# A Comparison of Power–Expected–Posterior Priors in Shrinkage Regression

G. Tzoumerkas[1], D. Fouskakis[1] and I. Ntzoufras[2]

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

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

Contributing authors: tzoumg@mail.ntua.gr; fouskakis@math.ntua.gr; ntzoufras@aueb.gr;

**Abstract**

The Power-Expected-Posterior (PEP) prior framework provides us a convenient and objective method to deal with variable selection problems, under the Bayesian perspective, in regression models. The PEP prior inherits all of the advantages of Expected-Posterior-Prior (EPP). Furthermore, it avoids the need of selection of imaginary data and mitigates their effect over the final posterior. Under the PEP prior methodology, an initial (usually default) baseline prior is updated using imaginary data. In this work, focus is given in normal regression models when the number of observations is smaller than the number of explanatory variables. We introduce the PEP prior methodology using different baseline shrinkage priors, we present a computational method and we perform comparisons in simulated and real life data-sets.

**Keywords:** Bayesian variable selection, imaginary training sample, MCMC, objective priors, shrinkage priors, sparse datasets

## 1 Introduction

We consider the variable selection problem for normal regression models, where the number of observations $n$ is smaller than the number of explanatory variables $p$. Suppose the model space is consisting of all combinations of available

covariates. Then for every model $M_\ell$ in model space $\mathcal{M}$ the likelihood is given by

$$\boldsymbol{y} \mid (X_\ell, \boldsymbol{\beta}_\ell, \sigma^2) \sim N_n(X_\ell \boldsymbol{\beta}_\ell, \sigma^2 I_n),$$

where $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ denotes the response data, $X_\ell$ is the $n \times p_\ell$ design matrix; where $p_\ell$ is the number of explanatory variables under model $M_\ell$, $\boldsymbol{\beta}_\ell$ is a vector of length $p_\ell$ of the effects of each covariate on the response variable, $I_n$ is the $n \times n$ identity matrix and $\sigma^2$ is the error variance. We assume that $\boldsymbol{y}$ and the columns of the design matrix of the full model (including all available explanatory variables) have been centred on their corresponding means, so no intercept is used in our model formulation.

Under the Bayesian perspective, we can use posterior odds [14] in order to compare any two models $M_1$ and $M_2$:

$$PO_{12} = \frac{\pi(M_1 \mid \boldsymbol{y})}{\pi(M_2 \mid \boldsymbol{y})} = \frac{m_1(\boldsymbol{y})}{m_2(\boldsymbol{y})} \times \frac{\pi(M_1)}{\pi(M_2)}, \tag{1}$$

where $\pi(M_\ell \mid \boldsymbol{y})$ is the posterior probability of model $M_\ell$, $\pi(M_\ell)$ is the prior probability of $M_\ell$, while $m_\ell(\boldsymbol{y})$ is the marginal likelihood of $M_\ell$ given by

$$m_\ell(\boldsymbol{y}) = \int f(\boldsymbol{y} \mid \boldsymbol{\beta}_\ell, \sigma, M_\ell) \pi(\boldsymbol{\beta}_\ell, \sigma \mid M_\ell) d\boldsymbol{\beta}_\ell d\sigma.$$

In the last expression $f(\boldsymbol{y} \mid \boldsymbol{\beta}_\ell, \sigma, M_\ell)$ is denoting the likelihood of model $M_\ell$, with model parameters $(\boldsymbol{\beta}_\ell, \sigma)$, having prior distribution $\pi(\boldsymbol{\beta}_\ell, \sigma \mid M_\ell)$. The ratio of the two marginal likelihoods, of the two models, which appears in equation (1), is called Bayes factor ($BF_{12}$), i.e. $BF_{12} = \frac{m_1(\boldsymbol{y})}{m_2(\boldsymbol{y})}$. The posterior probability of any model $M_\ell$, is given by

$$\pi(M_\ell \mid \boldsymbol{y}) = \frac{m_\ell(\boldsymbol{y})\pi(M_\ell)}{\sum_{M_k \in \mathcal{M}} m_k(\boldsymbol{y})\pi(M_k)}.$$

The model with the highest posterior probability (maximum a-posteriori (MAP) model) is often chosen as the optimal, under the Bayesian model choice problem. For large model spaces, we often use MCMC methods to estimate $\pi(M_\ell \mid \boldsymbol{y})$. These estimates have the disadvantage that they converge to the true quantities with a slow rate. As an alternative strategy, we could use the marginal posterior inclusion probabilities [11]. For each covariate $X_j, j = 1, \ldots, p$, the marginal posterior inclusion probability is defined as

$$\pi(\gamma_j = 1 \mid \boldsymbol{y}) = \sum_{M_\xi \in \mathcal{M}_j} \pi(M_\xi \mid \boldsymbol{y}),$$

where $\gamma_j$ is a binary indicator that takes the value 1 if covariate $X_j$ belongs to a model and 0 otherwise. In the above expression $\mathcal{M}_j = \{M_\ell \in \mathcal{M} : \gamma_j =$

$1\} \subset \mathcal{M}$ is defined as the set containing all models with $X_j$. Using the posterior variable inclusion probabilities we can define the median probability (MP) model which is the model containing only the covariates with marginal posterior inclusion probability above 0.5. The importance of the marginal posterior inclusion probabilities in Bayesian variable selection can be found in [2] where it is proven that the median probability (MP) model has better predictive properties than the maximum a-posteriori model under certain conditions.

As it is now clear, we must set priors both for the model space and the parameter space of each model. Regarding the prior on the model space, for sparsity reasons, we consider the uniform prior on model size, as a special case of the beta-binomial prior; see [23]. With respect to the prior distribution on the coefficients in each model, since we are not confident about any given set of regressors as explanatory variables, little prior information about their regression coefficients should be expected. This argument alone justifies the need for an objective model choice approach in which vague prior information is assumed. Furthermore, we need to use a prior capable to deal with the $n < p$ scenario. Finally, regarding the (common across models) error variance, the reference prior will be used, i.e. $\pi(\sigma^2) \propto \sigma^{-2}$.

## 1.1 Shrinkage priors

A common way to deal with normal regression problems, when $n < p$, is by using shrinkage methods. Under the Bayesian perspective this can be done using a shrinkage prior on the model coefficients. By the term shrinkage we refer to the behavior where non-important covariate effects will shrink towards zero. Shrinkage priors share eminent theoretical properties, compelling computational complexity and great empirical performance (e.g. [6], [21]).

A shrinkage prior, can often be conceived as a scale-mixture prior, which is placed on the regression coefficients of every possible model. Something that characterizes shrinkage priors, is their hyperparameters: the global shrinkage hyperparameter, that determines the overall sparsity in the whole parameter vector and the local shrinkage hyperparameter, where a distinct shrinkage parameter is considered specifically for every single effect and controls the shrinkage of this individual effect. Depending on the shrinkage prior, the global parameter or the local parameters may be absent from the formation.

By assuming a shrinkage prior, on the vector of regression coefficients $\boldsymbol{\beta}_\ell$, in most of the cases a prior with heavy mass around zero is being produced and by so, small effects will shrink towards zero. Furthermore, the necessity of heavy tails, is important, as it averts true non-zero effects to get shrinked. In Table 1, we mention the most popular shrinkage priors, where by $\tau$ we refer to local shrinkage hyperparameters and by $\lambda$ to global shrinkage hyperparameters. In all cases except the last two (Ridge g-prior and MG prior) independent priors for the coefficients of model $M_\ell$ are used, and thus $j = 1, \ldots, p_\ell$ (the vast majority of the shrinkage priors in the Bayesian literature are independent). Under the MG prior, we can say that $g$ takes the role of the global shrinkage parameter $\lambda$ and we have no local shrinkage parameters. This prior

**Table 1** List of shrinkage priors

| # | Name | conditional prior of $\boldsymbol{\beta}_\ell$ | shrinkage hyparameters |
|---|---|---|---|
| 1 | LASSO [19] | $\beta_j \mid \tau_j^2, \sigma^2 \sim N(0, \sigma^2\tau_j^2)$ | $\tau_j^2 \mid \lambda \sim Exp(\frac{\lambda^2}{2})$<br>$\lambda \sim HC(0,1)^*$ |
| 2 | Horseshoe [3] | $\beta_j \mid \lambda, \tau_j, \sigma^2 \sim N(0, \sigma^2\lambda^2\tau_j^2)$ | $\tau_j \sim HC(0,1)$<br>$\lambda \sim HC(0,1)$ |
| 3 | Ridge [13] | $\beta_j \mid \lambda, \sigma^2 \sim N(0, \sigma^2\frac{1}{\lambda})$ | $\lambda \sim HC(0,1)$ |
| 4 | Local Student's t [25] | $\beta_j \mid \tau_j^2, \sigma^2 \sim N(0, \sigma^2\tau_j^2)$ | $\tau_j^2 \mid \lambda \sim IG(\frac{k}{2}, \frac{k}{2\lambda})^{**}$<br>$\lambda \sim HC(0,1)$<br>$k = 1$ |
| 5 | Elastic Net [16] | $\beta_j \mid \lambda_2, \tau_j, \sigma^2 \sim N(0, \sigma^2\frac{1}{\lambda_2+\tau_j^2})$ | $\tau_j^2 \mid \lambda_1 \sim Exp(\frac{\lambda_1^2}{2})$<br>$\lambda_1, \lambda_2 \sim HC(0,1)$ |
| 6 | Beta Prime [1] | $\beta_j \mid \tau_j^2, \sigma^2 \sim N(0, \sigma^2\tau_j^2)$ | $\tau_j^2 \sim Inv-Beta(a,b)$<br>$a, b$ fixed*** |
| 7 | Ridge g-prior [12] | $\boldsymbol{\beta}_\ell \mid \lambda, \sigma^2 \sim N_{p_\ell}(\mathbf{0}, \sigma^2\Phi_\ell)$<br>$\Phi_\ell = g(X_\ell^T X_\ell + \lambda I_{p_\ell})^{-1}$ | $g = n$<br>$\lambda = 0.5$ |
| 8 | MG prior**** [18] | $Z_\ell^T\boldsymbol{\beta}_\ell \mid g, \sigma^2 \sim N_{p_\ell}(\mathbf{0}, \sigma^2\Psi_\ell(g))$ | $g \sim Inv-Beta(a,b)$<br>$a, b$ fixed |

*$HC(x_0, \gamma)$: truncated Cauchy distribution with location parameter $x_0$, scale parameter $\gamma$ and support $(x_0, \infty)$.

**$IG(\alpha, \beta)$: Inverse Gamma distribution with shape parameter $\alpha$ and scale parameter $\beta$.

***$a$ and $b$ are estimated using a data-adaptive method based on marginal maximum likelihood; see [1].

****Maruyama and George generalised g-prior. $Z_\ell$ is an orthogonal matrix, $\Psi_\ell(g)$ is a diagonal matrix. For the construction of these matrices and the default choices for $a, b$, see [18].

can be considered as a generalization of the Zellner's g-prior, which allows for $p > n$, and is placed on the rotated coefficients after orthogonalization (see [18]). To keep the notation as simple as possible, in the rest of the paper, when referring to the MG prior, we continue calling these rotated coefficients $\boldsymbol{\beta}_\ell$.

## 1.2 Power-Expected-Posterior prior

A principal approach to define objective priors is the use of random imaginary training data [5]. Power-Expected-Posterior (PEP) prior [7], [8], uses this methodology. In particular, for the normal linear regression model, the PEP prior is defined as

$$\pi_\ell^{PEP}(\boldsymbol{\beta}_\ell \mid \sigma^2, \delta, X_\ell^*) = \int \pi_\ell^N(\boldsymbol{\beta}_\ell \mid \boldsymbol{y}^*, \sigma^2, \delta, X_\ell^*) m_0^N(\boldsymbol{y}^* \mid \sigma^2, \delta, X_0^*) d\boldsymbol{y}^*, \quad (2)$$

$$\pi_\ell^{PEP}(\sigma^2) = \pi^N(\sigma^2) \propto \frac{1}{\sigma^2},$$

with

$$\pi_\ell^N(\boldsymbol{\beta}_\ell \mid \boldsymbol{y}^*, \sigma^2, \delta, X_\ell^*) \propto f_\ell(\boldsymbol{y}^* \mid \boldsymbol{\beta}_\ell, \sigma^2, \delta, X_\ell^*) \pi_\ell^N(\boldsymbol{\beta}_\ell \mid \sigma^2, X_\ell^*), \quad (3)$$

$$f_\ell(\boldsymbol{y}^* \mid \boldsymbol{\beta}_\ell, \sigma^2, \delta, X_\ell^*) = \frac{f_\ell(\boldsymbol{y}^* \mid \boldsymbol{\beta}_\ell, \sigma^2, X_\ell^*)^{1/\delta}}{\int f_\ell(\boldsymbol{y}^* \mid \boldsymbol{\beta}_\ell, \sigma^2, X_\ell^*)^{1/\delta} d\boldsymbol{y}^*} \tag{4}$$

and

$$m_0^N(\boldsymbol{y}^* \mid \sigma^2, \delta, X_0^*) = \int f_0(\boldsymbol{y}^* \mid \boldsymbol{\beta}_0, \sigma^2, \delta, X_0^*) \pi_0^N(\boldsymbol{\beta}_0 \mid \sigma^2, X_0^*) d\boldsymbol{\beta}_0,$$

with

$$f_0(\boldsymbol{y}^* \mid \boldsymbol{\beta}_0, \sigma^2, \delta, X_0^*) = \frac{f_0(\boldsymbol{y}^* \mid \boldsymbol{\beta}_0, \sigma^2, X_0^*)^{1/\delta}}{\int f_0(\boldsymbol{y}^* \mid \boldsymbol{\beta}_0, \sigma^2, X_0^*)^{1/\delta} d\boldsymbol{y}^*}.$$

In the above equations, we have set $\boldsymbol{y}^*$ to be the imaginary observations of size $n^*$ and $X_\ell^*$ the imaginary design matrix of model $M_\ell$. By $\pi_\ell^N(\boldsymbol{\beta}_\ell \mid \boldsymbol{y}^*, \sigma^2, \delta, X_\ell^*)$ we denote the conditional on $\sigma^2$ posterior of $\boldsymbol{\beta}_\ell$, using a baseline prior $\pi_\ell^N(\boldsymbol{\beta}_\ell \mid \sigma^2, X_\ell^*)$ and data $\boldsymbol{y}^*$. In equation (4) the likelihood of imaginary observations is raised to the power of $1/\delta$ and density normalized. By doing this we decrease the effect of the imaginary data. For $\delta = 1$, equation (2) results to the Expected-Posterior-Prior (EPP) [20]. In order to have a unit information interpretation [15], we set $\delta = n^*$ and in order to avoid any effect of the choice of imaginary design matrices, we set $n^* = n$. In equation (2), $m_0^N(\boldsymbol{y}^* \mid \sigma^2, \delta, X_0^*)$, is the prior predictive distribution (or the marginal likelihood), evaluated at $\boldsymbol{y}^*$, of the reference model $M_0$, given $\sigma^2$. For the calculation of this marginal likelihood we use the normalized power likelihood of the reference model. We denote by $X_0$ the design matrix and by $\boldsymbol{\beta}_0$ the coefficients of this reference model. As a reference model, in the rest of the paper, we consider, for reasons of parsimony, the model with only the intercept (null model). Finally, for every model $M_\ell$, the marginal likelihood under the baseline prior, given $\sigma^2$, is

$$m_\ell^N(\boldsymbol{y}^* \mid \sigma^2, \delta, X_\ell^*) = \int f_\ell(\boldsymbol{y}^* \mid \boldsymbol{\beta}_\ell, \sigma^2, \delta, X_\ell^*) \pi_\ell^N(\boldsymbol{\beta}_\ell \mid \sigma^2, X_\ell^*) d\boldsymbol{\beta}_\ell. \tag{5}$$

## 2 PEP-Shrinkage Prior Methodology

In the above formulation, by choosing as a baseline prior $\pi_\ell^N(\boldsymbol{\beta}_\ell \mid \sigma^2, X_\ell^*)$ a shrinkage prior (see Table 1), a PEP-Shrinkage prior is created and thus we can apply the PEP prior methodology in shrinkage problems.

PEP priors can be considered as fully automatic, objective Bayesian methods for model comparison in regression models (see for example [5] and [7]). They are developed through the utilization of the device of "imaginary" samples, coming from the simplest model under comparison. Therefore, PEP priors offer several advantages, among which they have an appealing interpretation based on imaginary training data coming from a prior predictive distribution and also provide an effective way to establish compatibility of priors among models (see [4]), through their dependence on a common marginal data distribution. Thus, the PEP methodology can be applied also with proper baseline

**Table 2** The matrix $\Omega_\ell$ for every baseline prior used in PEP-Shrinkage methodology

| shrinkage prior | $\boldsymbol{\theta}_\ell$ | $\Omega_\ell$ |
|---|---|---|
| LASSO | $(\lambda, \tau_1^2, \ldots, \tau_{p_\ell}^2)$ | $diag(\tau_1^2, \ldots, \tau_{p_\ell}^2)$ |
| Horseshoe | $(\lambda, \tau_1, \ldots, \tau_{p_\ell})$ | $diag(\lambda^2\tau_1^2, \ldots, \lambda^2\tau_{p_\ell}^2)$ |
| Ridge | $\lambda$ | $diag(\lambda^{-1}, \ldots, \lambda^{-1})$ |
| Local Student's t | $(\lambda, \tau_1^2, \ldots, \tau_{p_\ell}^2)$ | $diag(\tau_1^2, \ldots, \tau_{p_\ell}^2)$ |
| Elastic Net | $(\lambda_1, \lambda_2, \tau_1^2, \ldots, \tau_{p_\ell}^2)$ | $diag((\lambda_2 + \tau_1^2)^{-1}, \ldots, (\lambda_2 + \tau_{p_\ell}^2)^{-1})$ |
| Beta Prime | $(\tau_1^2, \ldots, \tau_{p_\ell}^2)$ | $diag(\tau_1^2, \ldots, \tau_{p_\ell}^2)$ |
| Ridge g-prior | $\lambda$ | $g(X_\ell^{*T} X_\ell^* + \lambda I_{p_\ell})^{-1}$ |
| MG prior | $g$ | $\Psi_\ell(g)$ |

prior distributions. Furthermore, by choosing the simplest model, as a reference model, to generate the imaginary samples, the PEP prior shares common ideas with the skeptical-prior approach described by [24].

In the following, under any model $M_\ell$, the likelihood is given by

$$f_\ell(\boldsymbol{y} \mid X_\ell, \boldsymbol{\beta}_\ell, \sigma^2) = f_{N_n}(\boldsymbol{y};\ X_\ell\ \boldsymbol{\beta}_\ell, \sigma^2\ I_n),$$

where $f_{N_d}(\boldsymbol{y};\ \boldsymbol{\mu}, \Sigma)$ is denoting the $d$-dimensional normal distribution with mean $\boldsymbol{\mu}$ and covariance matrix $\Sigma$. Under (4) the likelihood of the imaginary data $\boldsymbol{y}^*$, under model $M_\ell$, is given by

$$f_\ell(\boldsymbol{y}^* \mid X_\ell^*, \boldsymbol{\beta}_\ell, \sigma^2, \delta) = f_{N_{n^*}}(\boldsymbol{y}^*;\ X_\ell^*\ \boldsymbol{\beta}_\ell, \delta\ \sigma^2\ I_{n^*}).$$

From Table 1 it is obvious that all shrinkage priors that will use as baseline priors under the PEP methodology, have the following general form

$$\pi_\ell^N(\boldsymbol{\beta}_\ell \mid \boldsymbol{\theta}_\ell, \sigma^2) = f_{N_{p_\ell}}(\boldsymbol{\beta}_\ell;\ \mathbf{0}, \sigma^2\Omega_\ell),$$

where $\Omega_\ell \equiv \Omega_\ell(\boldsymbol{\theta}_\ell)$ is a $p_\ell \times p_\ell$ matrix; for more details see Table 2. In the rest of the paper, by $\boldsymbol{\theta}_\ell$ we denote the vector containing all the shrinkage hyperparameters (global and local) of model $M_\ell$, with a prior distribution denoting by $\pi(\boldsymbol{\theta}_\ell)$.

## 2.1 PEP-Shrinkage prior

The conditional posterior distribution $\pi_\ell^N(\boldsymbol{\beta}_\ell \mid \boldsymbol{y}^*, \sigma^2, \delta, X_\ell^*, \boldsymbol{\theta}_\ell)$, using the baseline prior and the imaginary data is given by

$$\begin{aligned}\pi_\ell^N(\boldsymbol{\beta}_\ell \mid \boldsymbol{y}^*, \sigma^2, \delta, X_\ell^*, \boldsymbol{\theta}_\ell) &\propto f_\ell(\boldsymbol{y}^* \mid X_\ell^*, \boldsymbol{\beta}_\ell, \sigma^2, \delta)\pi_\ell^N(\boldsymbol{\beta}_\ell \mid \boldsymbol{\theta}_\ell, \sigma^2) \\ &= f_{N_{n^*}}(\boldsymbol{y}^*;\ X_\ell^*\ \boldsymbol{\beta}_\ell, \delta\ \sigma^2\ I_{n^*}) f_{N_{p_\ell}}(\boldsymbol{\beta}_\ell;\ \mathbf{0}, \sigma^2\Omega_\ell)\end{aligned}$$

and so we have have that

$$\pi_\ell^N(\boldsymbol{\beta}_\ell \mid \boldsymbol{y}^*, \sigma^2, \delta, X_\ell^*, \boldsymbol{\theta}_\ell) = f_{N_{p_\ell}}(\boldsymbol{\beta}_\ell;\ \delta^{-1} W_\ell X_\ell^{*T} \boldsymbol{y}^*, \sigma^2 W_\ell),$$

where $W_\ell = [\delta^{-1} X_\ell^{*T} X_\ell^* + \Omega_\ell^{-1}]^{-1}$. Moreover, from equation (5), for any model $M_\ell$, the prior predictive distribution, under the baseline prior, conditional on $\sigma^2$ and $\boldsymbol{\theta}_\ell$ is

$$\begin{aligned} m_\ell^N(\boldsymbol{y}^* \mid \sigma^2, \delta, X_\ell^*, \boldsymbol{\theta}_\ell) &= \int f_\ell(\boldsymbol{y}^* \mid X_\ell^*, \boldsymbol{\beta}_\ell, \sigma^2, \delta) \pi_\ell^N(\boldsymbol{\beta}_\ell \mid \boldsymbol{\theta}_\ell, \sigma^2) d\boldsymbol{\beta}_\ell \\ &= \int f_{N_{n^*}}(\boldsymbol{y}^*;\ X_\ell^*\ \boldsymbol{\beta}_\ell, \delta\ \sigma^2\ I_{n^*}) f_{N_{p_\ell}}(\boldsymbol{\beta}_\ell;\ \mathbf{0}, \sigma^2 \Omega_\ell) d\boldsymbol{\beta}_\ell \\ &= f_{N_{n^*}}(\boldsymbol{y}^*;\ \mathbf{0}, \sigma^2 \Lambda_\ell), \end{aligned}$$

where $\Lambda_\ell = X_\ell^* \Omega_\ell X_\ell^{*T} + \delta I_{n^*}$. Thus, the conditional PEP-Shrinkage prior is

$$\begin{aligned} \pi_\ell^{PEP}(\boldsymbol{\beta}_\ell \mid \sigma^2, \delta, X_\ell^*, \boldsymbol{\theta}_\ell) &= \int \pi_\ell^N(\boldsymbol{\beta}_\ell \mid \boldsymbol{y}^*, \sigma^2, \delta, X_\ell^*, \boldsymbol{\theta}_\ell) m_0^N(\boldsymbol{y}^* \mid \sigma^2, \delta, X_0^*) d\boldsymbol{y}^* \\ &= \int f_{N_{p_\ell}}(\boldsymbol{\beta}_\ell;\ \delta^{-1} W_\ell X_\ell^{*T} \boldsymbol{y}^*, \sigma^2 W_\ell) f_{N_n}(\boldsymbol{y}^*;\ \mathbf{0}, \sigma^2 \Lambda_0) d\boldsymbol{y}^* \\ &= f_{N_{p_\ell}}(\boldsymbol{\beta}_\ell;\ \mathbf{0}, \sigma^2 V_\ell), \end{aligned}$$

where $V_\ell = [W_\ell^{-1} - \delta^{-2} X_\ell^{*T} Z_\ell X_\ell^*]^{-1}$ and $Z_\ell = [\delta^{-2} X_\ell^* W_\ell X_\ell^{*T} + \Lambda_0^{-1}]^{-1}$.

The final PEP-Shrinkage prior is then given by the following hierarchical structure

$$\pi_\ell^{PEP}(\boldsymbol{\beta}_\ell, \boldsymbol{\theta}_\ell, \sigma^2 \mid \delta, X_\ell^*) = \pi_\ell^{PEP}(\boldsymbol{\beta}_\ell \mid \sigma^2, \delta, X_\ell^*, \boldsymbol{\theta}_\ell) \pi(\boldsymbol{\theta}_\ell) \pi^N(\sigma^2).$$

**Example.** Let $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ be a random sample of size $n = 25$ from the normal distribution with mean $\theta$ and variance $\sigma^2 = 1$. We consider the hypothesis $H_0 : \theta = 0$ vs $H_1 : \theta \neq 0$. In order to visualize how the PEP shrinkage prior works, in Figure 1 we plot the PEP-Shrinkage prior (with $\delta = n^* = n = 25$), for the parameter $\theta$, under the alternative hypothesis, with a Horseshoe prior as baseline (PEP-Horseshoe), together with the Horseshoe prior (without using the PEP methodology) and the PEP-Horseshoe with $\delta = 1$ and $n^* = n = 25$. In all cases we assume that the local shrinkage parameter is equal to 1 and the global shrinkage parameter $\lambda \sim C^+(0, 1)$.

From Figure 1 it is distinct that the PEP-Horseshoe, with the recommended values $\delta = n^* = n = 25$, is also unbounded at 0, as the Horseshoe prior is, a property that yields massive shrinkage for zero effects. Higher amount of probability density is accumulating near the origin, in comparison to the Horseshoe prior, a property that yields more shrinkage for the true-zero effects. In addition, PEP-Horseshoe, although it has slightly lighter tails than the original Horseshoe prior, the tails are heavy enough to avoid any essential influence of the posterior distribution of true non-zero effects. Finally, as $\delta$ is getting smaller (see Figure 1 for $\delta = 1$) we observed a much higher amount of probability density near the origin, but also much lighter tails.

In Sections 1 and 2 of the Appendix, we have included theoretical proofs, under the setup of this example, of the behavior of the resulting PEP-Horseshoe prior (with $\delta = n^* = n = 25$), at the origin ($\theta \to 0$) and the tails ($\theta \to \infty$).

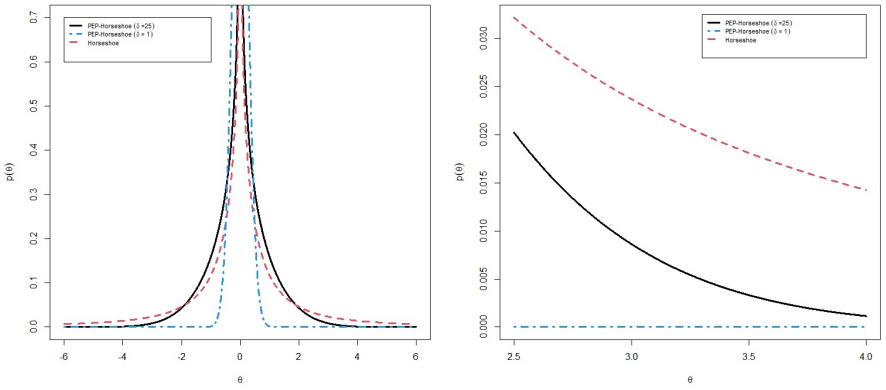


**Figure 1** Density function of the prior distribution under the PEP-Shrinkage methodology, with a Horseshoe prior as a baseline, in comparison with Horseshoe prior without using the PEP methodology, for a normal mean hypothesis testing.

## 2.2 Conditional posterior under the PEP-Shrinkage prior

The posterior distribution, under the PEP-Shrinkage prior, conditional on the shrinkage hyperparameters $\boldsymbol{\theta}_\ell$ of model $M_\ell$, is given by

$$\begin{aligned}\pi_\ell^{PEP}(\boldsymbol{\beta}_\ell, \sigma^2 \mid \boldsymbol{y}, \delta, X_\ell^*, X_\ell, \boldsymbol{\theta}_\ell) &\propto \pi_\ell^{PEP}(\boldsymbol{\beta}_\ell \mid \sigma^2, \delta, X_\ell^*, \boldsymbol{\theta}_\ell)\pi^N(\sigma^2) f_\ell(\boldsymbol{y} \mid X_\ell, \boldsymbol{\beta}_\ell, \sigma^2) \\ &= f_{N_{p_\ell}}(\boldsymbol{\beta}_\ell;\ \mathbf{0}, \sigma^2 V_\ell)\pi^N(\sigma^2) f_{N_n}(\boldsymbol{y};\ X_\ell\ \boldsymbol{\beta}_\ell, \sigma^2\ I_n).\end{aligned}$$

Using the reference prior for $\sigma^2$ (see Section 1), this joint posterior can be written as the product of

$$\pi_\ell^{PEP}(\boldsymbol{\beta}_\ell \mid \boldsymbol{y}, \sigma^2, \delta, X_\ell^*, X_\ell, \boldsymbol{\theta}_\ell) = f_{N_{p_\ell}}(\boldsymbol{\beta}_\ell;\ S_\ell\ X_\ell^T \boldsymbol{y}, \sigma^2 S_\ell)$$

and

$$\pi_\ell^{PEP}(\sigma^2 \mid \boldsymbol{y}, \delta, X_\ell^*, X_\ell, \boldsymbol{\theta}_\ell) = f_{IG}(\sigma^2;\ \alpha_\ell, b_\ell),$$

where $f_{IG}(x;\ \alpha, b)$ is denoting the Inverse Gamma distribution, with shape parameter $\alpha$ and scale parameter $b$. Furthermore, we have set $S_\ell = ({V_\ell}^{-1} + {X_\ell}^T X_\ell)^{-1}$, $\alpha_\ell = \frac{n}{2}$ and $b_\ell = \frac{\boldsymbol{y}^T[I_n + X_\ell V_\ell {X_\ell}^T]^{-1}\boldsymbol{y}}{2}$.

## 2.3 Marginal likelihood under the PEP-Shrinkage prior

The marginal likelihood, of model $M_\ell$, under the PEP-Shrinkage prior, given the shrinkage parameter $\boldsymbol{\theta}_\ell$ is given by

$$\begin{aligned}m_\ell^{PEP}(\boldsymbol{y} \mid \delta, X_\ell^*, X_\ell, \boldsymbol{\theta}_\ell) &= \int \pi_\ell^{PEP}(\boldsymbol{\beta}_\ell \mid \sigma^2, \delta, X_\ell^*, \boldsymbol{\theta}_\ell)\pi^N(\sigma^2) f_\ell(\boldsymbol{y} \mid X_\ell, \boldsymbol{\beta}_\ell, \sigma^2) d\boldsymbol{\beta}_\ell\ d\sigma^2 \\ &\propto |I_n + X_\ell V_\ell {X_\ell}^T|^{-\frac{1}{2}} (\boldsymbol{y}^T[I_n + X_\ell V_\ell {X_\ell}^T]^{-1}\boldsymbol{y})^{-\frac{n}{2}}. \qquad (6)\end{aligned}$$

Therefore in cases where the shrinkage parameters of the baseline prior are fixed (e.g. Ridge g-prior), the above marginal likelihood can be calculated

in closed form. The unknown normalizing constant, in the above expression, comes from the improper prior of the error variance, which is common in all compared models, and therefore we do not face any indeterminacy issues when calculating the Bayes factor.

When the shrinkage parameters are not fixed, the marginal likelihood, according to (6), is given by

$$m_\ell^{PEP}(\boldsymbol{y}) \equiv m_\ell^{PEP}(\boldsymbol{y} \mid \delta, X_\ell^*, X_\ell) = \int m_\ell^{PEP}(\boldsymbol{y} \mid \delta, X_\ell^*, X_\ell, \boldsymbol{\theta}_\ell)\pi(\boldsymbol{\theta}_\ell)d\boldsymbol{\theta}_\ell$$
$$\propto \int \left|I_n + X_\ell V_\ell {X_\ell}^T\right|^{-\frac{1}{2}} (\boldsymbol{y}^T[I_n + X_\ell V_\ell {X_\ell}^T]^{-1}\boldsymbol{y})^{-\frac{n}{2}}\pi(\boldsymbol{\theta}_\ell)d\boldsymbol{\theta}_\ell \quad (7)$$

If $\boldsymbol{\theta}_\ell$ is a single parameter, as in the ridge prior, the above integral can be numerically evaluated in a straightforward manner. Furthermore, in order to search the model space, when full-enumeration is computationally infeasible, $MC^3$ procedures [17] can be applied. If $\boldsymbol{\theta}_\ell$ is a multivariate vector, as in Horseshoe prior for example, we perform an $MC^3$ procedure, conditionally on $\boldsymbol{\theta}_\ell$, as in Algorithm 3 of [10], where each component of $\boldsymbol{\theta}_\ell$ is generated from its full conditional posterior distribution using a Metropolis-Hastings step. A detailed algorithmic procedure is presented in Section 3.

# 3 Computation

In common real life problems with $n < p$, it is almost sure that a full search of the model space containing all possible $2^p$ models, becomes computationally infeasible. Consequently, a full-enumeration of all marginal likelihoods can not be implemented. Alternatively, a procedure based on $MC^3$ has to be used instead, in order to search the model space for the "true" model. Furthermore, the marginal likelihood of model $M_\ell$, which is given by equation (7) cannot be expressed in closed form, for most of the baseline priors of PEP-Shrinkage methodology. In cases where $\boldsymbol{\theta}_\ell$ is one dimensional, an approximation of the integral required in the marginal likelihood can be obtained with the use of numerical methods. As the dimension of $\boldsymbol{\theta}_\ell$ increases, the need of approximate techniques, such as MCMC methods is substantial, in order to estimate the marginal likelihood of a model $M_\ell$ and derive marginal inclusion probabilities. In the following, we provide a detailed description of the adopted $MC^3$ algorithm we introduce for the computation of the relevant quantities (posterior distribution and marginal posterior inclusion probabilities) under the PEP-Shrinkage prior framework.

First of all, we introduce the usual binary vector $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_p)$, which indicates the variables that are included in a model, where $\gamma_j$ (for $j = 1, \ldots, p$) takes the value 1 if covariate $X_j$ belongs to a model and 0 otherwise. In this section we denote the competing models using this $\boldsymbol{\gamma}$ notation.

We set $\delta = n^* = n$, under the PEP prior methodology. We randomly start with model $\boldsymbol{\gamma}^{(0)}$, with local shrinkage parameter $\boldsymbol{\tau}_{\boldsymbol{\gamma}}^{(0)}$ and global shrinkage parameter $\boldsymbol{\lambda}^{(0)}$. We denote by $\boldsymbol{\theta} = (\boldsymbol{\lambda}, \boldsymbol{\tau})$ the set of all shrinkage parameters under the full model formulation (i.e. the model with $\gamma_j = 1$ for all $j = 1, \ldots, p$). Under the MG baseline prior we have that $\boldsymbol{\theta} = g$. Then for any model $\boldsymbol{\gamma}$, other than the full, the local parameters $\boldsymbol{\tau}_{\boldsymbol{\gamma}}$ can be split into two vectors: the ones for the covariates involved in $\boldsymbol{\gamma}$ given by $\boldsymbol{\tau}_{\boldsymbol{\gamma}} = (\tau_1^{\boldsymbol{\gamma}}, \ldots, \tau_{d_{\boldsymbol{\gamma}}}^{\boldsymbol{\gamma}})$ and the ones that are not included denoted by $\boldsymbol{\tau}_{\setminus\gamma}$ of dimension $p - d_{\boldsymbol{\gamma}}$; where $d_{\boldsymbol{\gamma}} = \sum_{j=1}^{p} \gamma_j$ is the dimension of $\boldsymbol{\tau}_{\boldsymbol{\gamma}}$ and it is equal to the number of active covariates included in model $\boldsymbol{\gamma}$. Furthermore, $\tau_j^{\gamma(t)}$ is the $j$-th local shrinkage parameter and $(\boldsymbol{\tau}_{\boldsymbol{\gamma}} \setminus \tau_j^{\boldsymbol{\gamma}})^{(t)}$ is the vector containing all of the shrinkage parameters except the $j$-th, of model $\boldsymbol{\gamma}$, under the $t$ MCMC iteration, for $j = 1, ..., d_{\boldsymbol{\gamma}}$. Local shrinkage parameters are updated sequentially, for every model $\boldsymbol{\gamma}$; we first update parameter $\tau_1^{\boldsymbol{\gamma}}$, then $\tau_2^{\boldsymbol{\gamma}}$ and so on. Therefore, at iteration $t$, when updating $\tau_j^{\boldsymbol{\gamma}}$ (after we have updated the previous $j-1$ parameters) the current vector of all the local shrinkage parameters except the $j$-th, for model $\boldsymbol{\gamma}$, is given by $(\boldsymbol{\tau}_{\boldsymbol{\gamma}} \setminus \tau_j^{\boldsymbol{\gamma}})^{(t)} \equiv (\tau_1^{\gamma(t)}, \ldots, \tau_{j-1}^{\boldsymbol{\gamma}\,(t)}, \tau_{j+1}^{\boldsymbol{\gamma}\,(t-1)}, \ldots, \tau_{d_{\boldsymbol{\gamma}}}^{\boldsymbol{\gamma}\,(t-1)})$.

We denote by $q_1\big(\boldsymbol{\lambda}^{(t)} \mid \boldsymbol{\lambda}^{(t-1)}\big)$ and $q_2\big(\tau_j^{\gamma(t)} \mid \tau_j^{\gamma(t-1)}\big)$ the proposal distributions for the new state of the global and for the $j$-th local shrinkage parameter under model $\boldsymbol{\gamma}$, respectively, given the current state, used in the Metropolis–Hastings step of the algorithm. We use the log-normal distribution for $q_2()$, with median value being the current state and variance chosen such that the acceptance rate is approximately 44%. The same proposal is used for $q_1()$ when $\boldsymbol{\lambda}$ is in all cases univariate, except in Elastic Net, where $\boldsymbol{\lambda}$ is bivariate; in this case $q_1()$ is the product of two independent log-normal distributions of the same type as before.

Finally, we denote by $f_{\boldsymbol{\gamma}}(\boldsymbol{\lambda} \mid \boldsymbol{y}, \boldsymbol{\tau}_{\boldsymbol{\gamma}})$ the posterior distribution of $\boldsymbol{\lambda}$, given $\boldsymbol{\tau}_{\boldsymbol{\gamma}}$, under model $\boldsymbol{\gamma}$, which is

$$f_{\boldsymbol{\gamma}}(\boldsymbol{\lambda} \mid \boldsymbol{y}, \boldsymbol{\tau}_{\boldsymbol{\gamma}}) \propto m_{\boldsymbol{\gamma}}^{PEP}(\boldsymbol{y} \mid \delta, X_{\boldsymbol{\gamma}}^*, X_{\boldsymbol{\gamma}}, \boldsymbol{\lambda}, \boldsymbol{\tau}_{\boldsymbol{\gamma}})\pi(\boldsymbol{\lambda}),$$

where $m_{\boldsymbol{\gamma}}^{PEP}(\boldsymbol{y} \mid \delta, X_{\boldsymbol{\gamma}}^*, X_{\boldsymbol{\gamma}}, \boldsymbol{\lambda}, \boldsymbol{\tau}_{\boldsymbol{\gamma}})$ is given in (6) and $\pi(\boldsymbol{\lambda})$ is the prior distribution of the global shrinkage parameter. Equivalently, we denote by $f_{\boldsymbol{\gamma}}(\tau_j^{\boldsymbol{\gamma}} \mid \boldsymbol{y}, \boldsymbol{\lambda}, (\boldsymbol{\tau}_{\boldsymbol{\gamma}} \setminus \tau_j^{\boldsymbol{\gamma}}))$ the full conditional posterior distribution of $\tau_j^{\boldsymbol{\gamma}}$, under model $\boldsymbol{\gamma}$. This density function is given by

$$f_{\boldsymbol{\gamma}}(\tau_j^{\boldsymbol{\gamma}} \mid \boldsymbol{y}, \boldsymbol{\lambda}, (\boldsymbol{\tau}_{\boldsymbol{\gamma}} \setminus \tau_j^{\boldsymbol{\gamma}})) \propto m_{\boldsymbol{\gamma}}^{PEP}(\boldsymbol{y} \mid \delta, X_{\boldsymbol{\gamma}}^*, X_{\boldsymbol{\gamma}}, \boldsymbol{\lambda}, \boldsymbol{\tau}_{\boldsymbol{\gamma}})\pi(\tau_j^{\boldsymbol{\gamma}} \mid \boldsymbol{\lambda}),$$

where $\pi(\tau_j^{\boldsymbol{\gamma}} \mid \boldsymbol{\lambda})$ is the prior distribution of the $j$-th shrinkage parameter of model $\boldsymbol{\gamma}$ conditional on the global parameters $\boldsymbol{\lambda}$ (in some methods this is equal to $\pi(\tau_j^{\boldsymbol{\gamma}})$). Algorithm 1 summarizes the computational scheme we use.

---

**Algorithm 1** Model search using MC$^3$ conditional on $\boldsymbol{\tau}$ and $\boldsymbol{\lambda}$ with a Metropolis–Hastings step

Start with model $\boldsymbol{\gamma}^{(0)}$, with local shrinkage parameter $\boldsymbol{\tau}_{\boldsymbol{\gamma}}^{(0)}$ and global shrinkage parameter $\boldsymbol{\lambda}^{(0)}$.

For iterations $t = 1, ..., T$:

1. Update of global shrinkage parameter $\boldsymbol{\lambda}$.[*]
   - Set $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(t-1)}$ and $\boldsymbol{\tau}_{\boldsymbol{\gamma}} = \boldsymbol{\tau}_{\boldsymbol{\gamma}}^{(t-1)}$.
   - Generate $\boldsymbol{\lambda}^{(t)}$ from the proposal distribution $q_1\big(\boldsymbol{\lambda}^{(t)} \mid \boldsymbol{\lambda}^{(t-1)}\big)$.
   - Accept the move with probability

$$\alpha_1 = \min\left(1, \frac{f_{\boldsymbol{\gamma}}(\boldsymbol{\lambda}^{(t)} \mid \boldsymbol{y}, \boldsymbol{\tau}_{\boldsymbol{\gamma}}) q(\boldsymbol{\lambda}^{(t-1)} \mid \boldsymbol{\lambda}^{(t)})}{f_{\boldsymbol{\gamma}}(\boldsymbol{\lambda}^{(t-1)} \mid \boldsymbol{y}, \boldsymbol{\tau}_{\boldsymbol{\gamma}}) q(\boldsymbol{\lambda}^{(t)} \mid \boldsymbol{\lambda}^{(t-1)})}\right)$$

     else set $\boldsymbol{\lambda}^{(t)} = \boldsymbol{\lambda}^{(t-1)}$.

2. Update of local shrinkage parameter $\boldsymbol{\tau}$.[†]
   - Set $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(t-1)}$ and $\boldsymbol{\lambda} = \boldsymbol{\lambda}^{(t)}$
   - For $j = 1, \ldots, d_{\boldsymbol{\gamma}}$, selected at random order, repeat
     - Generate $\tau_j^{\gamma(t)}$ from the proposal distribution $q(\tau_j^{\gamma(t)} \mid \tau_j^{\gamma(t-1)})$.
     - Accept the move with probability

$$\alpha_2 = \min\left(1, \frac{f_{\boldsymbol{\gamma}}\big(\tau_j^{\gamma(t)} \mid \boldsymbol{y}, (\boldsymbol{\tau}_{\boldsymbol{\gamma}} \setminus \tau_j^{\gamma})^{(t)}\big) q\big(\tau_j^{\gamma(t-1)} \mid \tau_j^{\gamma(t)}\big)}{f_{\boldsymbol{\gamma}}\big(\tau_j^{\gamma(t-1)} \mid \boldsymbol{y}, (\boldsymbol{\tau}_{\boldsymbol{\gamma}} \setminus \tau_j^{\gamma})^{(t)}\big) q\big(\tau_j^{\gamma(t-1)} \mid \tau_j^{\gamma(t)}\big)}\right)$$

       else set $\tau_j^{\gamma(t)} = \tau_j^{\gamma(t-1)}$.
   - Generate each element $\boldsymbol{\tau}_{\setminus\boldsymbol{\gamma}}^{(t)}$ from the prior distribution $f\big(\boldsymbol{\tau}_{\setminus\boldsymbol{\gamma}} \mid \boldsymbol{\lambda}^{(t)}\big)$.
   - Set $\boldsymbol{\tau}^{(t)} = \big(\boldsymbol{\tau}_{\boldsymbol{\gamma}}^{(t)}, \boldsymbol{\tau}_{\setminus\boldsymbol{\gamma}}^{(t)}\big)$.

3. Bayesian Variable Selection step.
   - Set $\boldsymbol{\gamma} = \boldsymbol{\gamma}^{(t-1)}$
   - For $j = 1, \ldots, p$, selected in random order, repeat
     - Set $\boldsymbol{\gamma}' = \boldsymbol{\gamma}$ and $\boldsymbol{\theta}_{\boldsymbol{\gamma}} = \big(\boldsymbol{\lambda}^{(t)}, \boldsymbol{\tau}_{\boldsymbol{\gamma}}^{(t)}\big)$.
     - Set $\gamma_j' = 1 - \gamma_j$ and $\boldsymbol{\theta}_{\boldsymbol{\gamma}'} = \big(\boldsymbol{\lambda}^{(t)}, \boldsymbol{\tau}_{\boldsymbol{\gamma}'}^{(t)}\big)$.
     - Compute the marginal likelihood

$$m_{\boldsymbol{\gamma}'}^{PEP}(\boldsymbol{y} \mid \boldsymbol{\theta}_{\boldsymbol{\gamma}'}) \equiv m_{\boldsymbol{\gamma}'}^{PEP}(\boldsymbol{y} \mid \delta, X_{\boldsymbol{\gamma}'}^{*}, X_{\boldsymbol{\gamma}'}, \boldsymbol{\theta}_{\boldsymbol{\gamma}'})$$

       of model $\boldsymbol{\gamma}'$, conditional on $\boldsymbol{\theta}_{\boldsymbol{\gamma}'}$ given by equation (6).

---

[*] *For the Beta Prime prior and the Ridge g-prior this step should be skipped since there are no global shrinkage parameters. For the Elastic Net, $\boldsymbol{\lambda}$ is bivariate. For the rest of the methods, a single univariate step is required. For the MG prior we perform the same step with $g$ instead of $\lambda$.*

[†] *This step should be skipped for Ridge prior, the Ridge g-prior and the MG prior, since no local shrinkage parameters exist.*

– Set $\boldsymbol{\gamma} = \boldsymbol{\gamma}'$ (i.e. accept proposed model $\boldsymbol{\gamma}'$) with probability

$$\alpha_3 = \min\left(1, \frac{m_{\boldsymbol{\gamma}'}^{PEP}(\boldsymbol{y} \mid \boldsymbol{\theta}_{\boldsymbol{\gamma}'})\pi(\boldsymbol{\gamma}')}{m_{\boldsymbol{\gamma}}^{PEP}(\boldsymbol{y} \mid \boldsymbol{\theta}_{\boldsymbol{\gamma}})\pi(\boldsymbol{\gamma})}\right),$$

where $\pi(\boldsymbol{\gamma}')$ is the prior probability of model $\boldsymbol{\gamma}'$.

- Set $\boldsymbol{\gamma}^{(t)} = \boldsymbol{\gamma}$.

# 4 Experimental Results

In this section we test the PEP-Shrinkage methodology on simulated and real-life data. We compare the performance of all PEP-Shrinkage priors, as those derived by using the different baseline shrinkage priors given in Table 1. Furthermore, we compare the results obtained by the different PEP methods with those obtained when using the shrinkage priors of Table 1 without applying the PEP methodology.

Our goal is to perform variable selection. To compare the performance of the different techniques, we follow a fully probabilistic approach and we report marginal posterior inclusion probabilities for every predictor. The vast majority of literature on shrinkage priors deals with estimation problems and for the variable selection step an "indirect" approach is usually followed: a variable is dropped if the posterior credible interval of its coefficient include the value of zero, or if the posterior mean for the shrinkage coefficient is below a threshold value (usually 1/2); for more details see [3]. By reporting marginal posterior inclusion probabilities we end up having a model-averaged weight of including a certain predictor in the model, given the observed responses. Therefore, we end up with an variable importance indicator quantifying how relevant a predictor is across all possible models. The median probability (MP) model can then be reported as the "optimal" choice, following the discussion of Section 1.

In Sections 3 and 4 of the Appendix we present results of an additional simulation study, as well as, another real-life example, respectively.

## 4.1 Simulation study

Here we test the PEP-Shrinkage methodology (with $\delta = n = n^*$, $X_\ell^* = X_\ell$ and the reference model to be the null one) on simulated data. As a baseline prior, we use the shrinkage priors listed on Table 1 and compare their results. Moreover, we contrast these results, with the ones obtained by using those shrinkage priors without the PEP-Shrinkage methodology.

We have simulated 100 different samples of length $n = 25$ with $p = 50$ predictors. The values of the explanatory variables have been generated from $N_{50}(\mathbf{0}, \Sigma)$, where a symmetric matrix $\Sigma$ with elements $\Sigma_{i,j} = (0.75)^{|i-j|}$, $i, j = 1, \ldots, 50$. For the predictor effects, we have set $(\beta_1, \beta_2, \beta_{10}) = (2, 0.8, 1.5)$ and for all of the rest, we set to be equal to zero. For the intercept, we have assumed

that $\beta_0 = 0.6$ and we have set $\boldsymbol{y} = \beta_0 1_n + X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_{25}(\mathbf{0}, \sigma^2 I_{25})$, for $\sigma^2 = 1.5$. Finally, we centre the values of the response variable, as well as the columns of the design matrix, on their corresponding means.

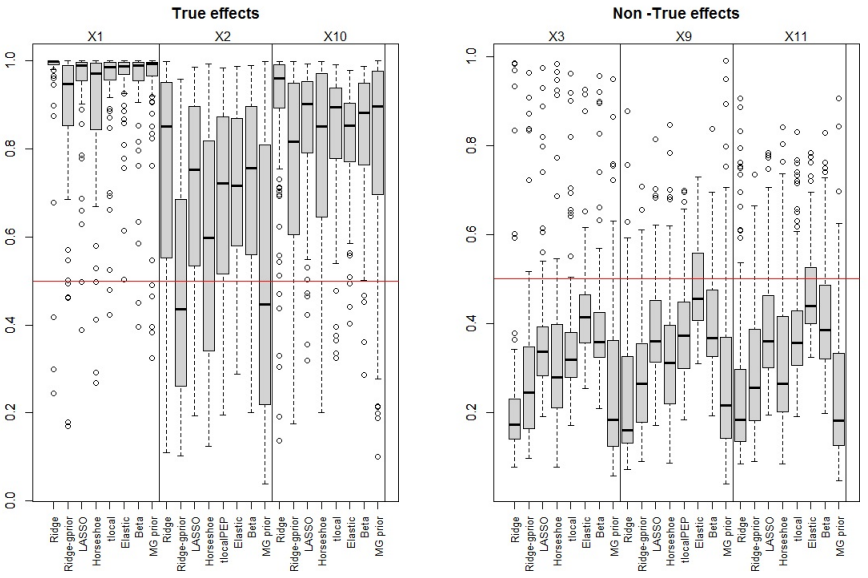


**Figure 2** Boxplots of posterior inclusion probabilities, across 100 simulated datasets, for the true effects - variables $X_1, X_2, X_{10}$ (left) and for some of the non true effects - variables $X_3, X_9, X_{11}$ (right) using the PEP-Shrinkage methodology, for different baseline prior (X-axis).

In Figure 2 (left), we present the boxplots of the marginal posterior inclusion probabilities, for the true non-zero effects, of the 100 different samples, for the seven different PEP-Shrinkage priors. Regarding the two most influential variables, $X_1$ and $X_{10}$, under every baseline prior, we obtained high marginal posterior inclusion probabilities with the majority of the cases to be above 0.5. Furthermore, for these two effects, PEP-Ridge seems to outperform every other PEP-Shrinkage prior, producing high marginal posterior inclusion probabilities with small variability. On the contrary, PEP-Ridge g-prior and PEP-Horseshoe prior give the least satisfactory results, with median marginal posterior inclusion probabilities lower than the ones produced by their competitors but still above 0.5. Also for these two methods we observe the highest variability in the quantity of interest. For predictor $X_2$, the median marginal posterior inclusion probabilities are above 0.5, for all baseline priors, except two. As before, PEP-Ridge gives the most satisfactory results, while PEP-Ridge g-prior and PEP-MG prior produce marginal posterior inclusion probabilities with a median value below 0.5.

For the true-zero effects, we present results only for covariates $X_3, X_9$ and $X_{11}$ (for brevity) in Figure 2 (right). The selected covariates are the ones with the higher correlations with the covariates with non-zero effects. For every selection of baseline prior, the posterior inclusion probabilities are below 0.5. Regardless the baseline prior we choose, only in a small percentage of occasions, the true-zero effects is indicated as important for the model (with posterior inclusion probabilities above 0.5). We notice that PEP-Ridge, followed by the PEP-MG prior, the PEP-Ridge g-prior and the PEP-Horseshoe manage to give, in general, very small marginal posterior inclusion probabilities. For the rest of the zero effects, we get similar results, with the PEP-Elastic Net to produce the highest values of the quantity of interest (but still with median value across all samples below 0.5).

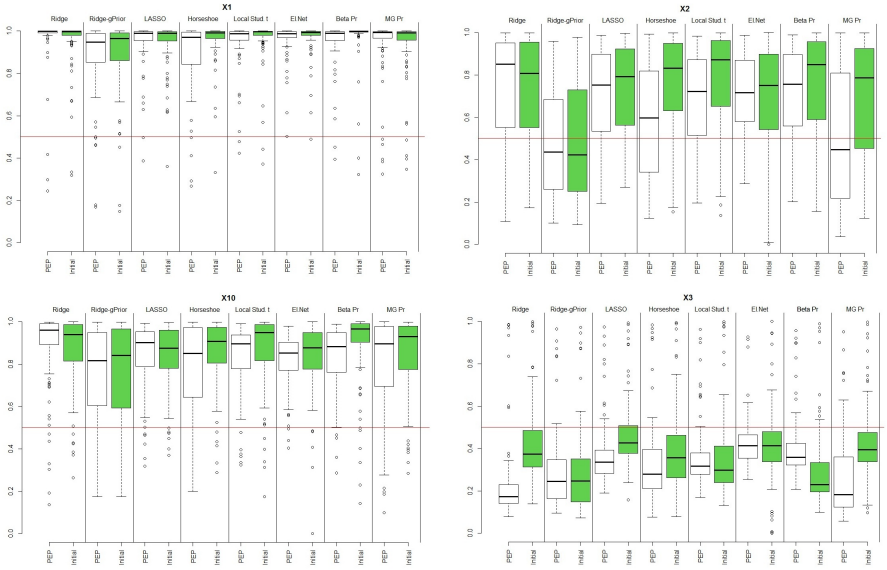


**Figure 3** Boxplots of posterior inclusion probabilities, across 100 simulated datasets, for the true effects - variables $X_1, X_2, X_{10}$ and for one non-true effect - variable $X_3$ using the PEP-Shrinkage priors and the shrinkage priors without the PEP methodology.

In Figure 3, we present boxplots of the posterior inclusion probabilities of the true non-zero effects ($X_1, X_2$ and $X_{10}$), as well as, for variable $X_3$ (true-zero effect). We compare the performance of all of the different PEP-Shrinkage priors mentioned so far, with the performance of the shrinkage priors without applying the PEP methodology. For variable $X_1$ we get similar results, among all pairwise comparisons. All methods (with and without applying the PEP methodology) correctly identify this variable as a true main effect. For variable $X_2$, the performance of the two approaches is again similar in most

cases. When applying the PEP methodology we obtain slightly lower posterior inclusion probabilities, under the Horseshoe, the Local Student's t, the Beta Prime prior and the MG prior. Under the MG prior when applying the PEP methodology, the median value of the posterior inclusion probabilities falls below 0.5, while at the same time the behavior of the MG prior without applying the PEP method is much better. It is also evident that when using the PEP-Ridge g-prior (with and without the PEP methodology) the majority of posterior inclusion probabilities are below 0.5. Finally, regarding variable $X_{10}$, all methods correctly identify it as a true non-zero effect. Under the PEP methodology, we obtain slightly better results under the Ridge prior and the LASSO, and slightly worse results under the Horseshoe, the Local Student's t, the Beta Prime prior and the MG prior. Regarding variable $X_3$ (zero effect), the PEP methodology outperforms its competitors under the Ridge, the LASSO, the Horseshoe and the MG prior, producing more parsimonious answers. Similar are the results, among all pairwise comparisons in the case of the Ridge g-prior and the Elastic Net prior, while only for the Beta Prime prior, results under the PEP methodology are slightly worse (but still with the majority of cases to have values for the quantity of interest below 0.5). The lowest marginal posterior inclusion probabilities are obtained under the PEP-Ridge, the PEP-MG prior and the Ridge g-prior, with and without applying the PEP methodology. Finally under the Ridge and the MG priors we observed the biggest improvement when applying the PEP methodology. Similar are the results for the remaining non-important effects, which are omitted for brevity reasons.

To conclude with, from this simulated study it seems that the best results are produced by the PEP-Ridge prior that manages to retain high posterior inclusion probabilities for the non-zero effects and low posterior inclusion probabilities for the zero effects. The PEP-Ridge g-prior, followed by the PEP-MG prior, are the most parsimonious methods, a property that can be very important for sparse data-sets; see section 4.2.

## 4.2 Real data example

In this section we use a real data-set, concerning the study for the relation of the level of gene TRIM32, with the expression levels of other genes. The motivation for this study is the fact that level of gene TRIM32 causes the Bardet-Biedl syndrome, a genetic condition that affects multiple areas of a patient's system. The data originate from the microarray experiments of mammalian-eye tissue samples (see [22]). The expression levels of different genes are used as explanatory variables and the level of TRIM32 as the response variable. The data-set we use has a sample size of $n = 120$ and $p = 200$ predictors and it is available in the `R` package `flare`. Finally, we centre the values of the response variable, as well as the columns of the design matrix, on their corresponding means.

In order to examine the predictive performance of the PEP-Shrinkage priors (with $\delta = n = n^*$, $X_\ell^* = X_\ell$ and the reference model to be the null one), we first

perform variable selection, where we find the median probability (MP) model, using each time a different baseline prior. A similar task is performed for all of the shrinkage priors without using PEP-Shrinkage methodology. Consequently, we randomly partition the data $N = 30$ times, to the modelling subsample $(M)$ of size $n_M = 90$ and the validation subsample $(V)$ of size $n_V = 30$. For each partition, we generate an MCMC sample ($T = 2000$ iterations) from the model of interest $M_\ell$, using the $M$ subsample and compute the root mean squares error for the data of the $V$ subset, given by:

$$RMSE_\ell = \sqrt{\frac{1}{T}\sum_{t=1}^{T}\frac{1}{n_V}\sum_{i\in V}(y_i - \hat{y}_{i|M_\ell}^{(t)})^2}.$$

In the above equation $y_i$ denotes the response data in the validation subsample and $\hat{y}_{i|M_\ell}^{(t)} = X_{\ell(i)}\boldsymbol{\beta}_\ell^{(t)}$ is the predicted value of $y_i$ under model $M_\ell$. To calculate this predicted value we use the $j$-th row of the design matrix $X_{\ell(i)}$, in the validation subsample, under model $M_\ell$ and the vector of parameters $\boldsymbol{\beta}_\ell^{(t)}$, under model $M_\ell$, for the $t$ iteration of the MCMC sample, using the modelling subsample.

**Table 3** Comparison of the predictive performance of PEP-Shrinkage and shrinkage priors methodology, using the MP model and the full model, in the TRIM32 data set.

| | $d_\ell$* | | *MP* | | *FULL* | |
|---|---|---|---|---|---|---|
| shrinkage prior | PEP | INIT** | PEP | INIT** | PEP | INIT** |
| LASSO | 55 | 90 | 0.0889 | 0.1044 | 0.1136 | 0.1135 |
| Horseshoe | 54 | 52 | 0.0900 | 0.0951 | 0.1189 | 0.1189 |
| Ridge | 54 | 55 | 0.0757 | 0.0837 | 0.1277 | 0.1277 |
| Ridge g-prior | 39 | 41 | 0.0669 | 0.0746 | 0.3108 | 0.2950 |
| Local Student's t | 53 | 73 | 0.0817 | 0.0721 | 0.1632 | 0.1502 |
| Elastic Net | 48 | 44 | 0.0935 | 0.1132 | 0.1373 | 0.1392 |
| Beta Prime | 51 | 49 | 0.0997 | 0.0916 | 0.1032 | 0.1067 |
| MG prior | 47 | 51 | 0.1057 | 0.1103 | 0.1226 | 0.1217 |

*$d_\ell$: Number of the variables accepted in the MP model.*

**INIT: The results given from a shrinkage prior without the use of PEP-Shrinkage methodology.*

In Table 3, we present the results of the mean, across the $N$ random data splits, $RMSE$, for the MP model, as well as, for the full model (including all 200 predictors), with and without using the PEP-Shrinkage methodology. Regarding the MP model, in all pairwise comparisons, except one (Local Student's t), the PEP methodology provided better predictive performance to that offered by the shrinkage priors without the usage of the PEP methodology. When using the Local Student's t shrinkage prior the PEP method produced an MP model with higher mean value of the $RMSE$. But even in this case it is

worthwhile noticing that the MP model under the PEP methodology is much more parsimonious, containing 20 predictors less than the MP model without the PEP approach. When using the LASSO prior, the PEP methodology produces an MP model with considerably lower dimension and predictive error. In all other cases, the dimension of the MP model is similar in the pairwise comparisons. Finally, the model with the best (among all comparisons) predictive performance is the MP model offered by the PEP-Ridge g-prior, both in terms of prediction accuracy and parsimony.

Regarding the full model, results are similar in all pairwise comparisons. This indicates that for estimation purposes, the PEP methodology works in a similar manner as the corresponding shrinkage priors. The full model with the best predictive performance is the one produced by the Beta prime prior (with and without the PEP methodology), while under the Ridge g-prior we get considerably higher prediction errors. On the other hand, as seen in the simulation study, the PEP-Ridge g-prior is the most parsimonious method among the competing ones. When applied as a variable selection method, especially in sparse data-sets (like the one considered here) it manages to drop the vast majority of the non-true effects, producing parsimonious models with good predictive performance.

# 5 Discussion

In this paper we present the model formulation, computation and results on simulated, as well as, real-life data, of an objective Bayesian prior distribution capable of dealing with variable selection problems in normal regression models when the number of observations is smaller than the number of explanatory variables. The proposed PEP-Shrinkage prior combines two approaches: the PEP prior methodology and the shrinkage priors. The resulting prior has a nice interpretation, based on imaginary data, and is compatible across models.

Based on the simulation study, presented here and in the Appendix, the PEP-Shrinkage priors, in the majority of cases, correctly identify the true model. In general, the PEP methodology seems to improve the initial shrinkage prior, by being more parsimonious, a property that is desirable on sparse regression problems. In the two real data examples presented in the main paper and in the Appendix we get slightly better predictive performance under the PEP-Shrinkage priors in almost all of the cases, when using the median probability model.

As a general conclusion, from the studies presented here and in the Appendix, it seems that the PEP-Ridge prior works better than each competitors, while for very sparse data-sets the PEP-Ridge g-prior manages to produce more parsimonious models with very good predictive performance. The PEP-Ridge g-prior is designed in a way that combines the good properties of the g-prior (via the mechanism of PEP prior and imaginary data) and of the ridge regression approach. The latter will help the posterior distribution of the regression coefficients to stabilize in cases of collinearity or when

facing situations with $p > n$. This explains why this prior generally works satisfactory in very sparse data-sets. Finally, for non-sparse data-sets, PEP-Local Student's t, under the MP model, provides better predictive performance.

There are several directions of future extensions. The main aim is to create a unified approach; i.e. a new class of PEP-Shrinkage priors, that includes all the cases mentioned in this paper. To achieve this goal our aim is to write the PEP-Shrinkage prior as a scale mixture of normal distribution, with the mixing distribution denoting the different baseline prior distributions used. This representation will offer several advantages: faster evaluation of posterior distributions and Bayes factors, under all approaches considered, as well as, computational tractability. The performance of this new class of shrinkage prior distributions then have to be assessed in relation to: a) computational efficiency, b) frequentist assessment, especially in terms of the speed of concentration of the posterior parameter distribution, or functional thereof, to the true value, and in terms of coverage of credible sets, c) ease of interpretation, d) default set of tuning hyperparameters in scientific applications. Moreover, a very important aspect is to check mathematical properties of the new class of prior distributions.

Another interesting topic for future research is to study the effect of the size of the imaginary data on the posterior results. In the same manner, we can study the sensitivity of the PEP methodology on the selection of different values of $\delta$, or even set a hyper-prior distribution for this parameter, as in [9]. Computational efficiency could be also improved, possibly with the use of an EM algorithm. Additional future extensions of our PEP-Shrinkage method may include the implementation in generalized linear models, where computation is more demanding.

## Acknowledgement

This work has received funding from the Research Committee of the National Technical University of Athens in Greece ($\Pi.E.B.E.$ 2020 Scheme).

## References

[1] Bai, R. and Ghosh, M.: On the Beta Prime Prior for Scale Parameters in High-Dimensional Bayesian Regression Models. Statistica Sinica. **31**, 843–865 (2021)

[2] Barbieri, M. and Berger, J.: Optimal predictive model selection. The Annals of Statistics. **32**, 870–897 (2004)

[3] Carvalho, C.M., Polson, N.G. and Scott, J.G.: The horseshoe estimator for sparse signals. Biometrika. **97**, 465–480 (2010)

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

[5] Consonni, G., Fouskakis, D., Liseo, B. and Ntzoufras, I.: Prior Distributions for Objective Bayesian Analysis. Bayesian Analysis. **13**, 627–679 (2018)

[6] Datta, J. and Ghosh, J.K.: Asymptotic Properties of Bayes Risk for the Horseshoe Prior. Bayesian Analysis. **8**, 111 – 132 (2013)

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

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

[9] Fouskakis, D., Ntzoufras I. and Perrakis K.: Power-expected-posterior priors in generalized linear models. Bayesian Analysis. **13**, 721–748 (2018)

[10] Fouskakis, D. and Ntzoufras, I.: Power-Expected-Posterior Priors as Mixtures of g-Priors in Normal Linear Models. Bayesian Analysis (accepted) (2021)

[11] George, E. and McCulloch, R.: Variable selection via Gibbs sampling. Journal of the American Statistical Association. **88**, 881–889 (1993)

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

[13] Hsiang, T. C.: A Bayesian view on ridge regression. The Statistician. **24**, 267–268 (1975)

[14] Jeffreys, H.: Theory of Probability. 3rd Edition, Clarendon Press, Oxford (1961)

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

[16] Kyung, M., Gill, J., Ghosh, M., and Casella, G.: Penalized regression, standard errors, and Bayesian lassos. Bayesian Analysis. **5**, 369 – 411 (2010)

[17] Madigan, D. and York, J.: Bayesian Graphical Models for Discrete Data. International Statistical Review. **63**, 215–232 (1995)

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

[19] Park, T. and Casella, G.: The Bayesian lasso. Journal of the American Statistical Association. **103**, 681–687 (2008)

[20] Pèrez, J.M. and Berger, J.O.: Expected - posterior prior distributions for model selection. Biometrika. **89**, 491–511 (2002)

[21] Polson, G. and Scott, J.: On the half-Cauchy prior for a global scale parameter. Bayesian Analysis. **7**, 887–902 (2011)

[22] Scheetz, T. E., Kim, K. Y., Swiderski, R. E., Philp, A. R., Braun, T. A., Knudtson, K. L., Dorrance, A. M., DiBona, G. F., Huang, J., Casavant, T. L., Sheffield, V. C. and Stone, E. M.: Regulation of gene expression in the mammalian eye and its relevance to eye disease. Proceedings of the National Academy of Sciences of the United States of America, **103**, 14429–14434 (2006)

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

[24] Spiegelhalter, D.J., Abrams, K.R., Myles, J.P.: Bayesian Approaches to Clinical Trials and Health-Care Evaluation. Wiley, Chichester (2004)

[25] Tipping, M.E.: Sparse Bayesian Learning and the Relevance Vector Machine. Journal of Machine Learning. **1**, 211–244 (2001)

# Appendix to "A Comparison of Power–Expected–Posterior Priors in Shrinkage Regression"

G. Tzoumerkas, D. Fouskakis and I. Ntzoufras

Contributing authors: tzoumg@mail.ntua.gr; fouskakis@math.ntua.gr; ntzoufras@aueb.gr;

## 1 Proof of the behavior of PEP-Horseshoe prior at the origin, under the setup of the Example of Section 2.1

Under the alternative hypothesis, we have that

$$f(\boldsymbol{y}^* \mid \theta, \delta) = f_{N_{n^*}}(\boldsymbol{y}^*;\ \theta 1_{n^*}, \delta I_{n^*}),$$

while under the null hypothesis

$$m_0^N(\boldsymbol{y}^* \mid \delta) = f_{N_{n^*}}(\boldsymbol{y}^*;\ \mathbf{0}, \delta I_{n^*}).$$

We have also assumed, under the alternative hypothesis, that $\pi^N(\theta \mid \lambda) = f_N(\theta;\ 0, \lambda^2)$ and $\pi(\lambda) = \frac{2}{\pi}\frac{1}{1+\lambda^2}$ (i.e. $\lambda \sim C^+(0,1)$). For the following we also assume that $\delta = n^* = n$.

Under the alternative hypothesis, the posterior distribution of $\theta$, conditional on $\lambda$, using the imaginary data and the baseline prior is

$$\begin{aligned} f(\theta \mid \boldsymbol{y}^*, \delta = n, \lambda) &\propto f(\boldsymbol{y}^* \mid \theta, \delta = n)\pi^N(\theta \mid \lambda) \\ &= f_{N_n}(\boldsymbol{y}^*;\ \theta 1_n, nI_n) f_N(\theta;\ 0, \lambda^2) \\ &= f_N\left(\theta;\ \frac{\lambda^2}{n(1+\lambda^2)} 1_n^T \boldsymbol{y}^*, \frac{\lambda^2}{1+\lambda^2}\right). \end{aligned}$$

Then, under the alternative hypothesis, the conditional (on $\lambda$) PEP-Horseshoe prior is given by

$$\begin{aligned}\pi^{PEP}(\theta \mid \delta = n, \lambda) &= \int f(\theta \mid \boldsymbol{y}^*, \delta = n, \lambda) m_0^N(\boldsymbol{y}^* \mid \delta = n) d\boldsymbol{y}^* \\ &= \int f_N(\theta;\ \frac{\lambda^2}{n(1+\lambda^2)} 1_n^T \boldsymbol{y}^*, \frac{\lambda^2}{1+\lambda^2}) f_{N_n}(\boldsymbol{y}^*;\ \mathbf{0}, nI_n) d\boldsymbol{y}^* \\ &= f_N\left(\theta;\ 0, \frac{2\lambda^4+\lambda^2}{(\lambda^2+1)^2}\right).\end{aligned}$$

Finally, under the alternative hypothesis, the PEP-Horseshoe prior is

$$\begin{aligned}\pi^{PEP}(\theta) \equiv \pi(\theta \mid \delta = n) &= \int_0^\infty \pi^{PEP}(\theta \mid \delta = n, \lambda)\pi(\lambda) d\lambda \\ &= \int_0^\infty f_N\left(\theta;\ 0, \frac{2\lambda^4+\lambda^2}{(\lambda^2+1)^2}\right)\pi(\lambda) d\lambda \\ &= \frac{\sqrt{2}}{\pi^{3/2}} \int_0^\infty \frac{\lambda^2+1}{\lambda\sqrt{2\lambda^2+1}} \frac{1}{1+\lambda^2} exp\left(-\frac{\theta^2}{2}\frac{(\lambda^2+1)^2}{2\lambda^4+\lambda^2}\right) d\lambda \\ &= \frac{\sqrt{2}}{\pi^{3/2}} \int_0^\infty \frac{1}{\lambda\sqrt{2\lambda^2+1}} exp\left(-\frac{\theta^2}{2}\frac{(\lambda^2+1)^2}{2\lambda^4+\lambda^2}\right) d\lambda.\end{aligned}$$

Let $z = 1/\lambda^2$. Then

$$\pi^{PEP}(\theta) = (2\pi^3)^{-1/2} \int_0^\infty \frac{1}{\sqrt{z(z+2)}} exp\left(-\frac{\theta^2}{2}\frac{(z+1)^2}{z+2}\right) dz. \tag{1}$$

Notice that for $z > 0$,

$$\begin{aligned}(z+1)^2 &= z^2+2z+1 > z^2+2z = z(z+2) \Rightarrow (z+1) > \sqrt{z(z+2)} \\ &\Rightarrow \frac{1}{z+1} < \frac{1}{\sqrt{z(z+2)}}.\end{aligned}$$

Furthermore,

$$\begin{aligned}z+1 < z+2 &\Rightarrow -\frac{1}{z+1} < -\frac{1}{z+2} \Rightarrow -\frac{(z+1)^2}{z+1} < -\frac{(z+1)^2}{z+2} \\ &\Rightarrow -(z+1) < -\frac{(z+1)^2}{z+2} \\ &\Rightarrow exp\left(-\frac{\theta^2}{2}(z+1)\right) < exp\left(-\frac{\theta^2}{2}\frac{(z+1)^2}{z+2}\right).\end{aligned}$$

By applying these inequalities to the last integral in equation (1) we have that

$$\pi^{PEP}(\theta) > (2\pi^3)^{-1/2} \int_0^\infty \frac{1}{z+1} exp\left(-\frac{\theta^2}{2}(z+1)\right) dz.$$

Let $u = 1 + z$. Then

$$\pi^{PEP}(\theta) > (2\pi^3)^{-1/2} \int_1^\infty \frac{1}{u} exp\left(-\frac{\theta^2}{2}u\right) du = (2\pi^3)^{-1/2} E_1(\frac{\theta^2}{2}),$$

where $E_1(\cdot)$ is the exponential integral function, which satisfies that $E_1(t) > \frac{exp(-t)}{2} log(1 + \frac{2}{t})$. So we have that

$$\pi^{PEP}(\theta) > (2\pi^3)^{-1/2} \frac{exp(-\theta^2/2)}{2} log(1 + \frac{4}{\theta^2}).$$

Notice that $\lim_{\theta \to 0} exp(-\theta^2/2) log(1 + \frac{4}{\theta^2}) = +\infty$, so from the last inequality we have that

$$\lim_{\theta \to 0} \pi^{PEP}(\theta) = +\infty.$$

## 2 Proof of the behavior of PEP-Horseshoe prior at the tails, under the setup of the Example of Section 2.1

First notice the following two inequalities, for $z > 0$:

$$(z+1)^2 > z(z+2) \Rightarrow \frac{(z+1)^2}{z+2} > z \Rightarrow exp\left(-\frac{\theta^2}{2}\frac{(z+1)^2}{z+2}\right) < exp\left(-\frac{\theta^2}{2}z\right),$$

$$z + 2 > 1 \Rightarrow \frac{1}{\sqrt{z+2}} < 1.$$

From applying them in equation (1), we find that

$$\pi^{PEP}(\theta) < (2\pi^3)^{-1/2} \int_0^\infty z^{-1/2} exp\left(-\frac{\theta^2}{2}z\right) dz. \tag{2}$$

Notice now that, the p.d.f. of a generalized gamma distribution, with parameters $a, d, p > 0$ is given by

$$f(x) = \frac{p/a^d}{\Gamma(d/p)} x^{d-1} exp\left(-\frac{x^p}{a^p}\right),$$

where $x > 0$. For $d = 1/2$, $p = 1$ and $a = \frac{2}{\theta^2}$ we have that

$$\int_0^\infty \frac{(\theta^2/2)^{1/2}}{\Gamma(1/2)} z^{1/2-1} exp\left(-\frac{\theta^2}{2}z\right) dz = 1$$

and thus

$$\int_0^\infty z^{-1/2} exp\left(-\frac{\theta^2}{2}z\right) dz = \frac{\Gamma(1/2)\sqrt{2}}{\mid \theta \mid},$$

which approaches 0 as $\theta \to \pm\infty$. From (2) it is clear that

$$\lim_{\theta \to -\infty} \pi^{PEP}(\theta) = 0 \text{ and } \lim_{\theta \to +\infty} \pi^{PEP}(\theta) = 0.$$

# 3 Second simulation study

Here we test again the PEP-Shrinkage methodology, on simulated data. The main difference with the simulation study found in the main paper is that in this example, the values of the explanatory variables have been generated independently.

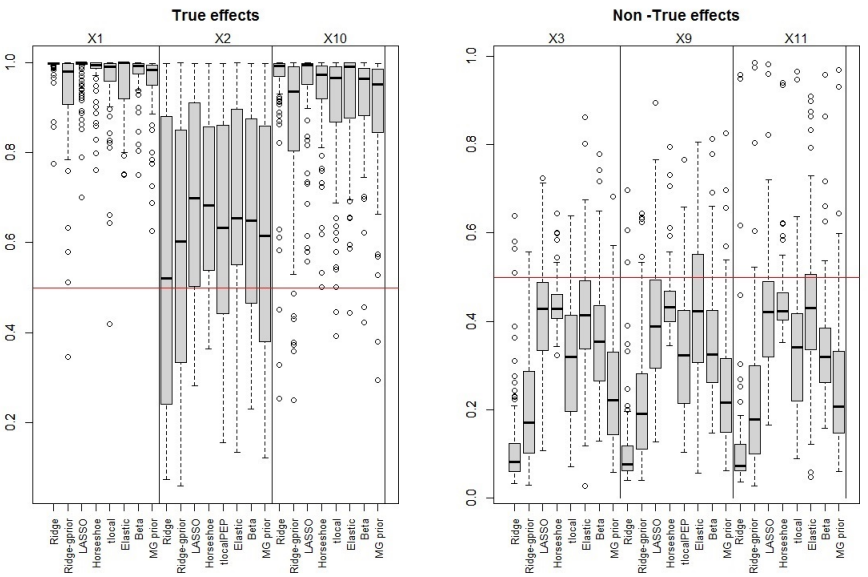


**Figure 1** Boxplots of posterior inclusion probabilities, across 100 simulated datasets, for the true effects - variables $X_1, X_2, X_{10}$ (left) and for some of the non true effects - variables $X_3, X_9, X_{11}$ (right) using the PEP-Shrinkage methodology, for different baseline prior (X-axis).

We have simulated 100 different samples of length $n = 25$ with $p = 50$ predictors. The values of the explanatory variables have been generated from

$N_{50}(\mathbf{0}, \Sigma)$, with $\Sigma = I_p$. For the predictor effects, we have set $(\beta_1, \beta_2, \beta_{10}) = (2, 0.8, 1.5)$ and for all of the rest, we set to be equal to zero. For the intercept, we have assumed that $\beta_0 = 0.6$ and we have set $\boldsymbol{y} = \beta_0 1_n + X\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim N_{25}(\mathbf{0}, \sigma^2 I_{25})$, for $\sigma^2 = 1.5$. Finally, we centre the values of the response variable, as well as the columns of the design matrix, on their corresponding means.

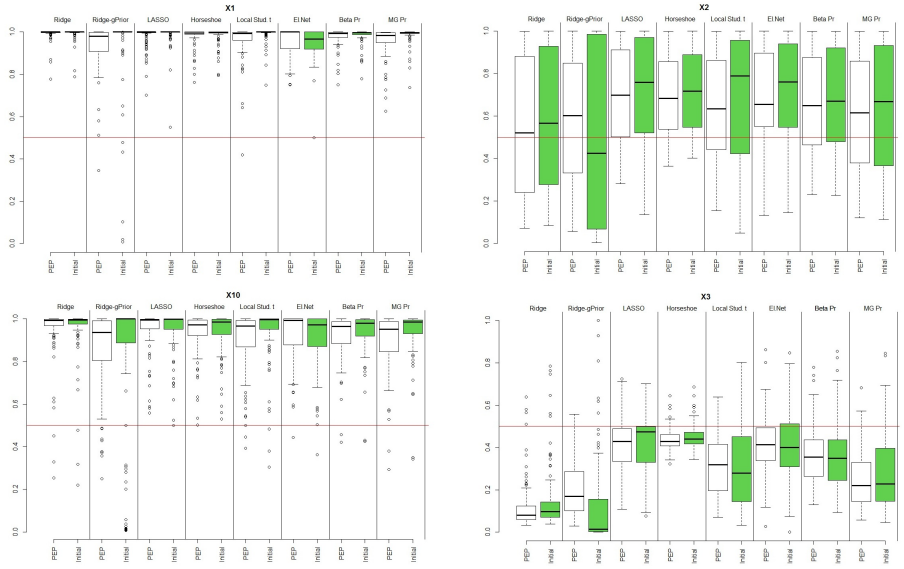


**Figure 2** Boxplots of posterior inclusion probabilities, across 100 simulated datasets, for the true effects - variables $X_1, X_2, X_{10}$ and for one non-true effect - variable $X_3$ using the PEP-Shrinkage priors and the shrinkage priors without the PEP methodology.

In Figure 1, we present the boxplots of the marginal posterior inclusion probabilities, for the true non-zero effects and for three of the non-true effects, of the 100 different samples, for the eight different PEP-Shrinkage priors. We observe quite similar results regarding the behavior of the choice of a shrinkage prior, as a baseline prior, in PEP methodology, as in the simulated study of the main paper. Regarding the two most influential variables ($X_1$ and $X_{10}$), in this simulated study, we get in general posterior inclusion probabilities closer to 1 (in comparison to the simulation study in the main paper), with the two depended shrinkage (baseline) priors (Ridge g-prior and MG) producing the least satisfactory results. For the variable $X_2$, all methods produces median marginal posterior inclusion probabilities above 0.5, in contrast with the first simulation study in the main paper, that under the PEP-Ridge g-prior and the PEP-MG prior the median marginal posterior inclusion probabilities were below 0.5.

For the non-true effects, presented in Figure 1, we observe, very similar results regarding the behavior of the choice of a shrinkage prior, as the ones in the simulation study presented in the main paper. The PEP-Ridge again produces the lowest values and shows an even better behavior compared to the one showed in the first simulation study. The two depended shrinkage (baseline) priors (Ridge g-prior and MG) produces also low values and outperform the other competitors.

In Figure 2, we compare PEP Shrinkage priors with the initial priors without the use of PEP methodology. We detect similar results regarding the behavior of the priors under investigation, as the simulated example in the main paper. For $X_2$, in this simulation study, the PEP-Ridge g-prior gives median values above 0.5, while the same prior without applying the PEP methodology produces median values below 0.5. For the non-true effect (variable $X_3$) under the PEP methodology we get less variability in general, and only under the Ridge g-prior and the Local Student's t we get slightly higher median values, but always below 0.5.

# 4 Second real data example

In this second real data-set example, we are interested in the disease progression for 442 diabetes patients. The explanatory variables used are the age, the sex, the body mass index, the average blood pressure and six blood serum measurements collected from every patient. The diabetes data-set [1] has a sample size is $n = 442$ and $p = 10$ predictors; thus we are facing a non-sparse regression problem with $p < n$. The data are available in the `R` package `care`, under the name `efron2004`, and have been standardized such that the means of all variables are zero, and all variances are equal to one.

In order to examine the predictive performance of the PEP-Shrinkage priors (with $\delta = n = n^*$, $X_\ell^* = X_\ell$ and the reference model to be the null one), we use same techniques as the ones under the real data example in the main paper. We first perform variable selection, where we find the median probability (MP) model, using each time a different baseline prior and perform the same task also for all of the shrinkage priors without using PEP methodology. Then, we randomly partition the data $N = 30$ times, to the modeling subsample $(M)$ of size $n_M = 300$ and the validation subsample $(V)$ of size $n_V = 142$. For each partition, we generate an MCMC sample ($T = 2000$ iterations) from the model of interest $M_\ell$, using the $M$ subsample and compute the root mean squares error $(RMSE)$ for the data of the $V$ subset.

In Table 1, we present the results of the mean $RMSE$, across the $N$ random data splits, for the MP model, as well as, for the full model, with and without using the PEP-Shrinkage methodology. Under the MP model, we get lower values when applying the PEP methodology in all pairwise comparisons, except the Horseshoe and the Beta Prime prior. Regarding the full model, results are similar in all pairwise comparisons and thus, like in the real life example of the main paper, this indicates that for estimation purposes, the PEP methodology

**Table 1** Comparison of the predictive performance of PEP-Shrinkage and shrinkage priors methodology, using the MP model and the Full model, in the diabetes data set.

| | $d_\ell$* | | *MP* | | *FULL* | |
|---|---|---|---|---|---|---|
| shrinkage prior | PEP | INIT** | PEP | INIT** | PEP | INIT** |
| LASSO | 6 | 5 | 0.7144 | 0.7459 | 0.7513 | 0.7527 |
| Horseshoe | 6 | 7 | 0.7291 | 0.7023 | 0.7171 | 0.7088 |
| Ridge | 4 | 5 | 0.7137 | 0.7169 | 0.7368 | 0.7303 |
| Ridge g-prior | 6 | 6 | 0.7028 | 0.7232 | 0.7102 | 0.7238 |
| Local Student's t | 5 | 3 | 0.6985 | 0.7331 | 0.7204 | 0.7368 |
| Elastic Net | 4 | 4 | 0.7231 | 0.7350 | 0.7512 | 0.7553 |
| Beta Prime | 6 | 6 | 0.7261 | 0.7055 | 0.7016 | 0.7329 |
| MG prior | 6 | 6 | 0.7432 | 0.7512 | 0.7693 | 0.7774 |

*$d_\ell$: *Number of the variables accepted in the MP model.*

**INIT: *The results given from a shrinkage prior without the use of PEP-Shrinkage methodology.*

works in a similar manner as the corresponding shrinkage priors. The full model with the best predictive performance is the one produced by the Beta prime prior with the PEP methodology. Finally, the model with the best (among all comparisons) predictive performance is the MP model offered by the PEP - Local Student's t prior.

# References

[1] Efron, B., Hastie, T., Johnstone, I., and Tibshirani, R.: Least angle regression. Annals of Statistics, **32**, 407–499 (2004)