. The result in (3.8) holds since

$$\begin{aligned} \phi(n) &= \log \xi - d_{\ell} \log n \\ &= d_{\ell} \log n + (d_{\ell} - d_0) \log \left[\frac{2n+1}{(n+1)^2}\right] \end{aligned} \tag{3.9}$$

is an increasing function of n and $\phi(2) > 0$; see Fouskakis & Ntzoufras (2016) for details. Hence, the DR-PEP prior is more dispersed than Zellner's g-prior with $g = n$, for any sample size $n \geqslant 2$, and consequently it leads to a more parsimonious variable selection procedure.

3.2.2 Concentrated-reference PEP prior

Under the CR-PEP approach, the prior predictive of the imaginary data, under the reference model, is given by

$$m_0^{\text{N},\mathcal{CR}}(\mathbf{y}^*|\sigma^2, \mathbf{X}_0) = f_{\text{N}_n}(\mathbf{y}^*; \mathbf{0}, \left[\boldsymbol{\Lambda}_0^{(\text{CR})}\right]^{-1}\sigma^2), \tag{3.10}$$

with

$$\left[\boldsymbol{\Lambda}_0^{(\text{CR})}\right]^{-1} = \mathbf{I}_n + g_0 n^{-1}\mathbf{1}_n\mathbf{1}_n^T \text{ and } \boldsymbol{\Lambda}_0^{(\text{CR})} = \mathbf{I}_n - \frac{g_0}{g_0+1} n^{-1}\mathbf{1}_n\mathbf{1}_n^T.$$

Combining (3.2) and (3.10), we obtain the CR-PEP prior which has the same form as the DR-PEP in (3.6) but with variance-covariance matrix

$$\mathbf{V}_{\ell}^{(\text{CR})} = \delta \left(\mathbf{X}_{\ell}^T \left[w^{-1}\mathbf{I}_n - (\delta\boldsymbol{\Lambda}_0^{(\text{CR})} + w\mathbf{H}_{\ell})^{-1}\right] \mathbf{X}_{\ell}\right)^{-1}.$$

As seen, the CR-PEP and DR-PEP priors differ only with respect to the variance-covariance matrix $\mathbf{V}_\ell$ appearing in (3.6) where $\mathbf{\Lambda}_0$ is substituted by $\mathbf{\Lambda}_0^{(\mathrm{CR})}$. The volume of dispersion is now given by

$$|\mathbf{V}_\ell^{(\mathrm{CR})}| = \xi \times |\mathbf{X}_\ell^T \mathbf{X}_\ell|^{-1} \text{ with } \xi = w^{d_\ell} (w+\delta)^{d_\ell - d_0} (w+\delta+wg_0)^{d_0}. \tag{3.11}$$

For the derivation of the result in (3.11) see A. Under the default setting $\delta = n$ and $g_0 = n^2$, the variance multiplier becomes

$$\xi = n^{2d_\ell} \left[\frac{n+2}{(n+1)^2}\right]^{d_\ell} \left[\frac{n^2+n+2}{n+2}\right]^{d_0}. \tag{3.12}$$

The log-ratio of the variance multipliers of the CR-PEP prior and the unit-information g-prior is given by

$$\begin{aligned} \phi(n) &= \log \xi - d_\ell \log n \\ &= d_\ell \log\left(\frac{n^2+2n}{n^2+2n+1}\right) + d_0 \log\left(\frac{n^2+n+2}{n+2}\right). \end{aligned} \tag{3.13}$$

For any model $M_\ell \supset M_0$ and under the restriction $1 \leqslant d_0 < d_\ell \leqslant n-1$ we obtain

$$\begin{aligned} d_\ell \log\left(\frac{n^2+2n}{n^2+2n+1}\right) + d_0 \log\left(\frac{n^2+n+2}{n+2}\right) &\geqslant \\ (n-1) \log\left(\frac{n^2+2n}{n^2+2n+1}\right) + \log\left(\frac{n^2+n+2}{n+2}\right) &= \phi^*(n), \end{aligned} \tag{3.14}$$

for $d_0 \in [1, n-2]$, $d_\ell \in [d_0+1, n-1]$ and any $n \in \{d_\ell+1, d_\ell+2, \ldots\}$. It can be proved that $\phi^*(n)$ is an increasing function of n and additionally $\phi^*(2) > 0$ and thus $\phi^*(n)$ is always positive. Therefore, the log-ratio of the variance multipliers in (3.13) will also be positive. Thus, the variance of the CR-PEP prior is larger than that of the g-prior, which means that CR-PEP prior will in general tend to favour less complex models. Additionally, by rewriting the variance multiplier in (3.12) as

$$\xi = n^{d_\ell} \left[\frac{n^2+2n}{n^2+2n+1}\right]^{d_\ell} \left[\frac{n^2+n+2}{n+2}\right]^{d_0}, \tag{3.15}$$

we can see that for relatively large n the first fraction in (3.15) tends to one while the second fraction tends to n. Assuming that $d_0 = 1$, the CR-PEP variance multiplier is then approximately equal to $n^{d_\ell+1}$ and the log-ratio in (3.13) will be $\phi(n) \approx \log(n^{d_\ell+1}/n^{d_\ell}) = \log(n)$. When the reference model M_0 is not the null model, the corresponding approximation is equal to $d_0 \log(n)$.

The comparison with respect to the DR-PEP prior, and consequently to the original conditional PEP approach, is more straightforward. In this case, the log-ratio of the CR-PEP variance multiplier (3.15) over the corresponding multiplier of the DR-PEP prior,

given in (3.8), is

$$\begin{aligned}\varphi(n) &= \log\left(\left[\frac{n+2}{(n+1)^2}\right]^{d_\ell}\left[\frac{n^2+n+2}{n+2}\right]^{d_0}\left[\frac{2n+1}{(n+1)^2}\right]^{d_0-d_\ell}\right)\\ &= \log\left([n+2]^{d_\ell-d_0}\,[2n+1]^{d_0-d_\ell}\left[\frac{n^2+n+2}{(n+1)^2}\right]^{d_0}\right)\\ &= \log\left(\left[\frac{n+2}{2n+1}\right]^{d_\ell-d_0}\left[\frac{n^2+n+2}{n^2+2n+1}\right]^{d_0}\right). \qquad (3.16)\end{aligned}$$

Both fractions appearing in (3.16) are equal to or smaller than one for any $n \geqslant 1$. Therefore, the log-ratio is always negative. This implies that the CR-PEP prior induces a variable selection procedure which is less parsimonious than the corresponding one under the DR-PEP prior.

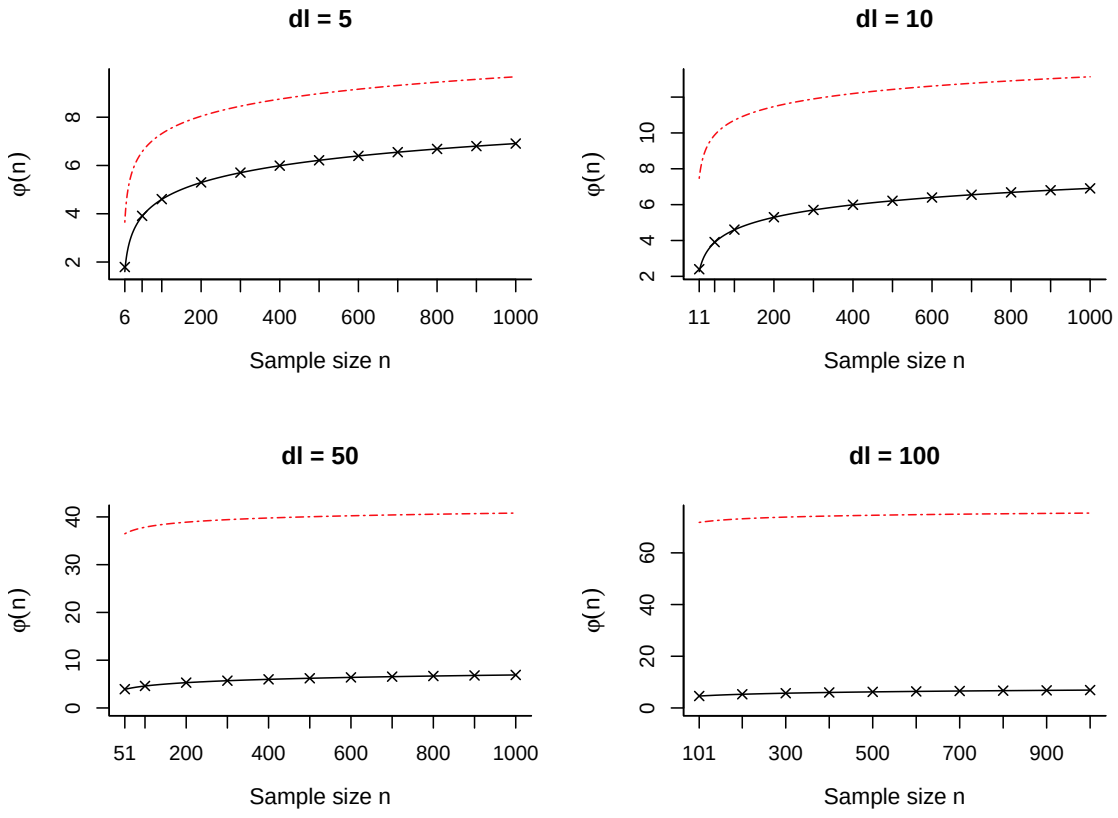

Figure 2: Log-ratios of the variance multipliers $\varphi(n)$ of the DR-PEP prior (dashed red line) and CR-PEP prior (solid black line) over the unit-information g-prior for $d_\ell = 5, 10, 50, 100$ and varying sample size; the crosses correspond to the approximation $\log(n)$.

3.2.3 Numerical illustrations

Here we provide some basic illustrations that highlight the behaviour of the variance multipliers of the CR-PEP and DR-PEP priors for varying sample size and number of predictors, given the restriction that $n \geqslant d_\ell + 1$ and assuming that $d_0 = 1$.

The log-ratios of the DR-PEP and CR-PEP prior multipliers over the unit-information g-prior multiplier (see respective Eqs. 3.9 and 3.13), for increasing sample size n and selected values of $d_\ell \in \{5, 10, 50, 100\}$, are illustrated in Figure 2. For both prior setups, the log-ratios are positive and increasing with the sample size n, with the DR-PEP being always more dispersed than the CR-PEP as expected according to Section 3.2.2. Additionally, the log-ratio of the DR-PEP prior over the g-prior increases as d_ℓ gets larger, whereas the ratio of the CR-PEP prior over the g-prior is not affected by d_ℓ as it remains constant, approximately equal to $\log(n)$.

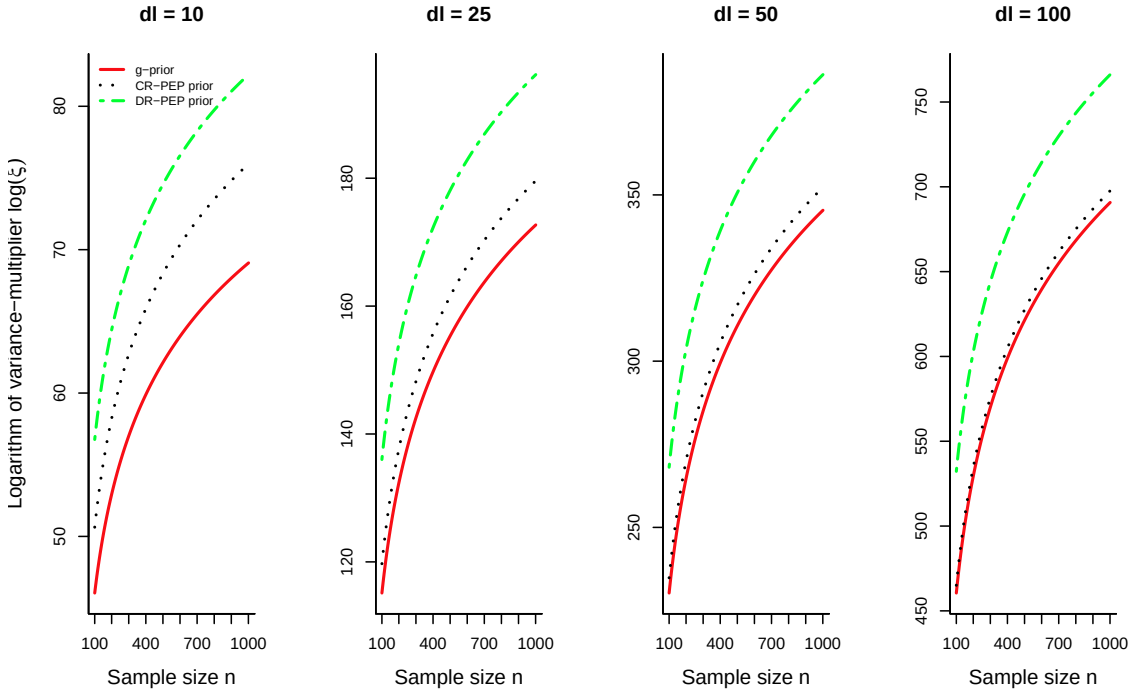

Figure 3: Log-scaled plots of the variance multipliers (ξ) of the DR-PEP, CR-PEP and Zellner's g priors for $d_\ell = 10, 25, 50, 100$ and sample size from 100 to 1000.

In Figure 3 we present on log-scale the variance multipliers of the CR-PEP, DR-PEP and the unit-information g priors for $d_\ell \in \{10, 25, 50, 100\}$ and sample size ranging from 101 (the minimum size required for $d_\ell = 100$) to 1000. As seen, as model dimensionality increases all priors become more dispersed; however, the distance between the variance multiplier of the DR-PEP prior and the corresponding multipliers of the CR-PEP and the g-prior is also increasing with d_ℓ. Potentially, this feature of the DR-PEP prior makes it more suitable for problems involving a large number of predictors and where the aim is to have a parsimonious model selection method.

3.3 Marginal likelihood and model selection consistency

In this Section we examine the limiting behaviour of the marginal likelihood of any model M_ℓ, under the DR-PEP and the CR-PEP priors, under the M-closed scenario. We have assumed that d_ℓ does not increases with the sample size n and additionally $d_\ell < n$.

The posterior distribution of $\boldsymbol{\beta}_\ell$ and σ^2 under either the conditional PEP prior of Fouskakis & Ntzoufras (2016) or the DR-PEP prior, examined here, is given by

$$\pi_\ell^{\text{DR-PEP}}(\boldsymbol{\beta}_\ell, \sigma^2|\mathbf{y}, \delta, \mathbf{X}_\ell) = f_{N_{d_\ell}}\left(\boldsymbol{\beta}_\ell; \widetilde{\boldsymbol{\beta}}_\ell, \widetilde{\boldsymbol{\Sigma}}_\ell\, \sigma^2\right) f_{IG}\left(\sigma^2; \widetilde{a}_\ell, \widetilde{b}_\ell\right), \tag{3.17}$$

where $\widetilde{\boldsymbol{\beta}}_\ell = \widetilde{\boldsymbol{\Sigma}}_\ell \mathbf{X}_\ell^T \mathbf{y}$, $\widetilde{\boldsymbol{\Sigma}}_\ell = \left(\mathbf{V}_\ell^{-1} + \mathbf{X}_\ell^T \mathbf{X}_\ell\right)^{-1}$, $\widetilde{a}_\ell = n/2$, $\widetilde{b}_\ell = \text{SS}_\ell/2$ with $\text{SS}_\ell = \mathbf{y}^T\left(\mathbf{I}_n + \mathbf{X}_\ell \mathbf{V}_\ell \mathbf{X}_\ell^T\right)^{-1}\mathbf{y}$, and $f_{IG}(\cdot\,;\, a, b)$ denotes the density of the inverse gamma distribution with shape parameter a and scale parameter b. In the above, $\mathbf{V}_\ell$ is given in Section 3.2.1.

Then, the marginal log-likelihood is given by

$$\begin{aligned} \log m_\ell^{\text{DR-PEP}}(\mathbf{y}|\delta, \mathbf{X}_\ell) &= C - \frac{1}{2}\log\left|\mathbf{I}_n + \mathbf{X}_\ell \mathbf{V}_\ell \mathbf{X}_\ell^T\right| - \\ &\quad \frac{n}{2}\log\left(\mathbf{y}^T(\mathbf{I}_n + \mathbf{X}_\ell \mathbf{V}_\ell \mathbf{X}_\ell^T)^{-1}\mathbf{y}\right), \end{aligned} \tag{3.18}$$

where C is a constant that does not depend on the structure of model M_ℓ. For large n, the marginal log-likelihood in (3.18) can be approximated by

$$\log m_\ell^{\text{DR-PEP}}(\mathbf{y}|\delta, \mathbf{X}_\ell) \approx C - \frac{1}{2}\text{BIC}_\ell. \tag{3.19}$$

Thus, the marginal likelihood under the DR-PEP prior has the same limiting behaviour as the BIC which is known to be consistent under a minor realistic assumption (Fernández et al. 2001, Liang et al. 2008). For a detailed proof of (3.19) see Fouskakis & Ntzoufras (2016).

Similarly to (3.18), the marginal log-likelihood under the CR-PEP prior is

$$\begin{aligned} \log m_\ell^{\text{CR-PEP}}(\mathbf{y}|\delta, \mathbf{X}_\ell) &= C - \frac{1}{2}\log\left|\mathbf{I}_n + \mathbf{X}_\ell \mathbf{V}_\ell^{(\text{CR})} \mathbf{X}_\ell^T\right| - \\ &\quad \frac{n}{2}\log\left(\mathbf{y}^T(\mathbf{I}_n + \mathbf{X}_\ell \mathbf{V}_\ell^{(\text{CR})} \mathbf{X}_\ell^T)^{-1}\mathbf{y}\right). \end{aligned} \tag{3.20}$$

Following the proof Fouskakis & Ntzoufras (2016, see Section D, Eqs. D.1–D.2 of the Supplementary Material), the first logarithmic term yields

$$\begin{aligned} |\mathbf{I}_n + \mathbf{X}_\ell \mathbf{V}_\ell^{(\text{CR})} \mathbf{X}_\ell^T| &= (1+\delta w)^{d_\ell} |\boldsymbol{\Lambda}_0^{(\text{CR})}|^{-1} \left|\boldsymbol{\Lambda}_0^{(\text{CR})} + \left(\frac{w^2}{1+\delta w}\right)\mathbf{H}_\ell\right| \\ &\approx (1+\delta)^{d_\ell} |\boldsymbol{\Lambda}_0^{(\text{CR})}|^{-1} \left|\boldsymbol{\Lambda}_0^{(\text{CR})} + \left(\frac{1}{1+\delta}\right)\mathbf{H}_\ell\right| \\ &\approx (1+\delta)^{d_\ell}. \end{aligned} \tag{3.21}$$

Note that the approximation is accurate for large n when $\delta = n$ and $g_0 = n^2$, so that $w = g_0/(g_0 + \delta) \approx 1$. Given these values, we can also approximate the second logarithmic term in (3.20) by

$$\mathbf{y}^T(\mathbf{I}_n + \mathbf{X}_\ell \mathbf{V}_\ell^{(\mathrm{CR})} \mathbf{X}_\ell^T)^{-1}\mathbf{y} \approx \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}_\ell \left(\mathbf{X}_\ell^T\mathbf{X}_\ell\right)^{-1} \mathbf{X}_\ell^T\mathbf{y} \equiv \mathrm{RSS}_\ell, \tag{3.22}$$

where RSS_ℓ is the usual residual sum of squared of model M_ℓ. The derivation for (3.22) is provided in B. Hence, the marginal log-likelihood under the CR-PEP prior is approximately given by

$$\begin{aligned} \log m_\ell^{\mathrm{CR-PEP}}(\mathbf{y}|\delta, \mathbf{X}_\ell) &\approx C - \frac{d_\ell}{2}\log(n+1) - \frac{n}{2}\log(\mathrm{RSS}_\ell) \\ &\approx C - \frac{d_\ell}{2}\log(n) - \frac{n}{2}\log(\mathrm{RSS}_\ell) \\ &\approx C - \frac{1}{2}\mathrm{BIC}_\ell, \end{aligned} \tag{3.23}$$

for $\delta = n$ and large n. Thus, variable selection, based on the CR-PEP prior with a g-prior as baseline, has also the same limiting behaviour as the BIC and is, therefore, consistent.

Following a comment of a referee, we have calculated the standardized inner product between the likelihood and the prior which provides a geometric perspective of the marginal likelihood. This quantity was defined by de Carvalho et al. (2019) and is given by

$$\kappa_{\pi_\ell, f_\ell} = \frac{m_\ell(\mathbf{y})}{||f_\ell(\mathbf{y}|\boldsymbol{\beta}_\ell, \sigma^2, \mathbf{X}_\ell)||\ ||\pi_\ell(\boldsymbol{\beta}_\ell, \sigma^2)||},$$

where $||g(\boldsymbol{\theta})|| = \sqrt{\int g(\boldsymbol{\theta})^2 d\boldsymbol{\theta}}$ for any function or distribution of the parameters under consideration $\boldsymbol{\theta}$ (here we consider the conflict/comparison between the prior and the likelihood). This measure provides a normalized version of the marginal likelihood and measures the agreement between the prior and the data likelihood. When $\kappa_{\pi_\ell, f_\ell}$ equals to one, the prior is in total agreement with the likelihood, while when $\kappa_{\pi_\ell, f_\ell} \to 0$ there is a large conflict between the prior and the likelihood. In terms of model selection, the posterior Bayes factor of Aitkin (1991) essentially leads to $\kappa_{\pi_\ell, f_\ell} = 1$ and generally it is not a desirable or realistic case in practice. On the other hand, within the objective Bayes framework, we wish to have a small $\kappa_{\pi_\ell, f_\ell}$ but well separated from zero since values very close to zero will indicate the activation of the Bartlett's (1957) paradox. In our case we use an improper prior (Jeffrey's prior) for the error variance σ^2, therefore, the above quantity cannot be calculated due to the unknown normalizing constant.

4 Simulation studies

In this Section we present comparisons based on simulations in order to study the performance of the proposed methods under various scenarios. In Section 4.1 we focus on one case of correct model specification and three cases of misspecified models, considering independent as well as correlated predictors and also increasing sample size. In Section

4.2 we investigate the setting where the model is correctly specified but the number of covariates grows with sample size, assuming again that covariates are either independent or correlated.

We compare DR-PEP and CR-PEP to "well-established" priors routinely used for Bayesian variable selection; namely, the g-prior (Zellner 1976), the hyper-g prior (Liang et al. 2008) and the hyper-g/n prior (Liang et al. 2008). All competing methods were implemented using the `BAS` package in `R`; we set $g = n$ in the g-prior in order to implement the unit-information prior (Kass & Wasserman 1995) and $\alpha = 3$ in the hyper-g and hyper-g/n priors as recommended by Liang et al. (2008). A beta-binomial prior on the model space, with both parameters equal to 1, is used. Each simulation described below is repeated 100 times. `R` reproducible code for all simulations is available at `https://github.com/kperrakis/PEP_variations`.

4.1 Simulations under M-open and M-closed scenarios

In these simulations we assume that dimensionality is $p = 10$ with three out of the ten covariates $(X_1, \cdots, X_{10})$ actually relating to the response; specifically, the assumed true relationship between the response variable and the influential predictors is

$$Y_i = 0.3X_{i1} + 0.5X_{i3} + X_{i4} + \varepsilon_i, \tag{4.1}$$

for $i = 1 \cdots, n$ and $n \in \{30, 50, 100, 500, 1000\}$. Similarly to the setup in Rossell & Rubio (2018) we consider the following four cases for the distribution of the random errors ε_i: (a) normal, (b) Laplace, (c) asymmetric normal, and (d) heteroscedastic normal. Case (a) is an M-closed case as the true model is included in the set of models under consideration; here the errors are generated as $\varepsilon_i \sim N(0, 1)$. In contrary, cases (b) - (d) are all M-open scenarios. Specifically, in case (b) we have heavier tails than in case (a) with errors $\varepsilon_i \sim L(0, 1/\sqrt{2})$, where $L(\mu, b)$ denotes the Laplace distribution with location μ and scale b; the latter parameter is set so that the variance equals one. In case (c) we utilize the asymmetric (or two-piece) normal distribution under the epsilon-skew parameterization (Mudholkar & Hutson 2000), denoted as $AN(\mu, \vartheta, \alpha)$, with location parameter $\mu \in \mathbb{R}$, scale parameter $\vartheta \in \mathbb{R}^+$ and asymmetry parameter $\alpha \in [-1, 1]$. Its probability density function is

$$f(x) = \frac{1}{\sqrt{2\pi\vartheta}}\left[\exp\left(-\frac{(x-\mu)^2}{\vartheta(1-\alpha)^2}\right)I(x < \mu) + \exp\left(-\frac{(x-\mu)^2}{\vartheta(1+\alpha)^2}\right)I(x \geqslant \mu)\right], \tag{4.2}$$

where $I(\cdot)$ is the indicator function and $x \in \mathbb{R}$. Here, we assume that $\varepsilon_i \sim AN(0, 1, 0.5)$ which gives approximately $\mathrm{Var}(\varepsilon_i) \approx 1.12$. Finally, for the heteroscedastic case (d) we simulate initially $\tilde{\varepsilon}_i \sim N(0, 1)$ and then set $\varepsilon_i = \exp(0.3X_{i1} + 0.5X_{i3} + X_{i4})\tilde{\varepsilon}_i/c$, where c is tuned in such a way so that $\mathrm{Var}(\varepsilon_i) = \mathrm{Var}(\tilde{\varepsilon}_i)$ in order to have comparable signal-to-noise ratio with the previous cases (Rossell & Rubio 2018).

4.1.1 Scenario 1: Independent covariates

In this first scenario the covariates are generated independently from a Gaussian distribution. Figure 4 depicts the between-samples distribution of the posterior probability of the true model for the Bayesian variable selection techniques under comparison and the four cases under consideration. Focusing initially on the M-closed case (a), it is clear that for small sample sizes all methods under consideration fail to provide high posterior evidence in favor of the true model. As the sample size gets larger, all methods increase their posterior support towards the true model, with DR-PEP to perform slightly better than CR-PEP and the Zellner's g-prior, followed by the hyper-g/n and the hyper-g priors. This is sensible since the two proposed methods together with the Zellner's g-prior, are converging to the same Bayes factors as n grows but with the proposed approaches constantly supporting more parsimonious models. On the other hand, the hyper-g and the hyper-g/n priors give the lowest support towards the true model due to their hierarchical structure which increases the posterior uncertainty on the model space. These two methods need larger sample size, than the rest of the approaches, in order to fully a-posteriori support the true generating mechanism. Interestingly, results are almost identical for the three M-open cases (b), (c) and (d); this provides some empirical evidence that the methods lead to robust model selection under the particular cases of model misspecification and when covariates are independent.

We now turn to inspection of posterior inclusion probabilities starting with case (a); these are presented in Figure 5. First, we observe that all methods successfully identify X_4 (with true effect equal to one) as an important component of the model, even for small sample sizes. Furthermore, the between-samples variability of the posterior inclusion probabilities reduces as the sample size increases. Similar is the picture for the posterior inclusion probabilities of the other two covariates with non-zero effects, X_1 and X_3, but with slower rates of convergence towards one. For covariate X_1 (with true effect equal to 0.3) we observe large between-samples uncertainty concerning the importance of this effect even with $n = 100$ under all methods. For $n \geqslant 500$, all methods successfully identify the importance of this covariate with almost zero between-samples variability. In general, the hyper-g method supports this covariate with slightly higher inclusion probabilities, compared to the other methods; with DR-PEP we observe slightly lower inclusion probabilities, compared to the other approaches, for small n. This is due to the characteristics of the two methods, with the first supporting more complicated models and the latter more parsimonious ones. Finally for covariate X_3 (with true effect equal to 0.5), all methods successfully identify it as important with almost zero between-samples variability for $n \geqslant 100$. For smaller sample sizes the methods behave similarly, showing large between-samples uncertainty concerning the importance of this effect and almost identical median values.

The between-samples distribution of the posterior inclusion probabilities for all covariates with zero true effects is similar; see Figure 5. It is noticeable that all methods identify, fairly fast, that these covariates should have low posterior inclusion probabilities with the between-samples variability decreasing as n gets larger. DR-PEP prior shows

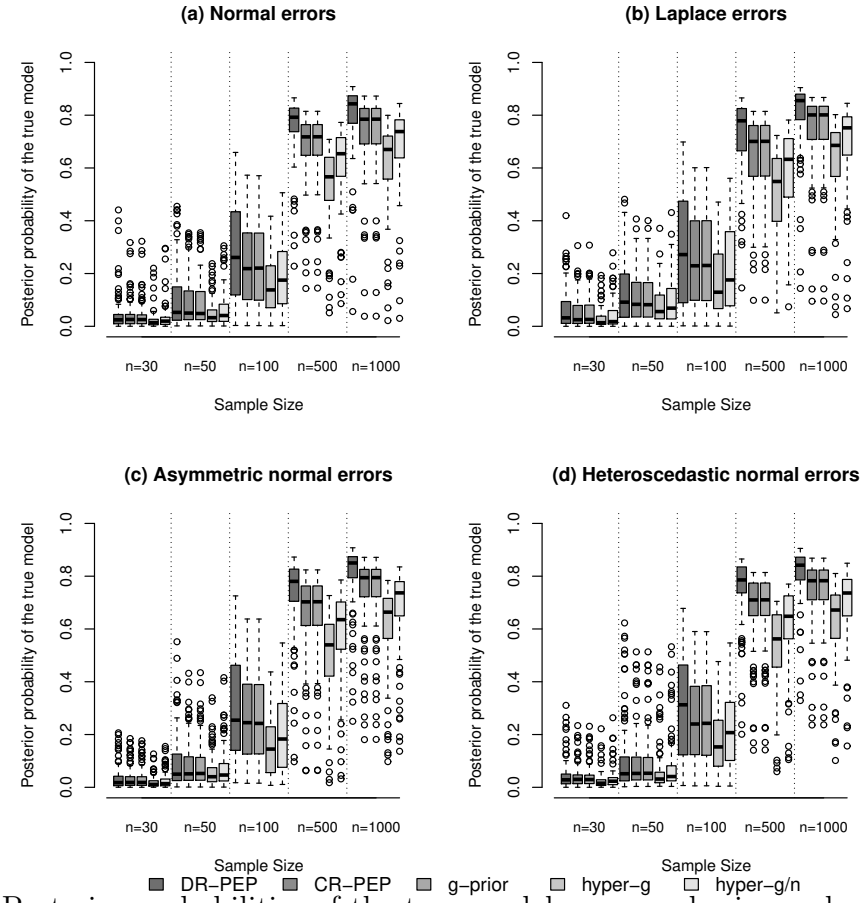

Figure 4: Posterior probabilities of the true model vs. sample size under the four model cases for Scenario 1 (independent covariates).

the best behaviour, followed by the CR-PEP and the g-prior that behave in a similar manner and then by the hyper-g/n and hyper-g priors. In general the posterior inclusion probabilities under the hyper-g prior are on average higher under small sample sizes (close to 0.4, for $n = 30$ for example). This increases the posterior uncertainty on the model space and results to lower probabilities of identifying the true model as the maximum a-posteriori model. Furthermore, it is also noticeable that the inclusion probabilities under the hyper-g and hyper-g/n priors seem to converge slower towards zero as n gets larger, both in terms of median values and in terms of between-samples variability. To sum-up, all methods identify the true model structure with increasing probability as n increases. All methods tend to select simpler models than the true model structure when they fail while the covariates identified falsely as important are relatively low even for small samples for the PEP and the g priors but considerably higher for the hyper-g priors (especially for small samples).

The corresponding results for the M-open cases (b), (c) and (d) can be found in C; see Figures C.1, C.2 and C.3, respectively. Overall, we cannot detect significant differences with respect to the results obtained under case (a); a finding which is generally in line with Figure 4.

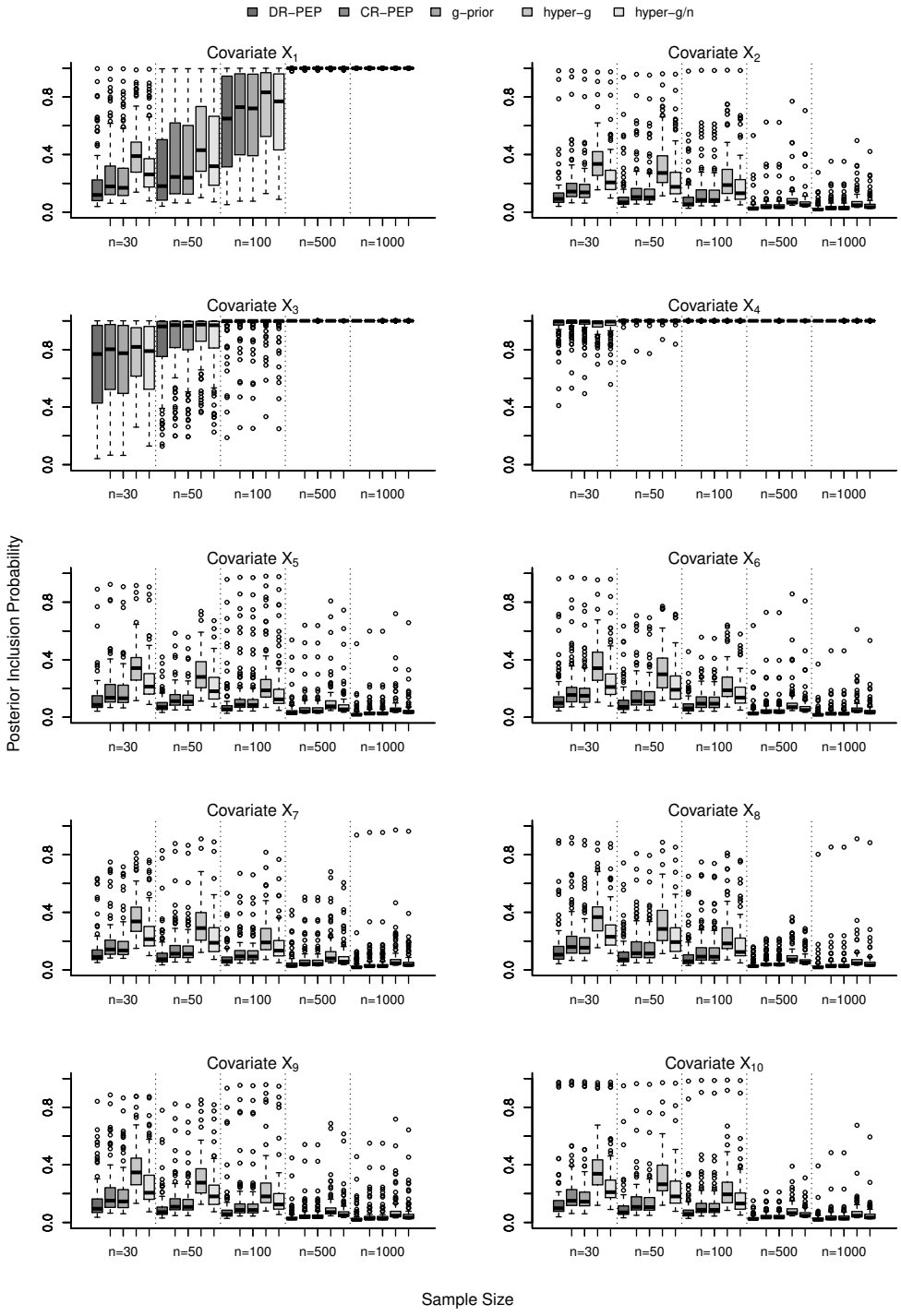

Figure 5: Posterior inclusion probabilities of each covariate for Scenario 1 (independent covariates) under normal errors (M-closed case).

To sum up, the main finding from this simulation is that the DR-PEP and CR-PEP

methods identify the true model structure with (slightly) higher posterior probability than the rest of the methods, with the DR-PEP prior to behave slightly better. They both provide posterior inclusion probabilities close to zero for non-important effects (even for small sample sizes) and high inclusion probabilities for the important effects (although these are slightly smaller than the ones obtained under the hyper-g and hyper-g/n priors for small sample sizes).

4.1.2 Scenario 2: Highly-correlated covariates

In this second simulation the covariates are drawn from a multivariate normal and have again marginally a zero mean and a unit variance, but this time we introduce strong correlations amongst them; namely, all correlations are set equal to 0.9. It is well known that this causes problems in variable selection methods, especially under small samples; see for example Ghosh & Ghattas (2015).

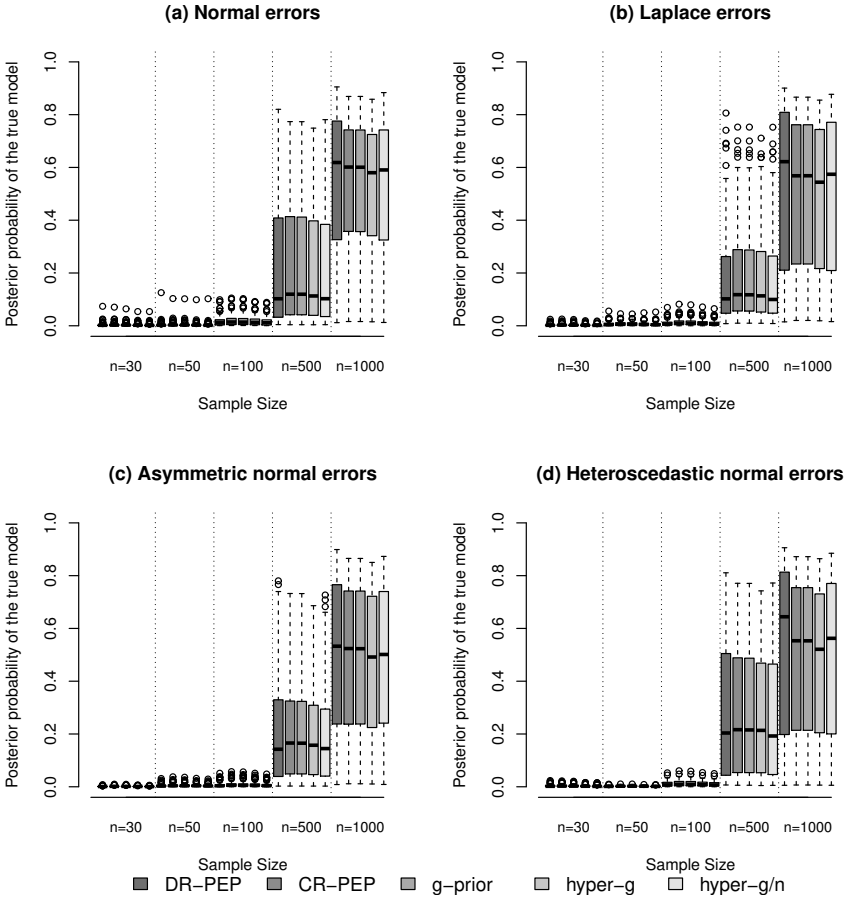

Figure 6: Posterior probabilities of the true model vs. sample size under the four model cases for Scenario 2 (highly-correlated covariates).

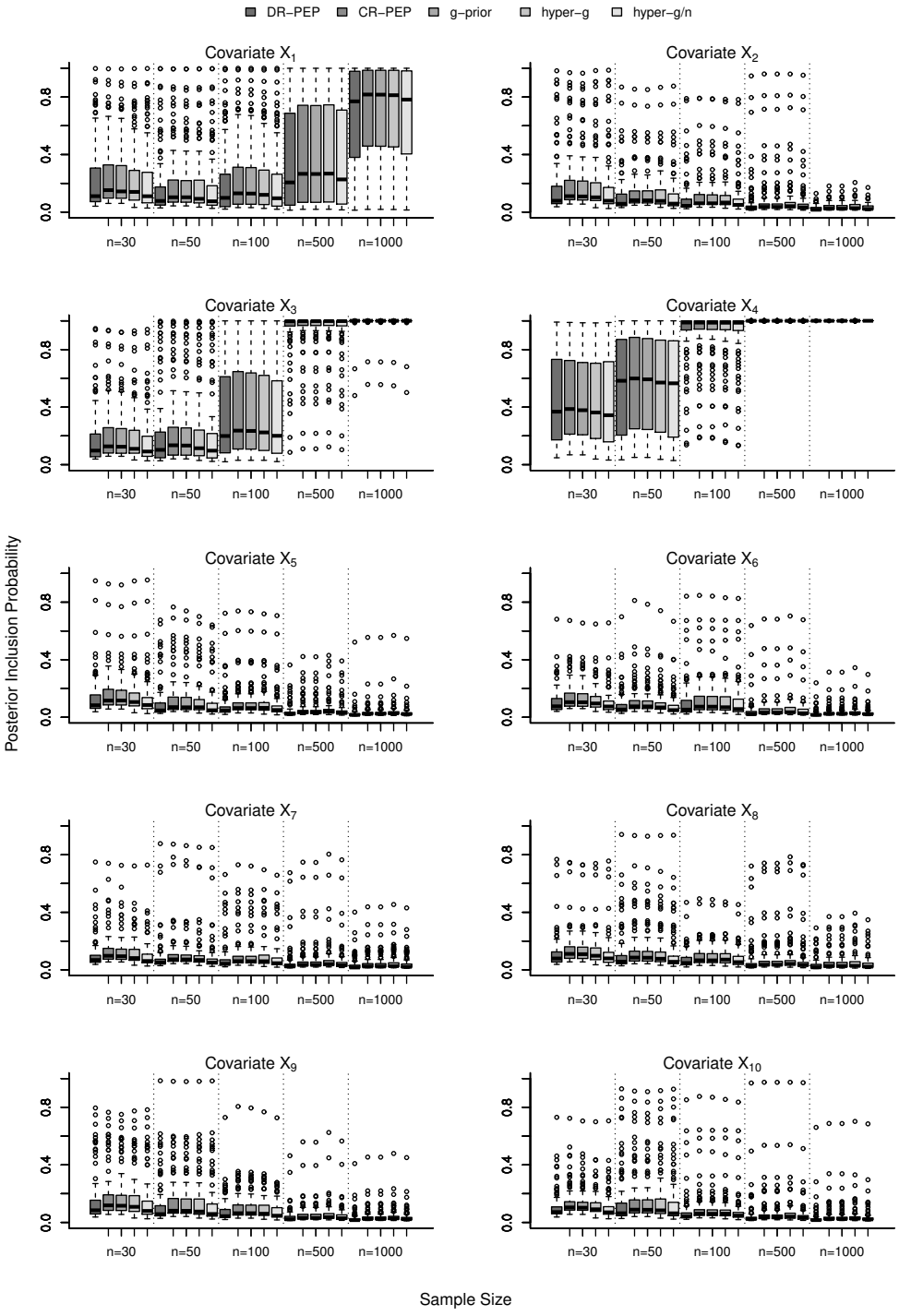

Figure 7: Posterior inclusion probabilities of each covariate for Scenario 2 (highly-correlated covariates) under normal errors (M-closed case).

Again, we start our analysis with the resulting posterior probabilities of the true model, initially focusing on case (a) in Figure 6. In general, all methods perform similarly

in this scenario; as expected a large sample size is needed for the identification of the true model due to the high correlations. Specifically, we see that all methods fail to provide high posterior evidence in favor of the true model even when $n = 500$. For $n = 1000$, posterior model probabilities are on average above 0.5 and all methods seem to perform equally good; however we still have high between-samples variability in the estimates. The situation is again similar under the M-open cases (b), (c) and (d) in Figure 6; overall there are no significant different with reference to case (a). We observe only some slight differences; specifically, in cases (b) and (c) we have a small drop in the posterior probabilities under $n = 500$ and, surprisingly, probabilities are a bit higher in case (d) under $n = 500$ and $n = 1000$.

The posterior inclusion probabilities under case (a) are presented in Figure 7. All methods successfully identify the importance of covariate X_4 for $n = 100$, with between-samples variability diminishing for larger n. In comparison, for covariate X_3 all methods require a larger sample size ($n = 500$ in our experiments) in order to reach high posterior inclusion probabilities (close to one), while for X_1 (which has the smallest effect) convergence to one is much slower as we reach median inclusion probabilities above 0.5 only when $n = 1000$, and even in this case between-samples variability is large. Concerning the non-influential covariates which have a zero effect to the response, their estimated posterior inclusion probabilities are quite similar across competing methods. Specifically, all methods identify fairly fast, i.e. even under small sample sizes, that these covariates should have low posterior inclusion probabilities. To sum up, all methods identify the true model with increasing posterior probability as n grows but, in this scenario, larger sample size is required to identify the true model in comparison with the independent covariates case. Moreover, here, zero coefficients are identified by all methods even for small samples. Hence, the behavior of selecting more parsimonious models is more intense in this case than in the scenario of Section 4.1.1.

The corresponding plots with the posterior inclusion probabilities under the three M-open cases can be found in D; specifically, see Figure D.1 for case (b), Figure D.2 for case (c) and Figure D.3 for case (d). Again, results are more or less the same to those obtained in the M-closed case. Correlations seem to affect more the heteroscedastic case (d), where we observe higher variance than usual in certain estimates; see Figure D.3.

4.2 Simulations with growing p and n

In this Section we focus on the M-closed case, allowing this time model dimension to increase with sample size. In particular, we assume the same true linear effects as in Eq. (4.1), with all covariates either independent (as in Section 4.1.1) or highly correlated (as in Section 4.1.2), for $n \in \{30, 50, 100, 500, 750\}$ and $p = \lceil n^{0.4} \rceil \in \{4, 5, 7, 13, 15\}$; where $\lceil x \rceil$ denotes the least integer greater than or equal to the real number x.

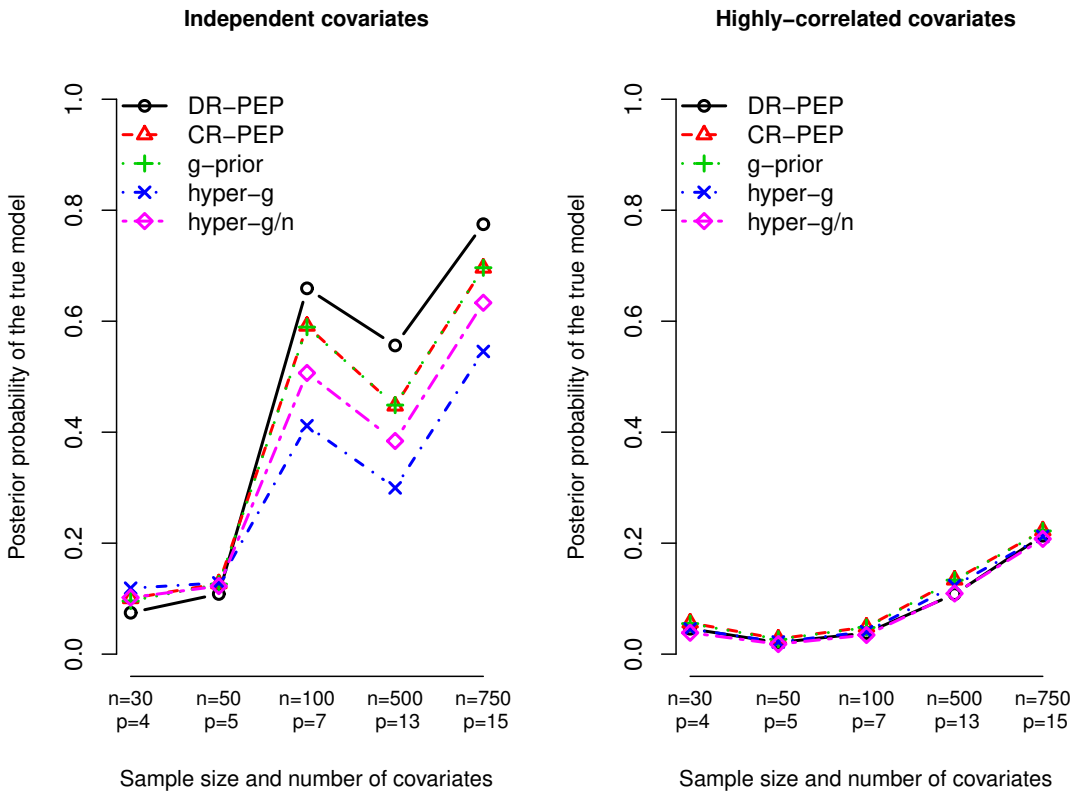

Figure 8: Posterior probabilities of the true model vs. sample size and model dimensionality under independent covariates (left) and highly-correlated covariates (right).

The posterior model probabilities of the true model are presented in Figure 8. Under independent covariates we see that with the DR-PEP prior we have a faster convergence to one, while the CR-PEP and g-prior (which exhibit identical behaviour) follow as second best. The differences in posterior model probabilities showcase the PEP properties, examined in this paper, in comparison to properties of the hyper-g prior. Under highly-correlated predictors we have an entirely different picture; here, as expected, convergence to one is very slow, irrespective of the prior. All methods behave similarly, with the CR-PEP and g-prior having just a very slight edge. In general, Figure 8 highlights the impact of high correlations on variable selection.

The resulting posterior inclusion probabilities (averages) from the independence and high-correlation scenarios are shown in Figures E.1 and E.2, respectively, in E. In the former case we generally observe similar patterns regarding the influential predictors, except for X_1 under $(n = 30, p = 4)$ and $(n = 50, p = 5)$; in this specific cases the hyper-g has a higher detection rate for this predictor. We also notice that the DR-PEP is more effective in identifying the covariates with a zero effect. In the second case (high-correlations), all methods are effective overall in detecting the importance of X_4, in the sense that this predictor would be included in the median probability model in all (n, p)

settings under consideration. This is also the case for covariate X_3, but to a lesser extend; i.e. not when $(n = 50, p = 5)$ and $(n = 100, p = 7)$. Here the problematic case is that of predictor X_1 which has the smallest non-zero effect; as seen this covariate never enters the median probability model regardless of method. The failure in detecting X_1 also explains the very low posterior model probabilities (of the true model) witnessed in Figure 8.

5 Discussion

In this paper we examined the properties of two new versions of the PEP prior, which have been recently proposed in the context of objective Bayesian variable selection (Fouskakis et al. 2018), namely the CR-PEP and DR-PEP priors. Specifically, we compared the dispersion of these priors and investigated the aspect of model selection consistency under each prior in the normal linear regression model. Under the M-closed scenario, consistency is proved theoretically, while under the M-open case, consistency is investigated via simulation, using three different model misspecification scenarios.

The main findings can be summarized as follows. In the Gaussian case, the DR-PEP prior coincides with the original conditional PEP prior of Fouskakis & Ntzoufras (2016), thus, sharing the same M-closed consistency and parsimony properties. On the other hand, the predictive distribution of the imaginary data used in the CR-PEP set-up, results in a PEP prior form which is less dispersed and, therefore, also less parsimonious than the DR-PEP prior. Nevertheless, the CR-PEP prior also leads to an M-closed consistent variable selection procedure. In addition, both priors have larger variances than the unit-information g-prior, which implies that they will support more parsimonious models than the g-prior. The DR-PEP prior in particular seems to be more suitable for large-p problems, as its variance ratio over the g-prior increases as the number of predictor variables becomes larger.

Generally, the resulting PEP priors can be viewed as extensions of the g-prior by considering imaginary, instead of fixed, data coming from a "suitable" predictive distribution. Specifically, they have an extra hierarchical level to the specification of the prior distribution that has an effect on both the prior mean and the prior variance, through the variability of the imaginary data. For more details see Fouskakis & Ntzoufras (2016).

Concerning the desiderata formulated in the important work of Bayarri et al. (2012), the two new versions of PEP prior satisfy the basic criterion $(C1)$ of propriety, the criterion of model consistency $(C2)$, under the M-closed normal linear regression setup (see Section 3.3), and the predictive matching criterion $(C5)$, under the general GLM setup (Fouskakis & Ntzoufras 2016). On the other hand both versions fail on the information consistency criterion $(C4)$; see Fouskakis & Ntzoufras (2016). The latter issue can be resolved through the hyper-δ PEP prior extension introduced in Fouskakis et al. (2018). Therefore, an additional interesting direction of future research is to examine the properties of those hyper-δ versions of PEP priors, under the normal linear regression setup.

Regarding the simulation experiments with the three M-open scenarios of model misspecification that we considered, the empirical findings suggest that PEP priors seem to

be quite robust as the results (posterior inclusion probabilities, posterior probability of true model) we obtained were very similar to those under the M-closed case. Notably, that was also the case for the competing methods (g, hyper-g and hyper-g/n prior). Studying analytically the theoretical properties of PEP priors under model misspecification is a further interesting future research direction.

Under the M-open case, using a proper inverse-gamma prior on σ^2, instead of the (improper) reference prior, would potentially be more effective in controlling the tails of the posterior of σ^2 in cases where the true generating mechanism has heavier tails and thus inflates the variance. In this work we preferred the standard (objective) choice of the reference prior because (a) the main aim was not estimation under model misspecification and (b) we focused on comparisons with other standard objective approaches (the g-prior and its extensions) that also use the reference prior. Generally, the reference prior is considered as a standard choice in the objective Bayesian variable selection literature (see for example in Liang et al. (2008)) since the error variance is treated as a common parameter under estimation.

Extension of the PEP prior methodology depends upon the specification of the power-likelihood. For Gaussian models, the power-likelihood can be written in a straightforward manner due to the property of the normal distribution which results again to a normal distribution with variance inflated by a factor of δ when it is raised to the power of δ. Hence, the PEP prior can be specified for any model that has in its core formulation the normal distribution; for example, Gaussian mixture models. For models with other sampling distributions this property is lost, as the corresponding power-likelihood can be of non-standard form. The PEP variations discussed here, initially introduced in Fouskakis et al. (2018) can be applied to generalized linear models and possibly models that have more complicated structures. Efficient computation in large spaces still remains an issue for future research for such cases. A possible solution can be found by approximating PEP priors by simpler normal distributions as the ones proposed by Rodriguez (2015). Of course, the BIC-consistency of the PEP-based marginal likelihood, should need to be explored in model-specific context. Nevertheless, the general result of Schwarz (1978) is still valid. Moreover, the unit-information interpretation of the PEP priors makes us intuitively believe that the results of Kass & Wasserman (1995) can be extended also for PEP priors.

Implementation of the PEP methodology in high dimensional problems can be achieved using standard model search tools, because the marginal likelihood under the proposed priors is readily available. Therefore, when full enumeration is computationally infeasible, model space exploration may be carried out by using the MC^3 algorithm (Madigan & York 1995), the strategies proposed in Johnson & Rossell (2012), the adaptive sampling scheme of Clyde et al. (2011) or any other adaptive model search approach; see for example in Ji & Schmidler (2013).

Finally, it is worth nothing that the simulation-based findings presented here do not necessarily provide evidence for cases of high–dimensional problems. According to a referee comment, robustness of posterior inclusion probabilities is expected in low dimensions, since the covariate effect detection rate depends on model specification as well as

model dimensionality (Rossell & Rubio 2018). In this work we considered misspecification of the distribution of the residual errors. Misspecification of the functional form of the covariates is a further interesting future research direction.

Acknowledgements/Funding

We wish to thank the Associate Editor and two referees for comments that greatly strengthened the paper.

This research was funded by the Research Centre of the Athens University of Economics and Business (Funding program Action 2 for the support of basic research).

Appendix

A Derivation of Equation 3.11

The determinant of the CR-PEP prior covariance matrix is

$$\left|\mathbf{V}_\ell^{(\mathrm{CR})}\right| = \delta^{d_\ell}\left|w^{-1}\mathbf{X}_\ell^T\mathbf{X}_\ell - \mathbf{X}_\ell^T(\delta\mathbf{\Lambda}_0^{(\mathrm{CR})} + w\mathbf{H}_\ell)^{-1}\mathbf{X}_\ell\right|^{-1}. \tag{A.1}$$

Based on the matrix determinant Lemma (Harville 1997, p.416), which states that $|\mathbf{A} + \mathbf{CBD}^T| = |\mathbf{A}||\mathbf{B}||\mathbf{B}^{-1} + \mathbf{D}^T\mathbf{A}^{-1}\mathbf{C}|$ for any square invertible matrices $\mathbf{A}$ and $\mathbf{B}$, we can write (A.1) as

$$\begin{aligned}
\left|\mathbf{V}_\ell^{(\mathrm{CR})}\right| &= \delta^{d_\ell}\Big(|w^{-1}\mathbf{X}_\ell^T\mathbf{X}_\ell||-(\delta\mathbf{\Lambda}_0^{(\mathrm{CR})} + w\mathbf{H}_\ell)|^{-1} \times \\
&\quad \times|-(\delta\mathbf{\Lambda}_0^{(\mathrm{CR})} + w\mathbf{H}_\ell) + w\mathbf{X}_\ell(\mathbf{X}_\ell^T\mathbf{X}_\ell)\mathbf{X}_\ell^T|\Big)^{-1} \\
&= \delta^{d_\ell}w^{d_\ell}|\mathbf{X}_\ell^T\mathbf{X}_\ell|^{-1}|\delta\mathbf{\Lambda}_0^{(\mathrm{CR})}|^{-1}|(\delta\mathbf{\Lambda}_0^{(\mathrm{CR})} + w\mathbf{H}_\ell)|.
\end{aligned} \tag{A.2}$$

Using repeatedly the matrix determinant Lemma on the last term of (A.2) yields

$$\begin{aligned}
|\delta\mathbf{\Lambda}_0^{(\mathrm{CR})} + w\mathbf{H}_\ell| &= |\delta\mathbf{\Lambda}_0^{(\mathrm{CR})}||\mathbf{X}_\ell^T\mathbf{X}_\ell|^{-1}\left|\mathbf{X}_\ell^T\mathbf{X}_\ell + \frac{w}{\delta}\mathbf{X}_\ell^T\left[\mathbf{\Lambda}_0^{(\mathrm{CR})}\right]^{-1}\mathbf{X}_\ell\right| \\
&= |\delta\mathbf{\Lambda}_0^{(\mathrm{CR})}||\mathbf{X}_\ell^T\mathbf{X}_\ell|^{-1}\left|\mathbf{X}_\ell^T\mathbf{X}_\ell + \frac{w}{\delta}\mathbf{X}_\ell^T\left(\mathbf{I}_n + g_0\mathbf{X}_0(\mathbf{X}_0{}^T\mathbf{X}_0)^{-1}\mathbf{X}_0{}^T\right)\mathbf{X}_\ell\right| \\
&= |\delta\mathbf{\Lambda}_0^{(\mathrm{CR})}||\mathbf{X}_\ell^T\mathbf{X}_\ell|^{-1}\left|\mathbf{X}_\ell^T\mathbf{X}_\ell + \frac{w}{\delta}\mathbf{X}_\ell^T\mathbf{X}_\ell + \frac{wg_0}{\delta}\mathbf{X}_\ell^T\mathbf{X}_0(\mathbf{X}_0{}^T\mathbf{X}_0)^{-1}\mathbf{X}_0{}^T\mathbf{X}_\ell\right| \\
&= |\delta\mathbf{\Lambda}_0^{(\mathrm{CR})}||\mathbf{X}_\ell^T\mathbf{X}_\ell|^{-1}\left|\frac{w+\delta}{\delta}\mathbf{X}_\ell^T\mathbf{X}_\ell + \frac{wg_0}{\delta}\mathbf{X}_\ell^T\mathbf{X}_0(\mathbf{X}_0{}^T\mathbf{X}_0)^{-1}\mathbf{X}_0{}^T\mathbf{X}_\ell\right| \\
&= |\delta\mathbf{\Lambda}_0^{(\mathrm{CR})}||\mathbf{X}_\ell^T\mathbf{X}_\ell|^{-1}\left(\frac{w+\delta}{\delta}\right)^{d_\ell}|\mathbf{X}_\ell^T\mathbf{X}_\ell||\mathbf{X}_0{}^T\mathbf{X}_0|^{-1} \times \\
&\quad \times\left|\mathbf{X}_0{}^T\mathbf{X}_0 + \frac{wg_0}{\delta}\mathbf{X}_0{}^T\mathbf{X}_\ell\left(\frac{w+\delta}{\delta}\mathbf{X}_\ell^T\mathbf{X}_\ell\right)^{-1}\mathbf{X}_\ell^T\mathbf{X}_0\right| \\
&= \left(\frac{w+\delta}{\delta}\right)^{d_\ell}|\delta\mathbf{\Lambda}_0^{(\mathrm{CR})}||\mathbf{X}_0{}^T\mathbf{X}_0|^{-1} \times \\
&\quad \times\left|\mathbf{X}_0{}^T\mathbf{X}_0 + \frac{wg_0}{w+\delta}\mathbf{X}_0{}^T\mathbf{X}_\ell\left(\mathbf{X}_\ell^T\mathbf{X}_\ell\right)^{-1}\mathbf{X}_\ell^T\mathbf{X}_0\right|
\end{aligned} \tag{A.3}$$

$$\begin{aligned}
&= \left(\frac{w+\delta}{\delta}\right)^{d_\ell}|\delta\mathbf{\Lambda}_0^{(\mathrm{CR})}||\mathbf{X}_0{}^T\mathbf{X}_0|^{-1}\left|\mathbf{X}_0{}^T\mathbf{X}_0 + \frac{wg_0}{w+\delta}\mathbf{X}_0{}^T\mathbf{X}_0\right| \\
&= \left(\frac{w+\delta}{\delta}\right)^{d_\ell}|\delta\mathbf{\Lambda}_0^{(\mathrm{CR})}||\mathbf{X}_0{}^T\mathbf{X}_0|^{-1}\left(\frac{w+\delta+wg_0}{w+\delta}\right)^{d_0}|\mathbf{X}_0{}^T\mathbf{X}_0| \\
&= (w+\delta)^{d_\ell-d_0}\,\delta^{-d_\ell}\,(w+\delta+wg_0)^{d_0}\,|\delta\mathbf{\Lambda}_0^{(\mathrm{CR})}|
\end{aligned} \tag{A.4}$$

Note that the transition from (A.3) to the following equation is due to the fact that ${\mathbf{X}_0}^T\mathbf{X}_\ell\left(\mathbf{X}_\ell^T\mathbf{X}_\ell\right)^{-1}\mathbf{X}_\ell^T\mathbf{X}_0 = {\mathbf{X}_0}^T\mathbf{H}_\ell\mathbf{X}_0 = {\mathbf{X}_0}^T\mathbf{X}_0$, since ${\mathbf{X}_0}^T\mathbf{H}_\ell = {\mathbf{X}_0}^T$ for any sub-matrix $\mathbf{X}_0$ of $\mathbf{X}_\ell$. From (A.2) and (A.4) we have that

$$|\mathbf{V}_\ell^{(\mathrm{CR})}| = w^{d_\ell}\,(w+\delta)^{d_\ell-d_0}\,(w+\delta+wg_0)^{d_0}\,|\mathbf{X}_\ell^T\mathbf{X}_\ell|^{-1}. \tag{A.5}$$

B Derivation of Equation 3.22

$$\begin{aligned}
&\mathbf{y}^T(\mathbf{I}_n + \mathbf{X}_\ell\mathbf{V}_\ell^{(\mathrm{CR})}\mathbf{X}_\ell^T)^{-1}\mathbf{y}\\
&= \mathbf{y}^Y\Big(\mathbf{1}_n - \mathbf{X}_\ell\big(\mathbf{V}_{\boldsymbol{\beta}_\ell}^{(\mathrm{CR})-1} + \mathbf{X}_\ell^T\mathbf{X}_\ell\big)^{-1}\mathbf{X}_\ell^T\Big)\mathbf{y}\\
&= \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}_\ell\left(\delta^{-1}\Big[\mathbf{X}^T\Big(w^{-1}\mathbf{I}_n - \big(\delta\boldsymbol{\Lambda}_0^{(\mathrm{CR})} + w\mathbf{H}_\ell\big)^{-1}\Big)\mathbf{X}\Big] + \mathbf{X}_\ell^T\mathbf{X}_\ell\right)^{-1}\mathbf{X}_\ell^T\mathbf{y}\\
&= \mathbf{y}^T\mathbf{y} - \delta\mathbf{y}^T\mathbf{X}_\ell\left(w^{-1}\mathbf{X}_\ell^T\mathbf{X}_\ell - \mathbf{X}_\ell^T\Big(\delta\boldsymbol{\Lambda}_0^{(\mathrm{CR})} + w\mathbf{H}_\ell\Big)^{-1}\mathbf{X}_\ell + \delta\mathbf{X}_\ell^T\mathbf{X}_\ell\right)^{-1}\mathbf{X}_\ell^T\mathbf{y}\\
&= \mathbf{y}^T\mathbf{y} - \delta\mathbf{y}^T\mathbf{X}_\ell\left(\frac{1+\delta w}{w}\mathbf{X}_\ell^T\mathbf{X}_\ell - \mathbf{X}_\ell^T\Big(\delta\boldsymbol{\Lambda}_0^{(\mathrm{CR})} + w\mathbf{H}_\ell\Big)^{-1}\mathbf{X}_\ell\right)^{-1}\mathbf{X}_\ell^T\mathbf{y}\\
&= \mathbf{y}^T\mathbf{y} - \delta\mathbf{y}^T\mathbf{X}_\ell\left(\frac{1+\delta w}{w}\mathbf{X}_\ell^T\mathbf{X}_\ell - \mathbf{X}_\ell^T\Big(\delta\Big(\mathbf{I}_n - \frac{g_0}{g_0+1}\mathbf{H}_0\Big) + w\mathbf{H}_\ell\Big)^{-1}\mathbf{X}_\ell\right)^{-1}\mathbf{X}_\ell^T\mathbf{y}\\
&= \mathbf{y}^T\mathbf{y} - \frac{w\delta}{1+w\delta}\mathbf{y}^T\mathbf{X}_\ell\left(\mathbf{X}_\ell^T\mathbf{X}_\ell - \frac{w}{1+w\delta}\mathbf{X}_\ell^T\Big(\delta\Big(\mathbf{I}_n - \frac{g_0}{g_0+1}\mathbf{H}_0\Big) + w\mathbf{H}_\ell\Big)^{-1}\mathbf{X}_\ell\right)^{-1}\mathbf{X}_\ell^T\mathbf{y}\\
&= \mathbf{y}^T\mathbf{y} - \frac{w\delta}{1+w\delta}\mathbf{y}^T\mathbf{X}_\ell\left(\mathbf{X}_\ell^T\mathbf{X}_\ell - \frac{w}{(1+w\delta)\delta}\mathbf{X}_\ell^T\Big(\mathbf{I}_n - \frac{g_0}{g_0+1}\mathbf{H}_0 + \frac{w}{\delta}\mathbf{H}_\ell\Big)^{-1}\mathbf{X}_\ell\right)^{-1}\mathbf{X}_\ell^T\mathbf{y},
\end{aligned} \tag{B.1}$$

where $\mathbf{H}_\ell = \mathbf{X}_\ell(\mathbf{X}_\ell^T\mathbf{X}_\ell)^{-1}\mathbf{X}_\ell^T$ and $\mathbf{H}_0 = \mathbf{X}_0(\mathbf{X}_0^T\mathbf{X}_0)^{-1}\mathbf{X}_0^T = \mathbf{1}_n(\mathbf{1}_n^T\mathbf{1}_n)^{-1}\mathbf{1}_n^T = n^{-1}\mathbf{1}_n\mathbf{1}_n^T$. For the derivation of the first expression, see Woodbury's matrix identity (Harville 1997, p. 423–426). For large values of δ and $g_0 >> \delta$ we have approximately $w \approx 1$, $\frac{w\delta}{1+w\delta} \approx 1$ and $\frac{w}{(1+w\delta)\delta} \approx 0$, which yields the approximation

$$\mathbf{y}^T(\mathbf{I}_n + \mathbf{X}_\ell\mathbf{V}_\ell^{(\mathrm{CR})}\mathbf{X}_\ell^T)^{-1}\mathbf{y} \approx \mathbf{y}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}_\ell\left(\mathbf{X}_\ell^T\mathbf{X}_\ell\right)^{-1}\mathbf{X}_\ell^T\mathbf{y} \equiv \mathrm{RSS}_\ell. \tag{B.2}$$

C Additional results from Section 4.1.1

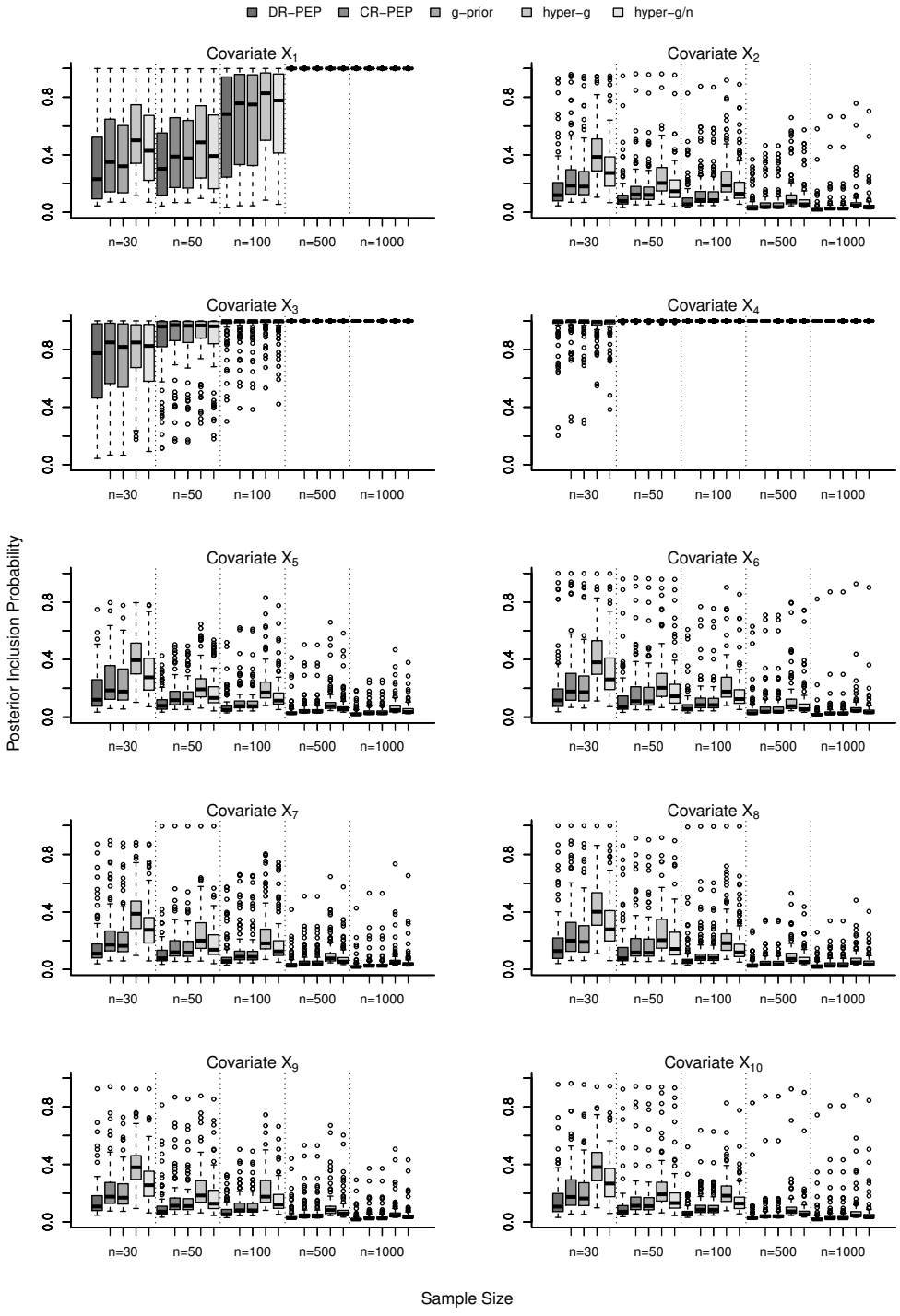

Figure C.1: Posterior inclusion probabilities of each covariate for Scenario 1 (independent covariates) under Laplace errors (M-open case).

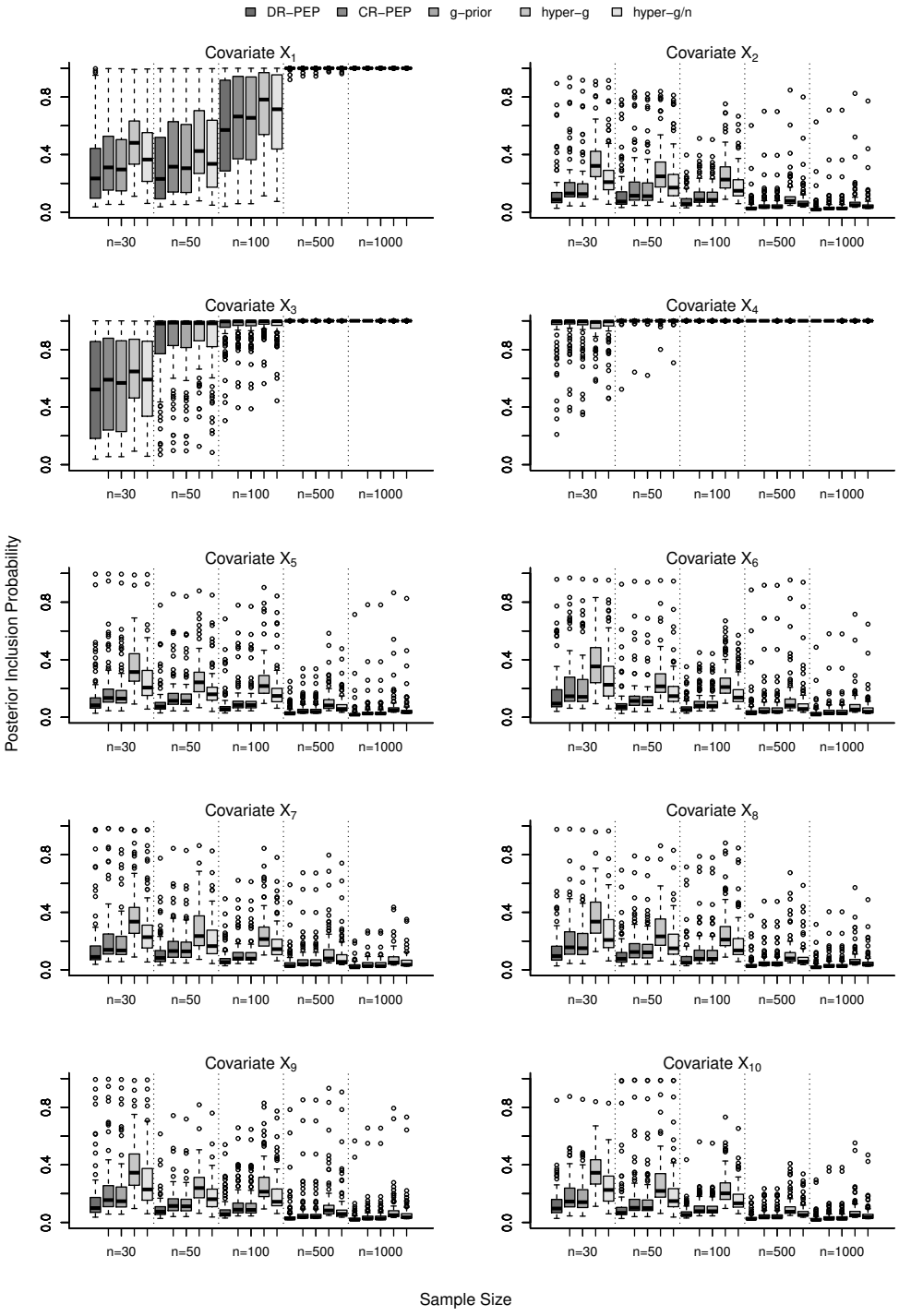

Figure C.2: Posterior inclusion probabilities of each covariate for Scenario 1 (independent covariates) under asymmetric normal errors (M-open case).

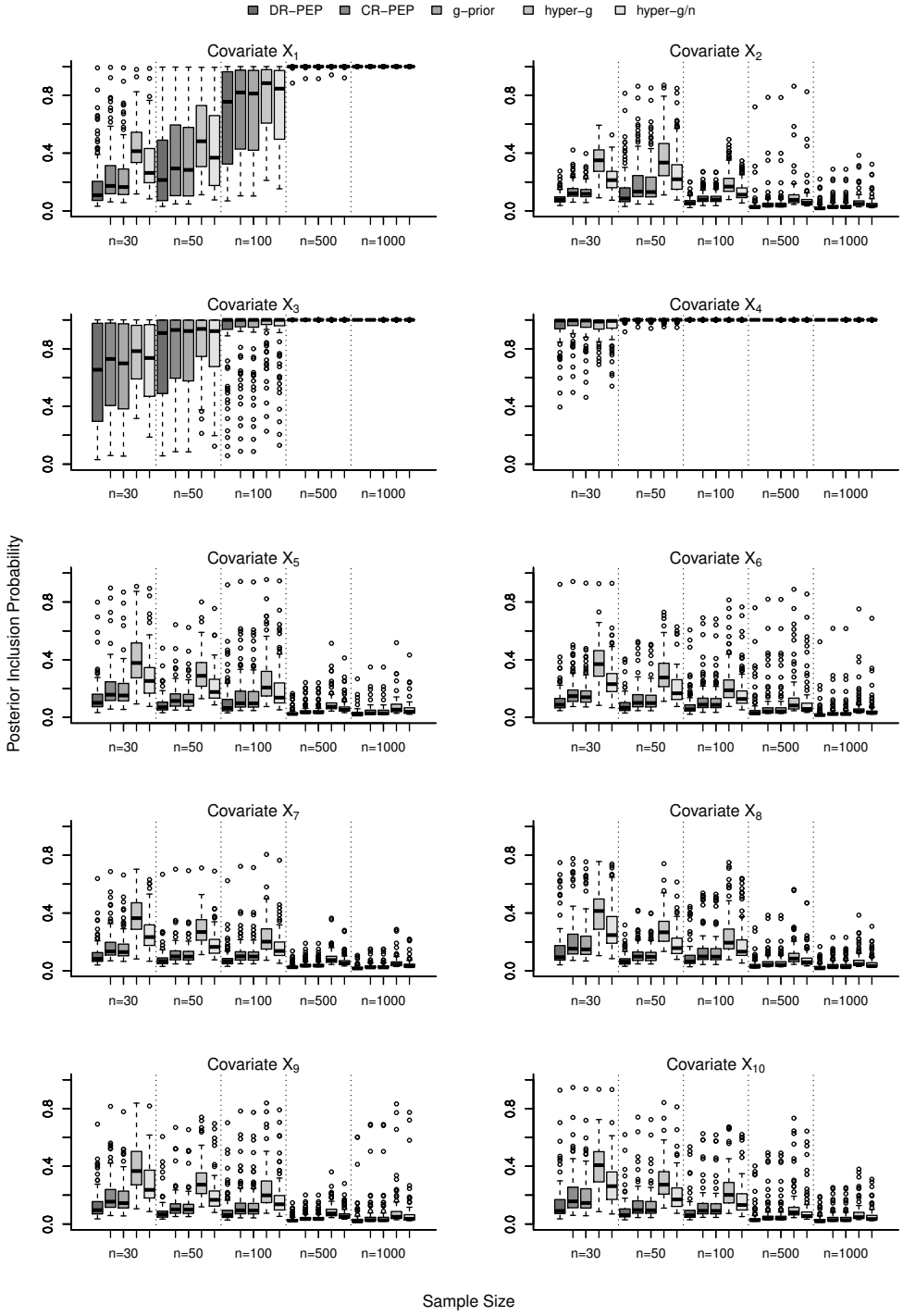

Figure C.3: Posterior inclusion probabilities of each covariate for Scenario 1 (independent covariates) heteroscedastic normal errors (M-open case).

D Additional results from Section 4.1.2

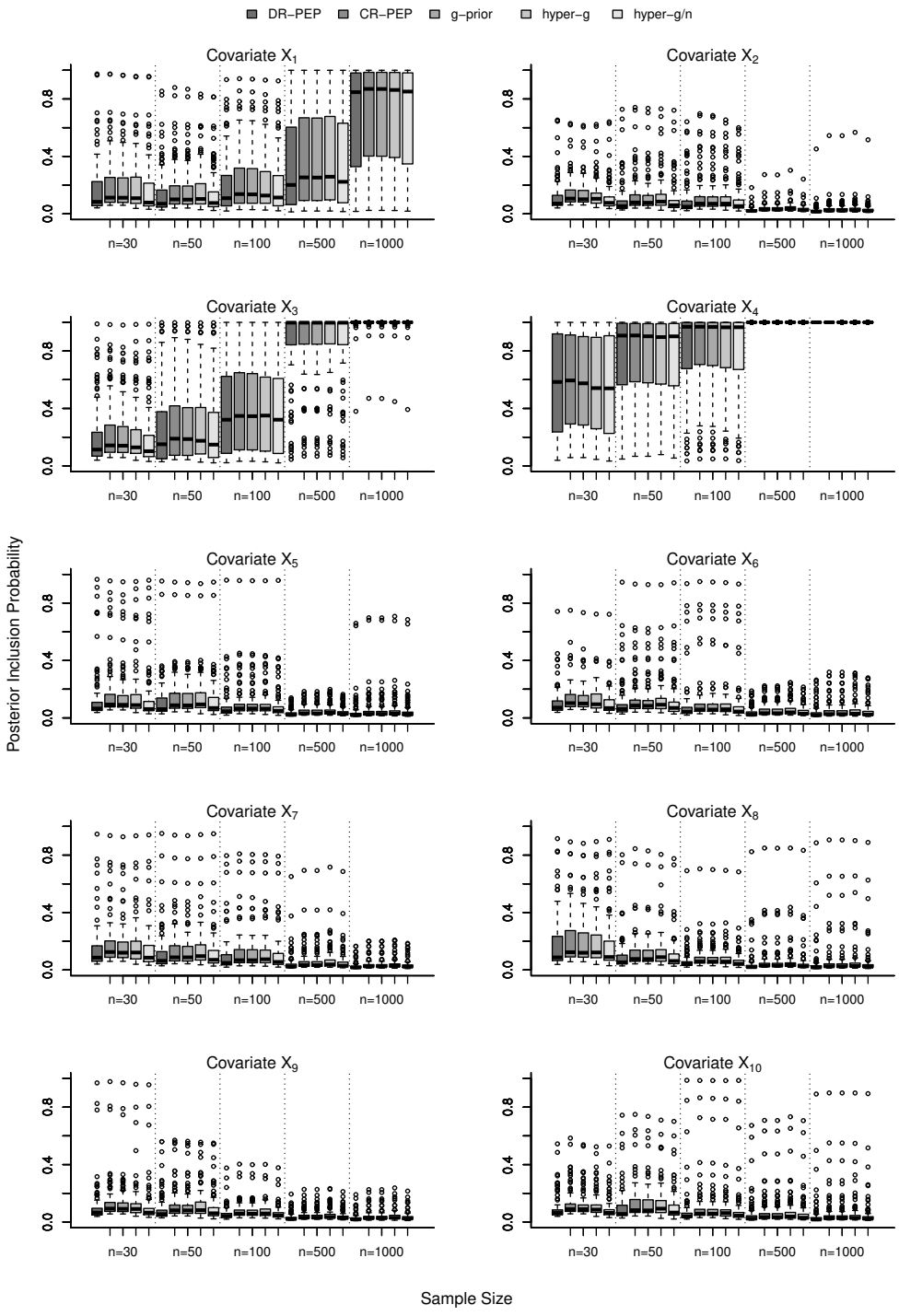

Figure D.1: Posterior inclusion probabilities of each covariate for Scenario 2 (highly-correlated covariates) under Laplace errors (M-open case).

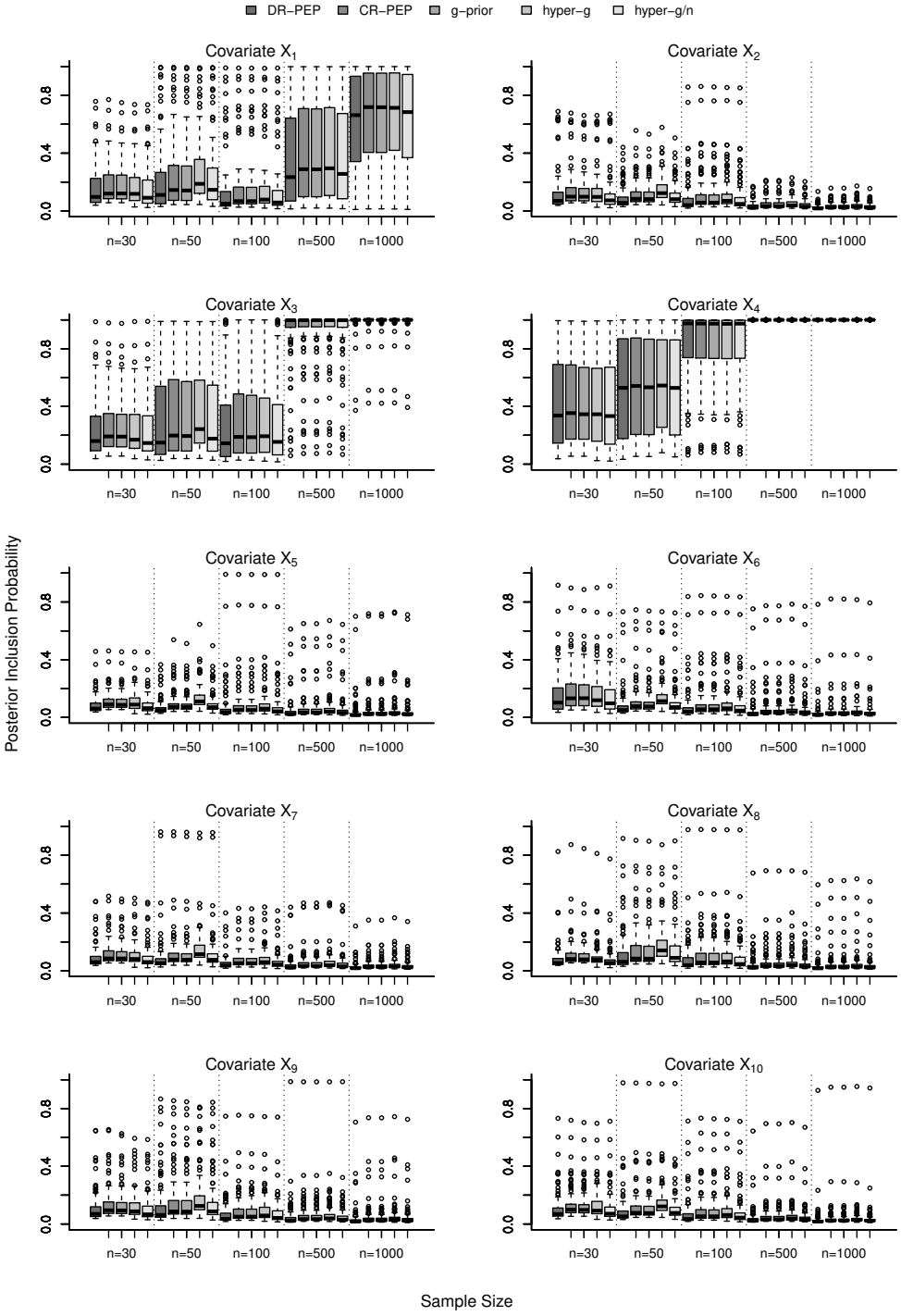

Figure D.2: Posterior inclusion probabilities of each covariate for Scenario 2 (highly-correlated covariates) under asymmetric normal errors (M-open case).

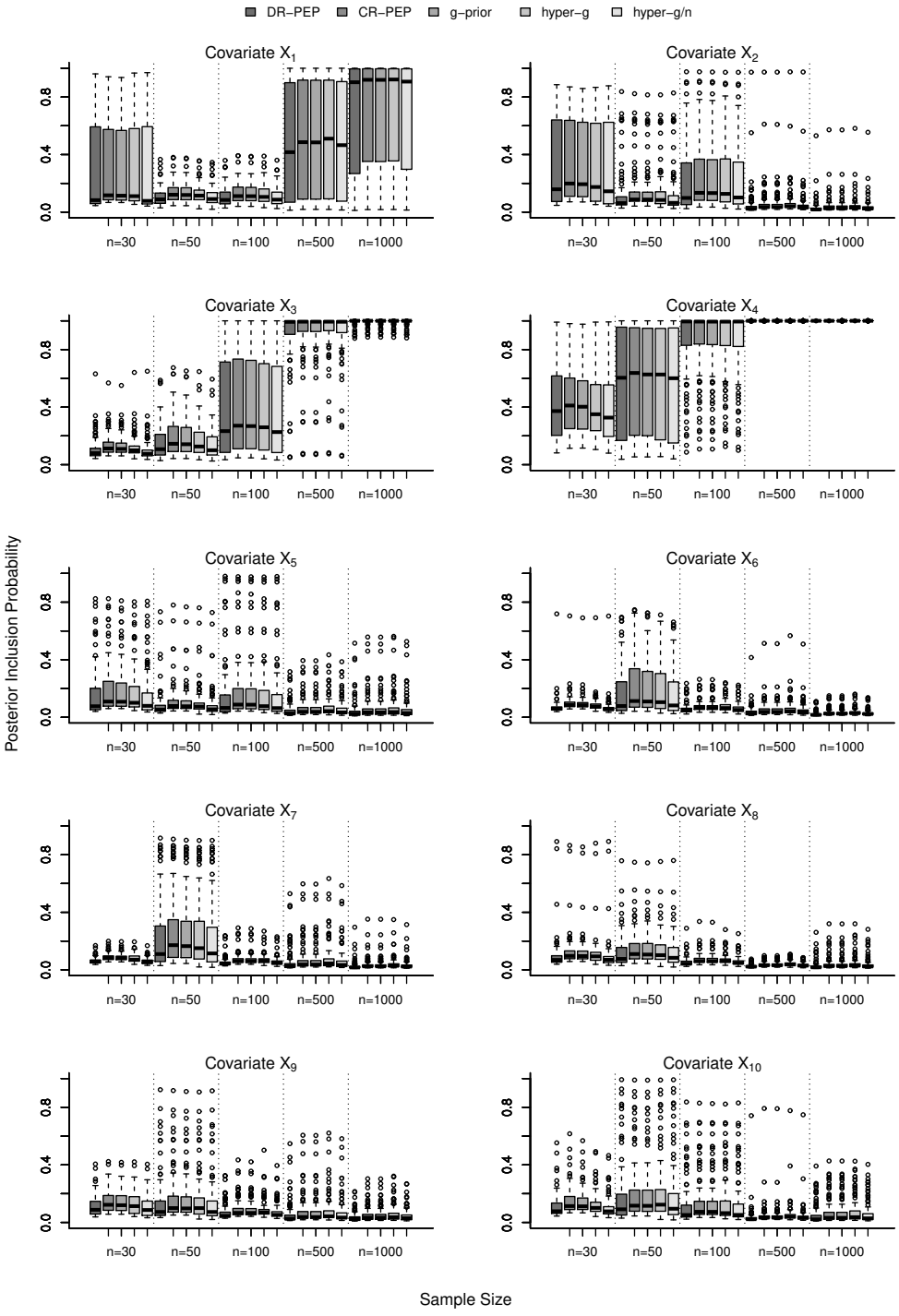

Figure D.3: Posterior inclusion probabilities of each covariate for Scenario 2 (highly-correlated covariates) heteroscedastic normal errors (M-open case).

E Additional results from Section 4.2

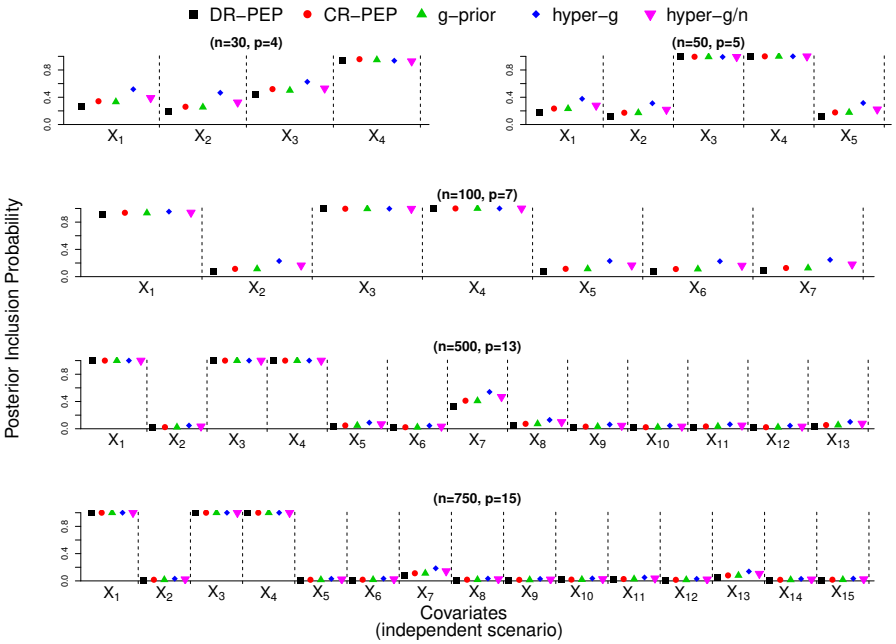

Figure E.1: Posterior inclusion probabilities of each covariate for different sample sizes and number of covariates for the independence scenario.

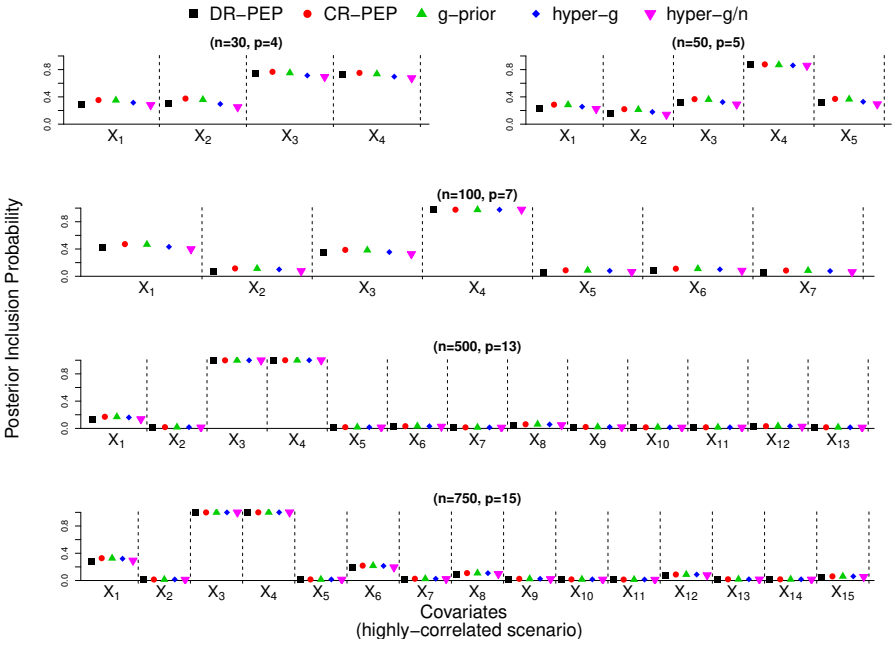

Figure E.2: Posterior inclusion probabilities of each covariate for different sample sizes and number of covariates for the high-correlation scenario.

References

Aitkin, M. (1991), 'Posterior Bayes Factors', *Journal of the Royal Statistical Society. Series B (Methodological)* **53**, 111–142.

Bartlett, M. (1957), 'Comment on D.V. Lindley's statistical paradox', *Biometrika* **44**, 533–534.

Bayarri, M. J., Berger, J. O., Forte, A. & García-Donato, G. (2012), 'Criteria for Bayesian model choice with application to variable selection', *The Annals of Statistics* **40**, 1550–1577.

Berger, J. & Pericchi, L. (2004), 'Training samples in objective model selection', *Annals of Statistics* **32**, 841–869.

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

Consonni, G., Fouskakis, D., Liseo, B. & Ntzoufras, I. (2018), 'Prior distributions for objective Bayesian analysis', *Bayesian Analysis* **13**, 627–679.

de Carvalho, M., Page, G. L. & Barney, B. J. (2019), 'On the Geometry of Bayesian Inference', *Bayesian Analysis (in press)* .

Dellaportas, P., Forster, J. J. & Ntzoufras, I. (2012), 'Joint specification of model space and parameter space prior distributions', *Statistical Science* **27**, 232–246.

Fernández, C., Ley, E. & Steel, M. F. J. (2001), 'Benchmark priors for Bayesian model averaging', *Journal of Econometrics* **100**, 381–427.

Forte, A., Garcia-Donato, G. & Steel, M. (2018), 'Methods and Tools for Bayesian Variable Selection and Model Averaging in Normal Linear Regression', *International Statistical Review* **86**, 237–258.

Fouskakis, D. & Ntzoufras, I. (2016), 'Power-conditional-expected priors: Using g-priors with random imaginary data for variable selection', *Journal of Computational and Graphical Statistics* **25**, 647–664.

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

Fouskakis, D., Ntzoufras, I. & Perrakis, K. (2018), 'Power-expected-posterior priors for generalized linear models', *Bayesian Analysis* **13**, 721–748.

Friel, N. & Pettitt, A. N. (2008), 'Marginal likelihood estimation via power posteriors', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 589–607.

Ghosh, J. & Ghattas, A. E. (2015), 'Bayesian variable selection under collinearity', *The American Statistician* **69**, 165–173.

Good, I. (2004), *Probability and the Weighting of Evidence*, Haffner, New York, USA.

Harville, D. (1997), *Matrix Algebra From a Statistician's Perspective*, Springer, New York, USA.

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

Ji, C. & Schmidler, S. C. (2013), 'Adaptive Markov Chain Monte Carlo for Bayesian Variable Selection', *Journal of Computational and Graphical Statistics* **22**, 708–728.

Johnson, V. E. & Rossell, D. (2012), 'Bayesian Model Selection in High-Dimensional Settings', *Journal of the American Statistical Association* **107**, 649–660.

Kass, R. E. & 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.

Liang, F., Paulo, R., Molina, G., Clyde, M. A. & Berger, J. O. (2008), 'Mixtures of g-priors for Bayesian variable selection', *Journal of the American Statistical Association* **103**, 410–423.

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

Mudholkar, G. S. & Hutson, A. D. (2000), 'The epsilon-skew-normal distribution for analyzing near-normal data', *Journal of Statistical Planning and Inference* **83**, 291–309.

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

Rodriguez, A. (2015), 'Discussion of "Power-Expected-Posterior Priors in Generalized Linear Models"', *O-Bayes15: 11th International Workshop on Objective Bayes Methodology* .

Rossell, D. & Rubio, F. J. (2018), 'Tractable Bayesian variable selection: beyond normality', *Journal of the American Statistical Association* **113**, 1742–1758.

Schwarz, G. (1978), 'Estimating the dimension of a model', *Annals of Statistics* **6**, 461–464.

Scott, J. G. & Berger, J. O. (2010), 'Bayes and empirical-Bayes multiplicity adjustment in the variable-selection problem', *The Annals of Statistics* **38**, 2587–2619.

Spiegelhalter, D. & Smith, A. (1982), 'Bayes Factors for Linear and Log-Linear Models with Vague Prior Information', *Journal of the Royal Statistical Society: Series B* **44**, 377–387.

Zellner, A. (1976), "Bayesian and non-Bayesian analysis of the regression model with multivariate Student-t error terms", *Journal of the American Statistical Association* **71**, 400–405.