Bayesian Analysis (2019) **TBA**, Number TBA, pp. 1–26

A Loss-Based Prior for Variable Selection in Linear Regression Methods

Cristiano Villa* and Jeong Eun Lee†

Abstract. In this work we propose a novel model prior for variable selection in linear regression. The idea is to determine the prior mass by considering the *worth* of each of the regression models, given the number of possible covariates under consideration. The *worth* of a model consists of the information loss and the loss due to model complexity. While the information loss is determined objectively, the loss expression due to model complexity is flexible and, the penalty on model size can be even customized to include some prior knowledge. Some versions of the loss-based prior are proposed and compared empirically. Through simulation studies and real data analyses, we compare the proposed prior to the Scott and Berger prior, for noninformative scenarios, and with the Beta-Binomial prior, for informative scenarios.

Keywords: Bayesian variable selection, linear regression, loss functions, objective priors.

1 Introduction

In this paper, we propose a method to derive model prior probabilities for variable selection problems in linear regression. The obtained prior, in its general form, is designed to penalise complex models and, therefore, to favour sparsity. We focus on the general case in which the total number of covariates, d, is smaller than the number of observations n. The prior we propose is based on losses and it is compared with existing options, including the Beta-Binomial prior (George and McCulloch, 1993), where a beta prior is defined over the (unknown) covariate inclusion probability, and a particular case of it known as the Scott and Berger prior (Scott and Berger, 2010). The priors are compared by using simulated data as well as real data sets: the Hald data (Woods et al., 1932) and the human micro-array gene expression data in colon cancer patients (Calon et al., 2012).

Variable selection problems, in the Bayesian framework, are in line with any other inferential procedure. That is, a posterior distribution for the space of models is obtained in order to represent the posterior uncertainty about the true regression model (see Gelman et al. (2004)). There may be instances where the above is not appropriate, for example if there are models with a negligible posterior probability, in which case a subset of all the possible regression models can be considered. With a prior distribution on the space of models, representing the model uncertainty related to variable selection, one way to proceed is by using Bayesian model averaging (Hoeting et al., 1999). When the

*University of Kent, School of Mathematics, Statistics and Actuarial Sciences, Canterbury, United Kingdom, C.Villa-88@kent.ac.uk

†University of Auckland, Department of Statistics, Auckland, New Zealand, kate.lee@auckland.ac.nz

 https://doi.org/10.1214/19-BA1162

model posterior distribution tends to be spread across many of the possible regression models, and when prediction is an important part of the statistical analysis, Raftery et al. (1997) show that Bayesian model averaging performs better than choosing the regression model with the highest posterior probability. Also, although one may decide to explore a different route than model averaging, Barbieri and Berger (2004) show that the median probability model (MPM), under certain conditions, has a predictive power at least as good as the one of the highest probability model (HPM). The MPM derives from an equivalent "answer" to the above problem obtained by estimating the marginal posterior probability that each variable has, independently from the others, of being included in the regression model.

An important component of the Bayesian variable selection approach is the definition of the prior for the regression coefficients, including the intercept, and the regression variance. In fact, we can comfortably say that the majority of literature related to variable selection is focused on the identification of appropriate prior distributions for the model specific parameters. Keeping in mind the importance of parameter specific priors, in this paper we will focus on the model prior distribution only, referring to the specific literature on the subject. See, for example, Bayarri et al. (2012) and the references therein.

The underlying idea of the proposed prior is that, given the total number of covariates, we associate a *worth* to each of the 2^d possible linear regression model, depending on the fact that model has been chosen to be part of the problem (Villa and Walker, 2015). The *worth* is determined by measuring what is lost if the model were to be excluded and it is the true model. The fact that a regression model has been chosen to be part of the model space (i) conveys information and (ii) induces complexity; as such, we can measure the loss in information carried by a model and the loss due to its complexity. These losses will then form the basis to determine the *worth* of the model and hence, the model prior probability. We will show that the loss in information one would incur in choosing the "wrong" model, due to the nature of the problem, is always zero, leaving the loss due to complexity only to form the basis for the prior. Furthermore, we will show that the prior exhibits, in its general form, an exponential decay which depends on a constant (namely, c); by calibrating the constant one may induce different degrees of sparsity in the prior. We will discuss some guidelines on how c can be calibrated.

The paper is organised as follows. In Section 2 we define the notation used throughout the paper and formalise the problem of variable selection for linear regression models in a Bayesian framework. In Section 3 we discuss the current model priors for variable selection (the Beta-Binomial prior and its particular case the Scott and Berger prior) and we present the proposed prior based on losses. The results of simulation studies are provided in Section 4. In the section we examine the performance of the considered priors on the basis of the frequentist results of the corresponding model size posterior distributions. Section 5 reports the analysis results using real data sets widely discussed in literature. The final Section 6 concludes and provides some discussion points on the proposed prior and its comparison with the other priors.

2 Notation and problem specification

Given the vector $\mathbf{y}$ of n responses, the design matrix $\mathbf{X}$ of size $n \times d$, an intercept α and a vector of coefficients $\boldsymbol{\beta}$ of dimension d, the response outcome y_i is expressed as

$$y_i = \alpha + \sum_{j=1}^{d} \beta_j X_{ij} + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

where $\varepsilon_i \sim N(0, 1/\phi)$ are the i.i.d. normally distributed errors with unknown variance $1/\phi$. We assume that the number of observations n is larger than the number of covariates (i.e. $n > d$), and the design matrix is of full rank. The variable selection problem can be seen as identifying which of the possible d covariates has impact on $\mathbf{y}$. In other words, we aim to identify which of the regression parameters β_js are different from zero. Let us consider the binary vector $\boldsymbol{\gamma}$, where the j-th element is zero if $\beta_j = 0$ and one if $\beta_j \neq 0$. Then, the generic Bayesian regression model is indicated by

$$M_{\boldsymbol{\gamma}} = \{f(\mathbf{y}|\alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi); \pi_{\boldsymbol{\gamma}}(\alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi)\}, \tag{2}$$

where

$$f(\mathbf{y}|\alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi) = N(\mathbf{y}|\alpha + \mathbf{X}_{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\boldsymbol{\gamma}}, \mathrm{I}/\phi),$$

and $\pi_{\boldsymbol{\gamma}}(\alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi)$ represents the prior distribution for the parameters of the model, the so-called *model-specific parameter prior*. Note that $|\boldsymbol{\gamma}|$ (the number of ones in $\boldsymbol{\gamma}$) indicates the number of covariates included in the model $M_{\boldsymbol{\gamma}}$. There are 2^d possible regression models and each one of them identified by $\boldsymbol{\gamma}$.

In the Bayesian framework, inference about model uncertainty is based on the model posterior probability

$$p(M_{\boldsymbol{\gamma}}|\mathbf{y}) \propto m(\mathbf{y}|M_{\boldsymbol{\gamma}})p(M_{\boldsymbol{\gamma}}),$$

where $p(M_{\boldsymbol{\gamma}})$ is the prior probability for model $M_{\boldsymbol{\gamma}}$ and

$$m(\mathbf{y}|M_{\boldsymbol{\gamma}}) = \int f(\mathbf{y}|\alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi)\pi_{\boldsymbol{\gamma}}(\alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi)\, d\alpha\, d\boldsymbol{\beta}_{\boldsymbol{\gamma}}\, d\phi, \tag{3}$$

is the marginal likelihood of the observations under model $M_{\boldsymbol{\gamma}}$. The model posterior distribution can then be used to either choose a specific regression model or perform model averaging.

The number of possible regression models grows exponentially with d. When d is large the model posterior probabilities are often small for most models, and posterior inclusion probabilities could give a better idea of the posterior uncertainty in comparison to model posterior probabilities. The posterior inclusion probability of the j-th covariate is defined as

$$\omega_j = \Pr(\gamma_j \neq 0|\mathbf{y}) = \sum_{\boldsymbol{\gamma}} p(M_{\boldsymbol{\gamma}}|\mathbf{y}) \cdot 1_{\gamma_j = 1}.$$

Prior and posterior inclusion probabilities originate from the common idea in Bayesian variable selection to consider variable inclusions as exchangeable Bernoulli trials, with $\omega \in [0, 1]$, implying

$$p(M_{\boldsymbol{\gamma}}|\omega) = \omega^{|\boldsymbol{\gamma}|}(1 - \omega)^{d - |\boldsymbol{\gamma}|}, \tag{4}$$

where $|\boldsymbol{\gamma}|$ represents the number of covariates included in the model $M_{\boldsymbol{\gamma}}$.

Model-specific parameter prior Prior choice on model-specific parameters has received much attention, and well-received priors include the Zellner–Siow prior (Zellner and Siow, 1980), the Zellner's g-prior (Zellner, 1986), the mixtures of g-priors (Liang et al., 2008), and the more recent *robust* prior by Bayarri et al. (2012), among others. The robust prior has the advantage of yielding closed-form marginal likelihoods and to not suffer from the information paradox (Liang et al., 2008). The robust prior is defined as

$$\begin{aligned} \pi_{\boldsymbol{\gamma}}(\alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi) &= \pi(\alpha, \phi) \times \pi(\boldsymbol{\beta}_{\boldsymbol{\gamma}}) \\ &= \phi^{1/2} \int_0^{\infty} N_{|\boldsymbol{\gamma}|}(\boldsymbol{\beta}_{\boldsymbol{\gamma}} | 0, g\Sigma_{\boldsymbol{\gamma}}) \pi_{\boldsymbol{\gamma}}(g)\, dg, \end{aligned} \tag{5}$$

where $\Sigma_{\boldsymbol{\gamma}}$ is the covariance matrix of the maximum likelihood estimator of $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$. The distribution of g is given by

$$\pi_{\boldsymbol{\gamma}}(g) = a[\rho_{\boldsymbol{\gamma}}(b+n)]^a (g+b)^{-(a+1)} \cdot 1_{\{g > \rho_{\boldsymbol{\gamma}}(b+n)-b\}},$$

where $a, b > 0$ and $\rho_{\boldsymbol{\gamma}} \geq b/(b+n)$. The prior (5) is called *robust* as its tails behave like the tails of a multivariate t density, therefore less sensible to outliers.

In this paper, as the focus is solely on model priors, most of the analysis are performed by using the same model-specific parameter prior, the robust prior with hyperparameter values $a = 1/2$, $b = 1$ and $\rho_{\boldsymbol{\gamma}} = 1/(d+1)$, as recommended in Bayarri et al. (2012), so differences in the results can be ascribed to differences in the model prior. For large d simulations and the large real data analysis we have used the g-Zellner prior, for computational convenience.

3 Model priors in objective variable selection

We begin this section with a short summary about the existing priors, then describe the loss-based prior in Section 3.1 followed by the choices for the penalty factor in Section 3.2 and the generalized version in Section 3.3.

In principle, the choice of the prior on the model space should incorporate any prior knowledge about the subsets of covariates which should be included in the model. A common way of achieving the above result is to subjectively determine ω in equation (4) (George and McCulloch, 1993). Furthermore, by subjectively fixing ω, so to represent the proportion of covariates that one believes should be included in the model, will induce multiplicity correction (Scott and Berger, 2010). To address this issue, Cui and George (2008) suggested a beta prior distribution on ω

$$p(M_{\boldsymbol{\gamma}}) = \int_0^1 p(M_{\boldsymbol{\gamma}} | \omega) \pi(\omega)\, d\omega = \frac{B(a + |\boldsymbol{\gamma}|, b + d - |\boldsymbol{\gamma}|)}{B(a, b)}, \tag{6}$$

where $B(\cdot)$ is the beta function and a, b are the parameters of the beta prior. Thus, by setting a and b, one could represent prior knowledge about the true proportion of the covariates that should be included in the model.

Alternatively, one may wish to adopt an objective approach and, to the best of our knowledge, the choice of model prior probabilities which convey minimal information is limited to two options: the uniform prior and the Scott and Berger prior.

The model uniform prior is obtained by assigning equal prior mass to each regression model, that is $p(M_{\boldsymbol{\gamma}}) = 1/(2^d)$ for any $\boldsymbol{\gamma}$, and it yields a prior inclusion probability of $\omega = 1/2$. Scott and Berger (2010) discuss the following model prior for variable selection

$$\begin{aligned} p(M_{\boldsymbol{\gamma}}) &= \int_0^1 p(M_{\boldsymbol{\gamma}}|\omega)\pi(\omega)\, d\omega \\ &= \frac{1}{d+1}\binom{d}{|\gamma|}^{-1}. \end{aligned} \tag{7}$$

Their model prior is obtained by assigning a beta prior to ω, with both the hyper-parameters equal to one, and then marginalising over ω. This prior was previously discussed in the literature, see for example Ley and Steel (2009), with the aim of representing prior minimal information. The motivation behind the choice of the model prior by Scott and Berger (2010) lies in its property to correct for multiplicity, which can be seen as an issue when model choice is performed by multiple statistical testing with respect to a reference model (typically the null model or the full model). The Scott and Berger prior induces a marginal prior inclusion probability of $\omega = 1/2$ for each covariate, same as the uniform model prior. However, as thoroughly discussed in Scott and Berger (2010), given that their prior is a function of d, it allows the multiplicity correction.

Both the uniform prior and the Scott and Berger prior assume that the expected number of covariates is $d/2$.

3.1 Model prior based on losses

The model prior based on losses has been introduced by Villa and Walker (2015), where model selection problems involving nested and non-nested models have been discussed. However, in Villa and Walker (2015) model complexity was not considered. The basic idea is that we can assign a *worth* to each model by objectively measuring what is lost if the model is removed from the space of models, and it is the true one. While in Villa and Walker (2015) the *worth* of a model was associated to a measure of loss in information only, in cases of variable selection it is sensible to include a component of loss due to the complexity of the model.

We introduce the idea to assign prior mass to models by means of the following illustration (from Villa and Walker (2015)). Let us consider three trivial models $M_j = \{f_j(x|\theta_j), \pi_j(\theta_j)\}$, for $j = 1, 2, 3$. Here each model represents a single density. We also assume that models M_1 and M_2 (and so densities f_1 and f_2) are very similar, and that the third model M_3 (density f_3) is significantly different from the other two. We do not question the rational behind this scenario set up, we just assume that there is one.

By analysing this scenario in the light of the utility of each model we note that the *worth* of models M_1 or M_2 is less than the one of model M_3. In fact, should we lose either M_1 or M_2, we would still have the remaining one to "represent" that position in the set of all possible models. On the other hand, M_3 would be more valuable, as its removal from the set of choices would lead to bad inference if it turned out to be the true model.

Having identified this approach to assign the mass to each model on the basis of *worth*, we see it takes into consideration the "position" of each model with respect to the others. The quantification of the *worth* comes from a result in Berk (1966) which says that, if the model is misspecified, the posterior distribution asymptotically tends to accumulate at the nearest model in terms of the Kullback–Leibler divergence. Therefore, if we were to remove model M_j from the set of possible models, and it is the true one, the loss we would incur is given by the Kullback–Leibler divergence from it to the nearest of $\{f_i\}$, $i \neq j$. Thus, by defining the Kullback–Leibler divergence between M_j and M_i by $D_{KL}(f_j \| f_i) = \int f_j \log(f_j/f_i)\, df_j$, the loss associated with model M_j would be

$$l(M_j) = l_j = -\min_{j \neq i} D_{KL}(f_j \| f_i). \tag{8}$$

That is, the larger the value of $\min_{j \neq i} D_{KL}(f_j \| f_i)$ the greater the utility (or, equivalently, the smaller the loss) of keeping the model.

If we consider the mass to be put on each model $P(M_j)$, this can be linked to the *worth* of the model via the *self-information* loss function. The *self-information* loss function (also known as the *log-loss* function in machine learning) measures the performance of a probability statement with respect to an outcome. Thus, for every probability assignment $P = \{P(A), A \in \Omega\}$, the *self-information* loss function is defined as

$$l(P, A) = -\log P(A).$$

More details and properties of this particular loss function can be found, for example, in Merhav and Feder (1998). Therefore, for each model M_j we have a measure of the information loss related to its *worth*, given by (8), and related to the *self-information*, given by $-\log P(M_j)$. We then equate the two losses, yielding

$$-\log P(M_j) = -\min_{j \neq i} D_{KL}(f_j \| f_i),$$

equivalently

$$P(M_j) \propto \exp\left\{\min_{j \neq i} D_{KL}(f_j \| f_i)\right\}. \tag{9}$$

In other words, the mass that we assign to each model is proportional to the exponential of the Kullback–Leibler divergence between the model and the nearest one in the set of options.

We can now take the basis of the idea to regression models, that is in a variable selection scenario.

If we consider the regression model $M_{\boldsymbol{\gamma}}$, with the simplified notation $\boldsymbol{\theta}_{\boldsymbol{\gamma}} = (\alpha, \boldsymbol{\beta}_{\boldsymbol{\gamma}}, \phi)$, the loss in information associated to the model, equivalent to equation (8), can be

written as

$$L_I(M_{\boldsymbol{\gamma}}) = -\int_{\boldsymbol{\theta}_{\boldsymbol{\gamma}}} \inf_{\boldsymbol{\theta}_{\boldsymbol{\gamma}'} \neq \boldsymbol{\theta}_{\boldsymbol{\gamma}}} D_{KL}\Big(f(\mathbf{y}|\boldsymbol{\theta}_{\boldsymbol{\gamma}}) \| f(\mathbf{y}|\boldsymbol{\theta}_{\boldsymbol{\gamma}'})\Big) \pi_{\boldsymbol{\gamma}}(\boldsymbol{\theta}_{\boldsymbol{\gamma}})\, d\boldsymbol{\theta}_{\boldsymbol{\gamma}},$$

where $f(\mathbf{y}|\boldsymbol{\theta}_{\boldsymbol{\gamma}'})$ represents the regression distribution of $M_{\boldsymbol{\gamma}'}$, that is, the regression model which is the most similar to $f(\mathbf{y}|\boldsymbol{\theta}_{\boldsymbol{\gamma}})$. We then see that the loss in information associated to model $M_{\boldsymbol{\gamma}}$ is the expected minimum Kullback–Leibler divergence between $M_{\boldsymbol{\gamma}}$ and the nearest one, where the expectation is taken with respect to the prior $\pi_{\boldsymbol{\gamma}}(\boldsymbol{\theta}_{\boldsymbol{\gamma}})$, representing the prior uncertainty about the true values of α, $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$ and ϕ.

As anticipated, to fully describe the *worth* of a regression model it is also necessary to take into considerations the complexity of the model. For the regression model $M_{\boldsymbol{\gamma}}$, we denote the loss due to complexity by $\mathrm{L}_C(M_{\boldsymbol{\gamma}})$, which it is determined as follows. If we keep model $M_{\boldsymbol{\gamma}}$ in the space of models, the loss would be proportional to the number of covariates that have to be considered and measured. Therefore, the loss of keeping a linear regression model increases with the number of covariates it contains, and we have

$$\mathrm{L}(\text{remove } M_{\boldsymbol{\gamma}}) = \mathrm{U}(\text{keep } M_{\boldsymbol{\gamma}}) = -c \cdot |\boldsymbol{\gamma}|,$$

and

$$\mathrm{L}(\text{keep } M_{\boldsymbol{\gamma}}) = c \cdot |\boldsymbol{\gamma}|, \qquad c > 0,$$

where $\mathrm{L}(\cdot)$ represents a loss and $\mathrm{U}(\cdot)$ a utility.

The loss component due to complexity is easily fit in our framework and the model prior for $M_{\boldsymbol{\gamma}}$ is

$$p(M_{\boldsymbol{\gamma}}) \propto \exp\left\{\int_{\boldsymbol{\theta}_{\boldsymbol{\gamma}}} \left[\min_{\boldsymbol{\gamma}' \neq \boldsymbol{\gamma}} D_{KL}\Big(f(\mathbf{y}|\boldsymbol{\theta}_{\boldsymbol{\gamma}}) \| f(\mathbf{y}|\boldsymbol{\theta}_{\boldsymbol{\gamma}'})\Big) \pi_{\boldsymbol{\gamma}}(\boldsymbol{\theta}_{\boldsymbol{\gamma}})\, d\boldsymbol{\theta}_{\boldsymbol{\gamma}}\right] - c \cdot |\boldsymbol{\gamma}|\right\}. \tag{10}$$

In other words, the prior is constructed based upon a cumulative loss with a component representing the loss in information and a component representing the loss due to complexity.

The following Theorem 3.1 (which proof is in Appendix A in the Supplementary Material (Villa and Lee, 2019)) shows the expression of the minimum Kullback–Leibler divergence between regression models.

Theorem 3.1. *Let* $M_{\boldsymbol{\gamma}} = \{f(\mathbf{y}|\boldsymbol{\theta}_{\boldsymbol{\gamma}}); \pi_{\boldsymbol{\gamma}}(\boldsymbol{\theta}_{\boldsymbol{\gamma}})\}$ *and* $M_{\boldsymbol{\gamma}'} = \{f(\mathbf{y}|\boldsymbol{\theta}_{\boldsymbol{\gamma}'}); \pi_{\boldsymbol{\gamma}'}(\boldsymbol{\theta}_{\boldsymbol{\gamma}'})\}$ *be linear normal regression models as in* (2)*, with design matrices, respectively,* $\mathbf{X}_{\boldsymbol{\gamma}}$ *and* $\mathbf{X}_{\boldsymbol{\gamma}'}$*. If* $\mathbf{X}_{\boldsymbol{\gamma}'}^T \mathbf{X}_{\boldsymbol{\gamma}'}$ *is invertible, the minimum Kullback–Leibler divergence between* $f(\mathbf{y}|\boldsymbol{\theta}_{\boldsymbol{\gamma}})$ *and* $f(\mathbf{y}|\boldsymbol{\theta}_{\boldsymbol{\gamma}'})$ *is*

$$\min_{\boldsymbol{\theta}_{\boldsymbol{\gamma}'}} D_{KL}\Big(f(\mathbf{y}|\boldsymbol{\theta}_{\boldsymbol{\gamma}}) \| f(\mathbf{y}|\boldsymbol{\theta}_{\boldsymbol{\gamma}'})\Big) = 0 \qquad \forall\, \boldsymbol{\gamma} \neq \boldsymbol{\gamma}'. \tag{11}$$

Theorem 3.1 shows that the minimum Kullback–Leibler divergence between any two linear regression models is zero, regardless to the number of covariates in the models. This means that, in variable selection problems for linear regression models, there is

no loss in information in selecting the "wrong" model, as such the model prior in (10) becomes

$$p(M_{\boldsymbol{\gamma}}) \propto \exp\{-c \cdot |\boldsymbol{\gamma}|\}. \tag{12}$$

It is therefore equation (12), for a suitable c, that will be used in the paper to represent the prior uncertainty on the space of regression models.

3.2 Setting the constant c

The proposed prior in (12) depends on the constant c, which can be interpreted as the penalty factor on the number of covariates included. So, the question is how to calibrate c, noting that it controls the rate at which the prior decreases as the model size increases, as well as the way the prior behaves. Here we discuss some ideas, intended to help the analyst, but we do not claim to be exhaustive. Our intention is to provide some insights on how the choice of c can be carried out on the basis, for example, whether prior information about the covariates to be included in the model is available, or if multiplicity correction is a matter of concern. There are fundamentally three options;

- c is fixed to a specific value. This can be either on the basis of any available initial knowledge about the number of covariates to be included in the model or in agreement to some criterion.
- c is a function of d.
- Adopting a hierarchical approach, a prior is assigned on c. Here as well it is possible to reflect any available prior information in the hierarchical structure or opt for a noninformative choice.

Fixed c If the constant c is fixed to a specific value which does not depend on the total number of covariates in the problem (see the next paragraph below), there is no multiplicity correction unless this reflects some prior knowledge (Scott and Berger, 2010). Although one may agree that multiplicity correction is one way to define model prior probabilities in an objective sense, it is argued that this is not a necessary condition for a model prior to satisfy, in particular if the problem of interest is related to prediction. Furthermore, this is not an issue should one believe that for variable-selection problems the approach should be in line with the Bayesian framework of having prior and posterior probability representing, respectively, prior and posterior uncertainty.

To understand how the constant c acts like a penalty factor, we note that as $c \to 0$, $p(M_{\boldsymbol{\gamma}}) \to 1/\{2^d\}$ for all models, yielding the uniform prior. As c gets larger, the prior assigns more and more mass to sparse models, and the rate at which the prior drops to zero increases.

The noninformative criterion we propose to set c is based on maximising the expected loss with respect to c. From equation (12), let $K = |\gamma|$; then the prior on model size $K = k$ given c is

$$P(k|c) \propto \binom{d}{k} e^{-ck},$$

which, normalised, gives

$$\begin{aligned} P(k|c) &= \binom{d}{k} e^{-ck} \Big/ \sum_{l=0}^{d} \binom{d}{l} e^{-cl} \\ &= \binom{d}{k} \frac{e^{-ck}}{(e^{-c}+1)^d}. \end{aligned} \tag{13}$$

Then, the expectation of K given $c > 0$ is

$$\mathbb{E}(K|c) = \sum_{k=0}^{d} kP(k|c) = \frac{d}{e^c+1},$$

and the expected loss, in terms of c, is given by

$$\mathbb{E}_L(K|c) = c\frac{d}{e^c+1}. \tag{14}$$

We note that (14) goes to 0 for $c \to 0$ and for $c \to \infty$. As such, we can maximise the expected loss in terms of c. Differentiating

$$\begin{aligned} \frac{d}{dc}\frac{cd}{e^c+1} &\propto [\log c - \log(e^c+1)] \\ &= \frac{1}{c} - \frac{e^c}{e^c+1}, \end{aligned}$$

which solution gives $c = W(e^{-1}) + 1 \approx 1.2785$, where $W(\cdot)$ is the Lambert W function.

In our simulation study, the choice is $c \approx 1.2785$ and the sensitivity raised by c is numerically examined using real data sets in Section 5. We noted the conservative aspect that the above choice induces a prior inclusion probability of $\omega = 0.22$, as it can be seen from equation (15) below.

Function of d The idea is to identify some specific functions of the total number of covariates under examination that have desirable properties. Having then c dependent on d, the prior corrects for multiplicity (Scott and Berger, 2010). First, we note that it is advisable to have a prior capable to produce sensible results even for large values of d. This is dictated to the obvious fact that modern regression models can easily include thousands (and more) covariates. One choice could simply be to set $c = d$; however, even for moderate values of d, the prior will exhibit an extremely fast decrease to zero with the consequence of assigning most of the mass to the very sparse models. Another possible choice would be to set $c = d^{-1}$. While this function would allow to avoid the previous undesirable behaviour, the consequence is that the prior will rapidly converge to a uniform prior, with all the negative caveats discussed in Scott and Berger (2010). We believe that a sensible choice is to have $c = \log(d)$. In a scenario of sparsity, most of the considered covariates will bring little (if not zero) information; as such, a desirable property of the prior would be to not "implode" as d grows, which is the case of, for example, the above choice of $c = d$. At the same time, it would be desirable to have a prior

that it is sensible to minimal changes for small d. A process that exhibit this behaviour is the logarithmic process, which shows rapid growth when it is small and slow growth when it is large. Hence, having $c = \log(d)$ appears to be a sensible choice, and results for simulated data in Section 4 appear to provide empirical support for this choice.

Hierarchical approach The third approach to calibrate c consists in assigning a prior distribution to c; thus, obtaining

$$P(M_{\boldsymbol{\gamma}}) = \int_0^\infty P(M_{\boldsymbol{\gamma}}|c)\pi(c)\, dc,$$

for a suitable density $\pi(c)$. Although setting, for example, $c \sim \text{Ga}(a, b)$, for some $a, b > 0$, represents a sensible choice, the resulting prior would be analytically intractable and, moreover, the calibration of the hyperparameters a and b is not straightforward. A more interpretable approach is as follows. First, we note that

$$\begin{aligned} P(M_{\boldsymbol{\gamma}}|c) &= e^{-c|\gamma|} \Big/ \sum_{l=0}^{d} \binom{d}{l} e^{-cl} \\ &= \frac{e^{-c|\gamma|}}{(1+e^{-c})^d} \\ &= \left(\frac{1}{1+e^c}\right)^{|\gamma|} \left(\frac{e^c}{1+e^c}\right)^{d-|\gamma|}, \qquad c > 0, k = 0, 1, \ldots, d. \end{aligned}$$

We then set $\omega = (1+e^c)^{-1}$, giving

$$P(M_{\boldsymbol{\gamma}}|c) = \omega^{|\gamma|}(1-\omega)^{d-|\gamma|}, \qquad \omega \in (0, 1/2). \tag{15}$$

By assigning the following Generalised Beta distribution to ω,

$$\pi(\omega) = \frac{2^p}{B(p,q)}\omega^{p-1}(1-2\omega)^{q-1}, \qquad p, q > 0, \tag{16}$$

that is $\omega \sim \text{GB}(a = 1, b = 1/2, c = 0, p, q)$, we have

$$P(M_{\boldsymbol{\gamma}}) = \int_0^{1/2} P(M_{\boldsymbol{\gamma}}|c)\pi(\omega)\, d\omega. \tag{17}$$

By inserting (15) and (16) in equation (17), we have

$$\begin{aligned} P(M_{\boldsymbol{\gamma}}) &= \int_0^{1/2} \omega^{|\gamma|}(1-\omega)^{d-|\gamma|} \frac{2^p}{B(p,q)}\omega^{p-1}(1-2\omega)^{q-1}\, d\omega \\ &= \frac{2^p}{B(p,q)} \int_0^{1/2} \omega^{|\gamma|+p-1}(1-\omega)^{d-|\gamma|}(1-2\omega)^{q-1}\, d\omega \\ &= \frac{2^p}{B(p,q)} \int_0^{1/2} \omega^{|\gamma|+p-1}(1-\omega)^{d-|\gamma|} \sum_{m=0}^{q-1} \binom{q-1}{m} (-2)^m \omega^m\, d\omega \end{aligned}$$

$$
\begin{aligned}
&= \frac{2^p}{B(p,q)} \sum_{m=0}^{q-1} \binom{q-1}{m} (-2)^m \int_0^{1/2} \omega^{m+|\gamma|+p-1} (1-\omega)^{d-|\gamma|}\, d\omega \\
&= \frac{2^p}{B(p,q)} \sum_{m=0}^{q-1} \binom{q-1}{m} (-2)^m B_{1/2}(m+|\gamma|+p, d-|\gamma|+1), \qquad (18)
\end{aligned}
$$

where $B_{1/2}(\cdot,\cdot)$ is the incomplete Beta function defined over $(0, 1/2)$.

The hyperparameters p and q can be chosen to reflect prior information about ω (and hence about c), if available. A noninformative option would be to assign a uniform prior to ω, corresponding to a Generalised Beta distribution with $p = q = 1$, giving $\pi(\omega) = 2$, for $\omega \in (0, 1/2)$. The prior of the regression model, from (18) above, $M_{\boldsymbol{\gamma}}$ becomes

$$
\begin{aligned}
P(M_{\boldsymbol{\gamma}}) &= 2 \int_0^{1/2} \omega^{|\gamma|} (1-\omega)^{d-|\gamma|}\, d\omega \\
&= 2B_{1/2}(|\gamma|+1, d-|\gamma|+1).
\end{aligned}
$$

Should one be interested in expressing the prior distribution with respect to the number of covariates included in the regression model, $K = k$, one has to consider

$$
\begin{aligned}
P(k|c) &= \binom{d}{k} \left(\frac{1}{1+e^c}\right)^k \left(\frac{e^c}{1+e^c}\right)^{d-k} \\
&= \binom{d}{k} \omega^k (1-\omega)^{d-k},
\end{aligned}
$$

with $\omega = (1+e^c)^{-1}$. We have, by assigning a Generalised Beta distribution to ω,

$$
P(k) = \binom{d}{k} \frac{2^p}{B(p,q)} \sum_{m=0}^{q-1} \binom{q-1}{m} (-2)^m B_{1/2}(m+k+p, d-k+1),
$$

and, for $p = q = 1$, that is the noninformative choice,

$$
P(k) = \binom{d}{k} 2B_{1/2}(k+1, d-k+1).
$$

Figure 1 compares the proposed prior with $c = 1.2785$, $c = \log(d)$ and the hierarchical loss-based to the Scott and Berger prior for $d = 30$. In other words, it compares noninformative choices of $P(k)$. Whilst Scott and Berger prior has a symmetrical behaviour, the proposed prior assigns more mass to the more simple models than to the more complex ones, as expected from expression (12). We note how the different choice of c impacts the rate at which the prior mass decreases as the model becomes more complex. In particular, for $c = \log(d)$ the prior assigns the highest mass to the null model with a quick drop in the prior probability for already moderate values of the number of covariates. While the choice of setting $c = 1.2785$ allows for a more distributed prior mass among the sparse model. The hierarchical approach, as it can be seen from the plot, yields a prior distribution that is relatively high for values of the number of covariates smaller than $d/2$, to decrease towards zero afterwards. It is the

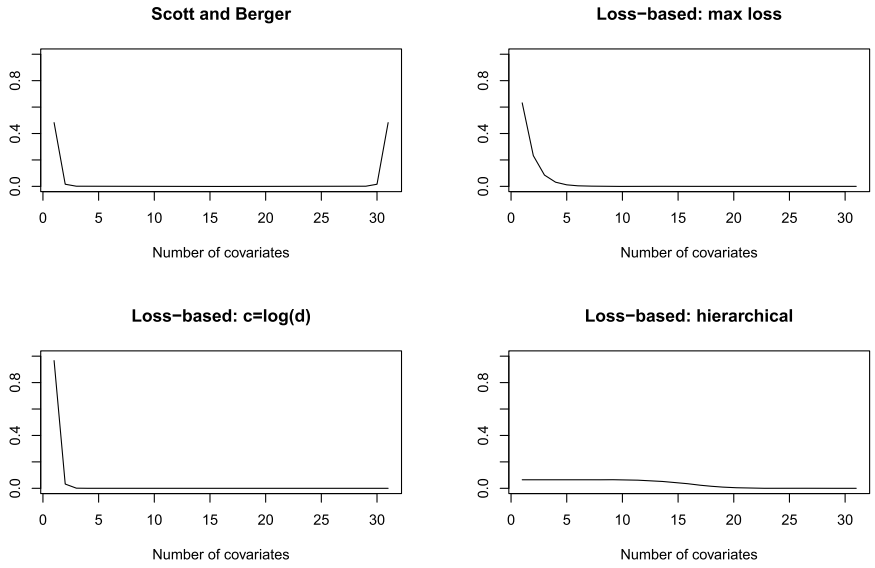

Figure 1: Prior model comparison when the full model contains $d = 30$ covariates.

above characteristics of the proposed prior of assigning more pass to the lower region of the parameter space (although with different behaviours) that makes the loss-based prior more suitable in scenarios where preference is put on the more sparse models, when compared to the Scott and Berger prior.

Figure 2 below shows examples of the loss-based prior for different values of the parameters p and q reflecting, respectively, an approximate prior mean of 1, 3, 5 and

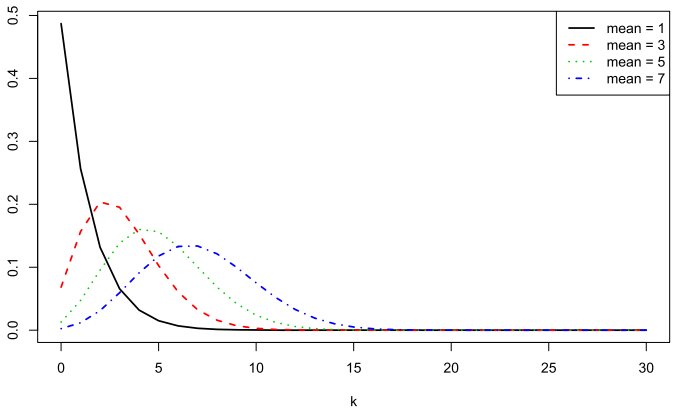

Figure 2: Loss-based prior obtained by using the hierarchical approach with the Generalised Beta.

7 (with $d = 30$). The parameters, numerically obtained, of the Generalised Beta are, respectively, $p = 1, 6, 8, 8$ and $p = 14, 23, 16, 9$.

4 Simulation study

In this section we present the results of simulation studies designed to assess the performance of the proposed prior and to compare it to the Scott and Berger's prior. We consider different scenarios, in terms of number of covariates, sample size as well as whether prior information is available or not.

It is well known that a variable selection problem is driven by both the choice of the model prior and of the model-specific parameter prior. However, here the interest is in the effects on variable selection determined by the prior probability on the space of models. As such, the simulation exercise described has the purpose to analyse the frequentist properties of the posterior distribution on the model size and, to minimise any possible effects of the model-specific parameters prior, we choose to use the robust prior for the parameter of the regression.

In the first simulation study, the performances of the priors for various scenarios are compared considering the posterior means squared error from the mean (MSE), the coverage of the posterior 95% credible interval and the frequency rate for identifying the true model by the HPM. The detailed results are reported in Tables in Appendix B (included in the Supplementary Material), while graphical analysis of the MSE and the coverage will be presented in the current Section. The second simulation study is limited to relatively sparse models and it consider the case where prior information, in terms of the mean number of covariates the model should include, is available and it is reflected in the choice of prior hyperparameter(s). The simulation considers correct prior information as well as inaccurate prior knowledge.

4.1 Non-informative simulation

The study involves 250 experiments and each experiment uses four data sets, one for each $d = 5, 10, 15, 30$, and repeated for $n = 50$ and $n = 100$. The procedure of each experiment follows;

- Generate a design matrix $\mathbf{X}$ of size $n \times d$ where each element is an independent realisation of a standard normal distribution;
- Generate a binary vector $\boldsymbol{\gamma}$ from a sequence of d independent Bernoulli experiments with probability of success equal to ω;
- From the robust prior in (5), generate the vector of coefficients $\boldsymbol{\beta}_{\boldsymbol{\gamma}}$;
- Generate the response vector from the regression model in (1), considering $\phi = 1$;
- Using the above values of the design matrix and the vector of responses, compute the necessary quantities, including the marginal likelihoods, the model posteriors and the model size posterior distribution.

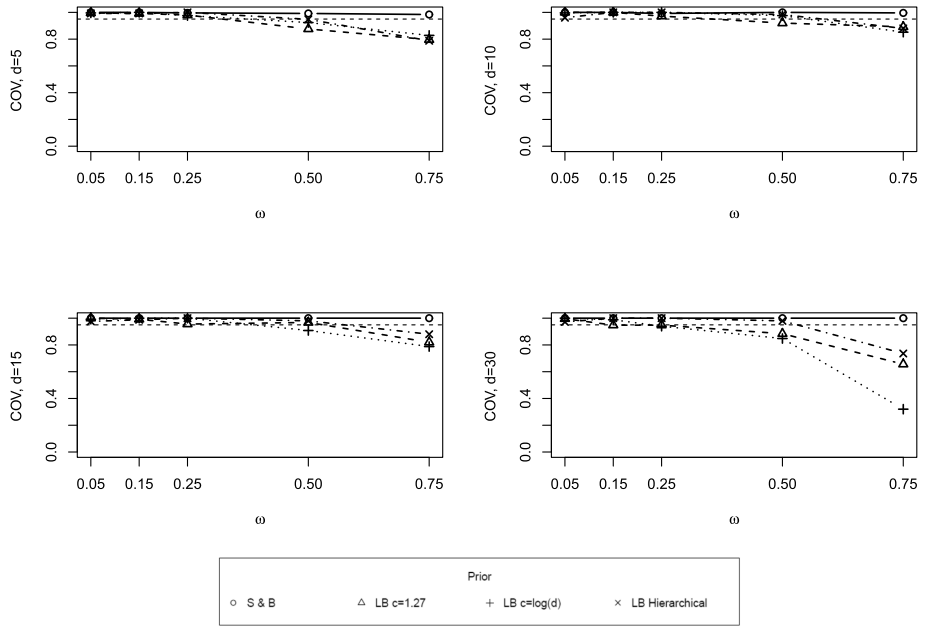

Figure 3: Coverage of the 95% posterior credible interval under the Scott and Berger prior and the loss-based prior, for the three methods of calibration of c. The plots represent the posterior summary statistic for different values of d and for $n = 50$.

The last step of the above procedure has been performed under the Scott and Berger prior and the loss-based prior with the three proposed methods to calibrate c, that is setting $c = 1.2785$ (so to maximise the expected loss), $c = \log(d)$ and the hierarchical approach as described in the previous section, where the parameters of the Generalised Beta have been both set to one. These choices for c are made for the case in which no prior information about the true model size is available, but sparsity is expected.

The simulation result using the loss-based prior and Scott and Berger prior are summarized in Appendix B (included in the Supplementary Material). Five values of ω are considered to examine small to large model sizes. For each model prior, the coverage of the 95% credible interval of the posterior (*Coverage*), the mean squared error of posterior mean (*MSE Mean*) and frequency rate of the HPM equals to the true model (*Freq. True*) from 250 experiments are estimated. Finally, we note that for $d = 30$ we have used a Gibbs search to explore the space of models resulting in relatively low frequencies for the true model, as well as a larger variability in the different performances of the priors.

The coverage (Figures 3 and 4) under the Scott and Berger prior appears to be the most stable across sample size and model size, although it is mostly above the nominal value of 95%. The prior we propose appears to have a relatively low coverage when $\omega = 0.75$; this is particularly obvious for the prior with $c = \log(d)$ set up, and it is due

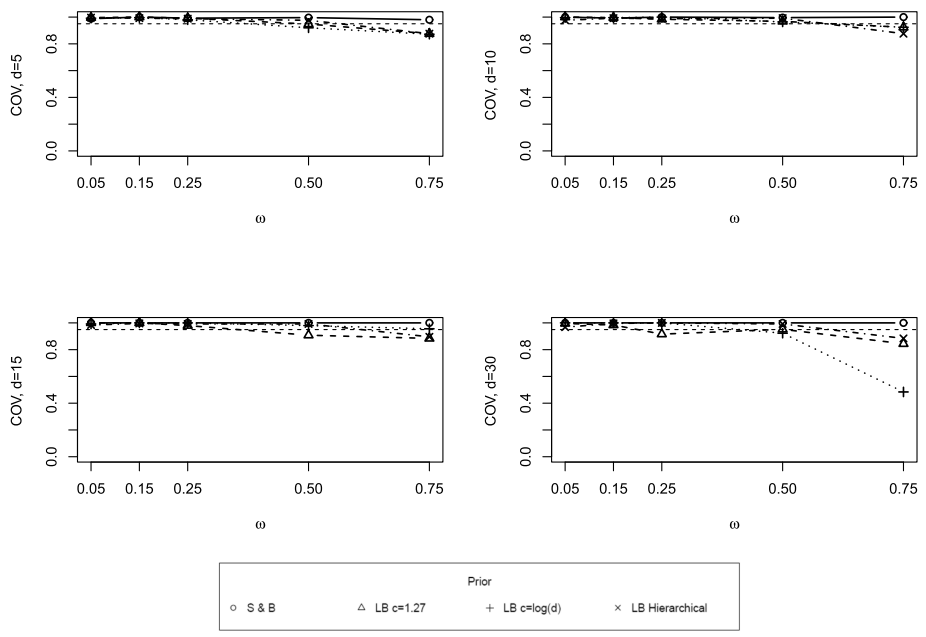

Figure 4: Coverage of the 95% posterior credible interval under the Scott and Berger prior and the loss-based prior, for the three methods of calibration of c. The plots represent the posterior summary statistic for different values of d and for $n = 100$.

to the fact that the distribution goes to zero (for large k) much faster than the other two loss-based options. However, the loss-based prior shows an overall performance in the coverage closer to the nominal value than the Scott and Berger's. For $n = 100$, as one would expect, the differences between the priors becomes smaller in comparison to the case $n = 50$; again, we notice that the loss-based prior with $c = \log(d)$ has the worst performance when $\omega = 0.75$.

Considering the HPM when it refers to the true model, we note that all the priors exhibit a similar pattern (Appendix B in the Supplementary Material). In fact, for any value of d, the frequency the true model is identified decreases as the average model size (i.e. ω) increases. For $n = 50$ the frequencies between the priors have more variability when compared to the case $n = 100$. Furthermore, more variability in the performance is observed for $d = 30$; as in this case the inference is performed through a random exploration of the model space (i.e. Gibbs), a higher degree of uncertainty is included in the process by the fact that the model space itself is very large.

Comparing the MSE in Figure 5 and Figure 6, we note the following. First, as expected, the MSE is larger for $n = 50$ then for $n = 100$ as the information in the data increases. For $d = 5$, that is for a small model space, the priors tend to be quite similar, overall. We note that the Scott and Berger prior tends to perform better for relatively large models, while the loss-based priors with $c = 1.2785$ and $c = \log(d)$ have

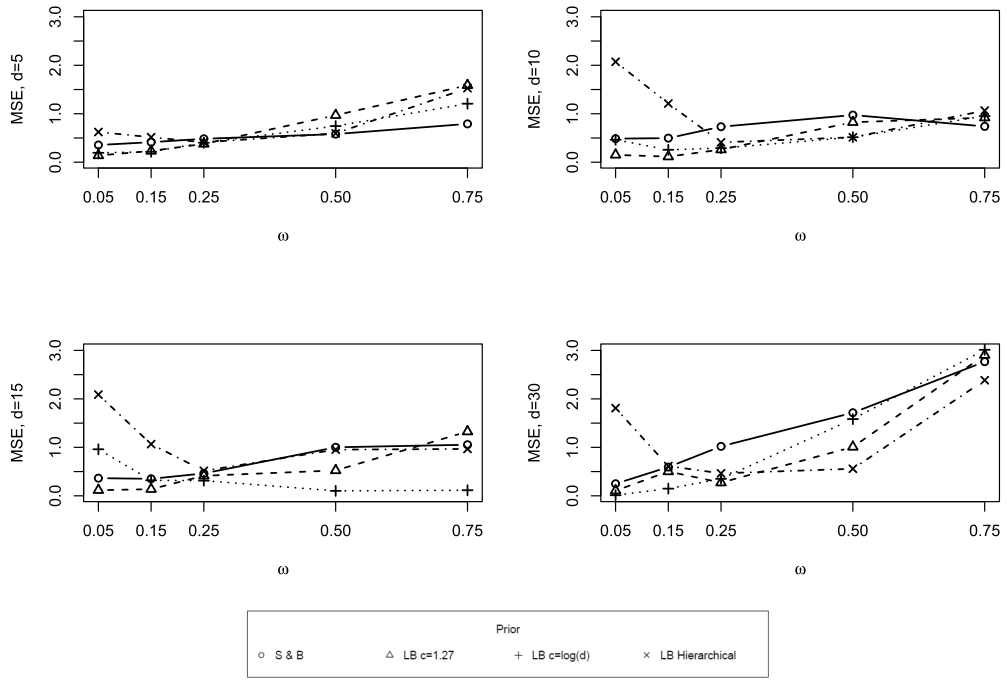

Figure 5: MSE under the Scott and Berger prior and the loss-based prior, for the three methods of calibration of c. The plots represent the posterior summary statistic for different values of d and for $n = 50$.

smaller MSE for $\omega = 0.05$. The loss-based prior with a hierarchical approach has the largest MSE for relatively smaller models, and this behaviour is consistent for any n and any d. In fact, its behaviour of spreading most of its mass evenly in the lower part of the parameter space of k, renders it weak in dealing with models with a relatively small number of covariates. When d is either 10 or 15, we note a similar pattern in the remaining priors (the Scott and Berger's and the loss-based with $c = 1.2785$ and $c = \log(d)$) as to when $d = 5$. Differences are a bit more accentuated, in particular for small sample sizes, but the overall performances are quite stable. As we have observed for the frequency of "guessing" the true model, when $d = 30$ the random search contributes in increasing the variability of the differences. For example, the Scott and Berger prior does not have the smallest MSE for $\omega = 0.75$.

Undoubtedly, the Scott and Berger prior will have a higher degree of efficiency when the true model is the full model (or a true model that contains a very large proportion of possible covariates). In fact, if one suspects that the true model is quite large, possibly this is the prior to employ. On the other hand, when the model tends to be sparse (or when it is thought that this is the case), the loss-based approach appears to give better results; at least, when we calibrate $c = 1.2785$ or $c = \log(d)$. The latter option might be preferable should one be concerned with multiplicity correction. The

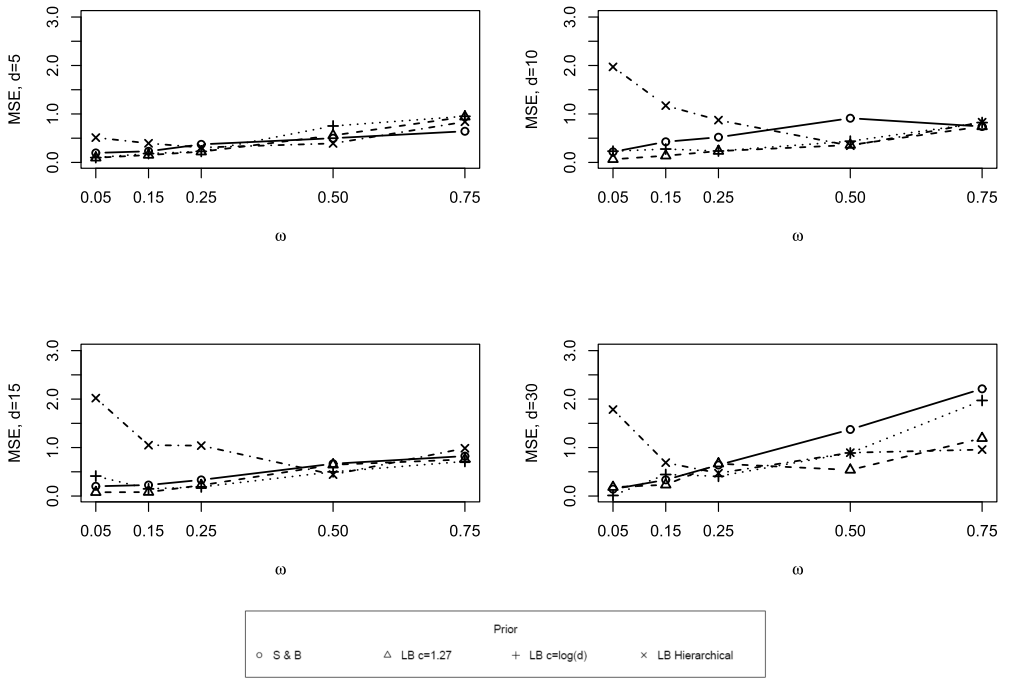

Figure 6: MSE under the Scott and Berger prior and the loss-based prior, for the three methods of calibration of c. The plots represent the posterior summary statistic for different values of d and for $n = 100$.

hierarchical approach here explored, when non information is included in the prior, is the less preferable, as it has been shown by the summary statistics considered.

4.2 Informative simulation

In this simulation study we analyse the performance of the loss-based prior when prior information about the true size of the model is available and we compare its performance with the Beta-Binomial prior with same mean and variance. For the coefficients, we have employed the 'robust' prior in (5). We have simulated 200 data sets of different sizes, $n = 50, 100$, with $d = 5, 10, 15$, covariates. We have started by considering the vector of coefficients $\boldsymbol{\beta} = (0.5, -0.5, 0, 0, -1)$ for $d = 5$ and added 5 and 10 extra null coefficients for the simulations with, respectively $d = 10$ and $d = 15$. As all scenarios yielded the same conclusion that the priors perform in the same manner, we show the details of the more complex ones only, that is for $n = 50, 100$ and $d = 15$. We proceeded as follows.

We assume to have prior information about the mean number of covariates that the true model includes, say m, and we look at scenarios where this information is correct, i.e. $m = 3$, and where it is inaccurate, i.e. $m = 1, 5, 7$. With this piece of information,

m	c	$\text{Var}(K)$	a	b
1	2.64	0.93	131	1828
3	1.39	2.40	672	2687
5	0.69	3.33	233	466
7	0.13	3.73	3658	4181

Table 1: Values of c, a and b for the simulation study to compare the loss-based prior with the Beta-Binomial prior for $d = 15$.

we obtain the value of c for the loss-based prior in (12) by setting

$$\mathbb{E}(K|c) = \frac{d}{e^c + 1} = m. \tag{19}$$

With the obtained value of c, we then derive the variance of the loss-based prior by applying

$$\begin{aligned} \text{Var}(K|c) &= \mathbb{E}(K^2|c) - [\mathbb{E}(K|c)]^2 \\ &= \frac{d(e^c + d)}{(e^c + 1)^2} - \frac{d^2}{(e^c + 1)^2} \\ &= \frac{de^c}{(e^c + 1)}. \end{aligned} \tag{20}$$

For example, for $d = 5$ and $m = 1$, we have $c = 2.64$ and $\text{Var}(K) = 0.31$. We then obtain the values of a and b of the Beta-Binomial prior by equating the expectation and the variance of the distribution to m and $\text{Var}(K)$, respectively, and solve with respect to the two parameters. Table 1 shows the values associated with the simulation study for $d = 15$ and $m = 1, 3, 5, 7$. The results for $n = 30$ and $n = 100$ are reported in Appendix C (included in the Supplementary Material) in graphical form. We note that the loss-based prior and Beta-Binomial prior, when they have same means and same variances, is almost identical (within the same scenario and for the same coefficient). As one would expect, the posterior inclusion probability reflects more accurately the true nature of a coefficient when n is relatively large; or when the true value is relatively different from zero (i.e. for β_5). If we consider the priors performance on the basis of the accuracy of the prior information m, we see that the posterior inclusion probabilities of the non-null coefficients is better for a large (although inaccurate) m in the case of $n = 50$. This is a consequence of having a prior that puts more mass on relatively large models; however, there is also a larger inclusion of null coefficients. For $n = 100$ the above effect is drastically reduced, in the sense that for $m = 3$ (the true value), the inclusion posterior probabilities, overall, reflect the true status of the regression model.

The Figures in Appendix C (refer to the Supplementary Material) contain also the results when the hierarchical loss-based prior is used (right violin plots). The parameters p and q of the Generalised Beta density have been chosen so to have mean m and variance similar to the one in Table 1 (as mentioned at the end of Section 3 above). We note that, for $n = 50$, the hierarchical loss-based prior has similar performances to the

other priors when the coefficient is either 0 or -1 (i.e. β_3 to β_{15}). When $\beta = \pm 0.5$, i.e. β_1 and β_2, its performance appear to be better when the prior information about the mean is either accurate or larger than the true one. In fact, we note that the means of the posterior inclusion probabilities are above 0.5. For $m = 1$, although the posterior inclusion probability yielded (in mean) is below the 0.5 threshold, it has a larger value when any of the other two priors is used. For $n = 100$, the above differences are still noticeable, but with the magnitude that is diminished by the increase in information in the data.

A remark is that the loss-based prior has only one parameter that has to be calibrated. This allows a single piece of information (e.g. the mean) to be easily reflected in the prior, while the Beta-Binomial prior and the hierarchical loss-based would need to fix one of the to parameters "freely". It is true that two parameters would allow to include an extra piece of information, such as variability, but in practice this information is not known or, at least to a practitioner, it is not easy to define.

5 Illustrative examples with real data sets

In this section we investigate the properties of objective model priors for variable selection in real data sets. The first considered data set is the Hald data (Woods et al., 1932), which have been extensively used in the literature (see Kubinyi (1996) and Liang et al. (2008), for example). The second data set considered concerns with the study of gene expression data in colon cancer patients (Calon et al., 2012). For both examples we compare the Scott and Berger prior with the three approaches for the loss-based prior discussed in Section 3.2.

5.1 Hald data

The Hald data set contains $n = 13$ observations with $d = 4$ covariates, and it concerns an engineering application to study the cement composition (Woods et al., 1932). In particular, the study considers the effect on the heat evolved per gram of cement (in calories) by the amount of tricalcium aluminate, the amount of tricalcium silicate, the amount of tricalcium aluminio ferrite and the amount of dicalcium silicate.

The summary statistics of the model size posterior distributions are presented in Table 2. The corresponding histograms of the posterior distributions are represented in Figure 7. The loss-based priors appear to yield similar posteriors for the number of covariates. In fact, the posterior distributions have very similar statistics and histograms. The Scott and Berger prior is more spread with a slightly higher mean. Although the above differences, all priors basically point in the direction of the same regression model for the Hald data set.

The above conclusion is supported by the information in Table 3, where we note that the MPM is the same under each prior with posterior inclusion probabilities that clearly suggest the inclusion of both the Tricalciums and the exclusion of the remaining two covariates. In the table we have also reported the posterior probabilities associated to the

Model prior	Mean	Median	SD	95% C.I.	HPM	MPM
Scott & Berger	2.41	2	0.81	(2,4)	2	2
Loss-based ($c = 1.2785$)	2.09	2	0.72	(2,3)	2	2
Loss-based ($c = \log(d)$)	2.08	2	0.72	(2,3)	2	2
Loss-based (Hierarcical)	2.12	2	0.71	(2,3)	2	2

Table 2: Comparison of the posterior summary statistics for the Hald data set. Four statistics for the number covariates are measured: mean, median, standard deviation (SD) and the 95% confidence interval (95% C.I.). The number of covariates included in the HPM and in the MPM are reported. The hierarchical version of the Loss-based prior is as discussed in Section 3.2 and it is based on the Generalised Beta distribution with parameters $p = q = 1$.

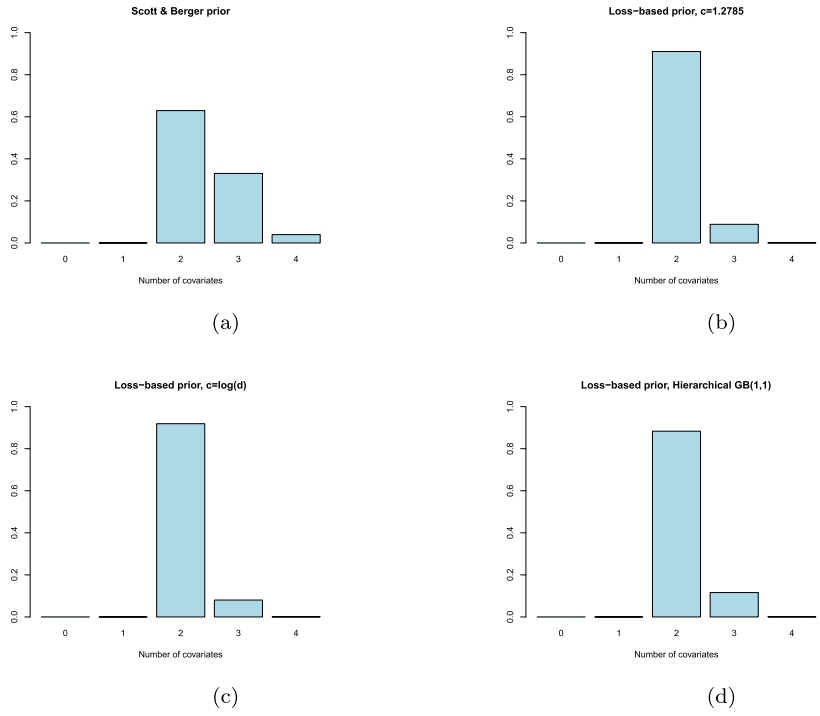

Figure 7: Posterior distribution of the number of covariates for the Hald data set. Four model priors are considered: Scott and Berger, loss-based proposed prior with $c = 1.2785$ (to maximise the expected loss), the loss-based prior with $c = \log(d) = \log(4) = 1.39$ and the loss-based prior with the hierarchical approach based on the Generalised Beta distribution GB(1,1).

highest posterior density (HPD) interval. Again, the conclusions are in direction on the above model, but we note that the loss-based prior (under each method) yields a posterior probability definitely larger than the once yielded by the Scott and Berger prior.

	Posterior Inclusion Probability			
Covariate	S&B	LB-ml	LB-lg	LB-gb
Ticalcium aluminate	**0.98•**	**0.99•**	**0.99•**	**0.99•**
Tricalcium silicate	**0.76•**	**0.75•**	**0.75•**	**0.75•**
Tetracalcium alumino ferrite	0.26	0.09	0.06	0.08
Dicalcium silicate	0.42	0.29	0.28	0.30
HPM Posterior Prob.	0.47	0.69	0.69	0.67

Table 3: Posterior inclusion probabilities for the Hald data set. The covariates with a posterior inclusion probability greater than 1/2 are highlighted in bold, and a dot notation represents the covariate included in the highest posterior probability model. The prior compared are: S&B (Scott and Berger), LB-ml (loss-based with $c = 1.2785$, to maximise the expected loss), LB-lg (loss-based with $c = \log(d)$) and LB-gb (hierarchical loss-based using the Generalised Beta GB(1,1)). The table includes the posterior probability of the HPM under each prior.

5.2 Large data set analysis

Data sets with a large number of covariates are, nowadays, widespread. It is therefore important to illustrate how the proposed method deals with a practical problem with a large size, in terms of covariates. We illustrate the prior performance with the human micro-array gene expression data in colon cancer patients. This data set was originally discussed in Calon et al. (2012) and it consists of $d = 172$ predictors for a total of $n = 262$ patients. The aim is to identify which genes have an effect on the expression of transforming growth factor-beta (TGFB). Although the whole data set consists of $10,172$ genes, we limit the dimension as we are working under the assumption that $n > d$, and the first 172 genes are the key ones for a preliminary analysis (Rossell and Telesca, 2017). The analysis has been performed by running 10000 simulations, with a burn-in period of 500, using the Gibbs-search mechanism implemented into the 'BayeVarSel' R-package under the *g-Zellner* prior with $g = d^2$. Posterior statistics are summarised in Table 4, while Table 5 shows the probeset identifiers (which can be associated to genes ESM1, GAS1, HIC1, CILP and IGFBP3) contained in the HPMs under each prior distribution.

When we consider a large data set we note important differences in the priors. Both Scott and Berger and the loss-based prior with $c = \log(d)$ give similar results. In particular, in Table 4 we see that the posterior statistics are virtually the same and, from Table 5, that they both identify as the "best" model the one with four probesets. When we consider the loss-based prior with a value of c chosen so that the expected loss is maximised, the posterior statistics appear to suggest a slightly larger model than the previous one, which is supported by the fact that the extra probeset '212143_s_at' is included in the model; although the inclusion posterior probability is only 0.51. Differently, when the loss-based hierarchical prior is adopted, the inferential process results in what is a different outcome; in fact, the posterior statistics show a wider posterior distribution for the number of covariates and probeset '212143_s_at' is included but with a larger posterior probability (0.75) compared to the above one.

Model prior	Mean	Median	SD	95% C.I.	HPM	MPM
Scott & Berger	4.20	4	0.85	(4,5)	4	4
Loss-based ($c = 1.2785$)	5.71	6	1.38	(4,8)	5	5
Loss-based ($c = \log(d)$)	3.92	4	0.87	(3,5)	4	4
Loss-based (Hierarcical)	8.86	9	2.07	(4,13)	5	5

Table 4: Comparison of the posterior summary statistics for the gene expression data set. Four statistics for the number covariates are measured: mean, median, standard deviation (SD) and the 95% confidence interval (95% C.I.). The number of covariates included in the HPM and in the MPM are reported. The hierarchical version of the Loss-based prior is as discussed in Section 3.2 and it is based on the Generalised Beta distribution with parameters $p = q = 1$.

Covariate	Posterior Inclusion Probability			
(Probeset)	S&B	LB-ml	LB-lg	LB-gb
208394_x_at	**0.82•**	**0.84•**	**0.76•**	**0.84•**
204457_s_at	**0.99•**	**0.99•**	**0.99•**	**0.99•**
230218_at	**0.95•**	**0.93•**	**0.95•**	**0.86•**
206227_at	**0.98•**	**0.99•**	**0.94•**	**0.99•**
212143_s_at	0.23	**0.51•**	0.19	**0.75•**

Table 5: Posterior inclusion probabilities for the gene expression data set. The covariates with a posterior inclusion probability greater than 1/2 are highlighted in bold, and a dot notation represents the covariate included in the highest posterior probability model. The prior compared are: S&B (Scott and Berger), LB-ml (loss-based with $c = 1.2785$, to maximise the expected loss), LB-lg (loss-based with $c = \log(d)$) and LB-gb (hierarchical loss-based using the Generalised Beta GB(1,1)). The table includes the posterior probability of the HPM under each prior.

6 Discussion

This paper introduces a novel prior distribution for the model space in variable selection for linear regression. The prior is based on the idea that, if the "wrong" model is chosen, we incur in a cumulative loss with two components: one represents the loss in information and one related to the complexity of the model expressed by its size. The proposed prior, in its general form, exhibits an exponential decay which depend on the number of covariates included in the model and that can be controlled by the calibration of a constant c. It is therefore possible to reflect any prior information into the prior by setting the constant accordingly. For example, in Section 4.2 we discuss how the constant c can be set up so to reflect a prior expected number of covariates that should be included in the regression model. We also discuss how prior information can be included through a hierarchical approach, in particular by means of a Generalised Beta hyperprior.

It is also possible to represent minimal information in the prior by choosing particular values of the parameters of the prior. In the paper, besides showing through simulation how choices of c perform, we have discussed some general guidelines on how it can be

fixed, considering or not considering correction for multiplicity, and how c can be either directly fixed or modelled through an appropriate prior density in a hierarchical set up.

The simulation studies are carried out to show frequentist performance of the proposed model prior relative with selective choices on c and they are compared to the Scott and Berger prior (Scott and Berger, 2010), when the true model is relatively sparse. In the case where prior information about the model size is available, we show how the constant c can be easily calibrated so to reflect, for example, prior information about the mean number of covariates one believes should be included in the model.

When it comes to real data analysis, we note a fair closeness of the results obtained by using the proposed prior with the one of Scott and Berger's when the size of the problem is small (Hald data). Both MPM and HPM represent the same regression model under each prior, although the loss-based priors result in a higher posterior probability for the HPM than Scott and Berger's.

We have also analysed a relatively large data set (colon cancer data), in terms on number of covariates. Here we note that the loss-based prior with $c = \log(d)$ and the Scott and Berger prior give very similar results. The loss-based prior with $c = 1.2785$ and the hierarchical loss-based prior, on the other hands, identifies a slightly larger model.

Supplementary Material

A loss-based prior for variable selection in linear regression methods. Supplementary Material (DOI: 10.1214/19-BA1162SUPP; .pdf). The Supplementary Material of "A loss-based prior for variable selection in linear regression" contains the Appendices A, B and C.

References

Barbieri, M. and Berger, J. O. (2004). "Optimal predictive model selection." *Annals of Statistics* **32**, 870–897. MR2065192. doi: https://doi.org/10.1214/009053604000000238. 2

Bayarri, M. J., Berger, J. O., Forte, A. and García-Donato, G. (2012). "Criteria for Bayesian model choice with application to variable selection." *Annals of Statistics* **40**, 1550–1577. MR3015035. doi: https://doi.org/10.1214/12-AOS1013. 2, 4

Bogdan, M., Ghosh J. and Tokar, S. T. (2008). "Selecting explanatory variables with the modified version of the Bayesian information criterion." *Quality and Reliability Engineering International* **24**, 627–641.

Berger, J. O. and Molina, G. (2005). "Posterior model probabilities via path-based pairwise priors." *Statistica Neerlandica* **59**, 3–15. MR2137378. doi: https://doi.org/10.1111/j.1467-9574.2005.00275.x.

Berk, R. H. (1966). "Limiting behaviour of posterior distributions when the model is incorrect." *Annals of Mathematical Statistics* **37**, 51–58. MR0189176. doi: https://doi.org/10.1214/aoms/1177699477. 6

Bernardo, J. M. and Smith, A. F. M. (1994). *Bayesian Theory*. New York: John Wiley & Sons. MR1274699. doi: https://doi.org/10.1002/9780470316870.

Brown, P. J., Vannucci, M. and Fearn, T. (1998). "Bayesian wavelength selection in multi-component analysis." *Journal of Chemometrics* **12**, 173–182.

Calon, A., Espinet, E., Palomo-Ponce, S., Tauriello, D. V. F., Iglesias, M., Céspedes, M. V., Sevillano, M., Nadal, C., Jung, P., Zhang, X. H. F., Byrom, D., Riera, A., Rossell, D., Mangues, R., Massague, J., Sancho, E. and Batlle, E. (2012). "Dependency of colonrectal cancer on the tgf-beta-driven programme in stromal cells for metastasis initiation." *Cancer Cell* **22**, 571–584. 1, 19, 21

Carlin, B. and Louis, T. (2000). "Empirical Bayes: Past, present and future." *Journal of the American Statistical Association* **95**, 1286–1289. MR1825277. doi: https://doi.org/10.2307/2669771.

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

Clyde, M. A. and George, E. I. (2004). "Model uncertainty." *Statistical Science* **19**, 81–94. MR2082148. doi: https://doi.org/10.1214/088342304000000035.

Cui, W. and George, E. I. (2008). "Empirical Bayes vs. fully Bayes variable selection." *Journal of Statistical Planning and Inference* **138**, 888–900. MR2416869. doi: https://doi.org/10.1016/j.jspi.2007.02.011. 4

Fernández, C., Ley, E. and Steel, M. F. J. (2001). "Benchmark priors for Bayesian model averaging." *Journal of Econometrics* **100**, 381–427. MR1820410. doi: https://doi.org/10.1016/S0304-4076(00)00076-2.

Garćia-Donato, G. and Martínez-Beneito, M. A. (2013). "On sampling strategies in Bayesian variable selection problems with large model spaces." *Journal of the American Statistical Association* **108**, 340–352. MR3174624. doi: https://doi.org/10.1080/01621459.2012.742443.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. and Rubin, D. B. (2004). *Bayesian Data Analysis*. Chapman and Hall/CRC. MR2027492. 1

George, E. I. and Foster, D. P. (2000). "Calibration and empirical Bayes variable selection." *Biometrika* **87**, 731–747. MR1813972. doi: https://doi.org/10.1093/biomet/87.4.731.

George, E. I. and McCulloch, R. E. (1993). "Variable selection via Gibbs sampling." *Journal of the American Statistical Association* **88**, 881–889. 1, 4

Hoeting, J. A., Madigan, D., Raftery, A. E. and Volinsky, C. T. (1999). "Bayesian model averaging: a tutorial." *Statistical Science* **14**, 382–401. MR1765176. doi: https://doi.org/10.1214/ss/1009212519. 1

Jeffreys, H. (1961). *Theory of Probability*. London: Oxford University Press. MR0187257.

Kass, R. E. and Raftery, A. E. (1995). "Bayes factors." *Journal of the American*

Statistical Association **90**, 773–795. MR3363402. doi: https://doi.org/10.1080/01621459.1995.10476572.

Bubinyi, H. (1996). "Evolutionary variable selection in regression and PLS analyses." *Chemometrics* **10**, 119–133. 19

Kullback, S. and Leibler, R. A. (1951). "On information and sufficiency." *Annals of Mathematical Statistics* **22**, 79–86. MR0039968. doi: https://doi.org/10.1214/aoms/1177729694.

Ley, E. and Steel, M. F. (2009). "On the effect of prior assumptions in Bayesian model averaging with applications to growth regression." *Journal of Applied Econometrics* **24**, 651–674. MR2675199. doi: https://doi.org/10.1002/jae.1057. 5

Liang, F., Paulo, R., Molina, G., Clyde, M. and Berger, J. O. (2008). "Mixtures of g-priors for Bayesian variable selection." *Journal of the American Statistical Association* **103**, 410–423. MR2420243. doi: https://doi.org/10.1198/016214507000001337. 4, 19

Merhav, N. and Feder, M. (1998). "Universal prediction." *IEEE Transactions on Information Theory* **44**, 2124–2147. MR1658815. doi: https://doi.org/10.1109/18.720534. 6

Nott, D. J. and Kohn, R. (2005). "Adaptive sampling for Bayesian variable selection." *Biometrika* **92**, 747–763. MR2234183. doi: https://doi.org/10.1093/biomet/92.4.747.

O'Hara, R. B. and Sillanpää, M. J. (2009). "A Review of Bayesian Variable Selection Methods: What, How and Which." *Bayesian Analysis* **4**, 85–118. MR2486240. doi: https://doi.org/10.1214/09-BA403.

Raftery, A. E., Madigan, D. and Hoeting, J. A. (1997). "Bayesian model averaging for linear regression models." *Journal of the American Statistical Association* **92**, 179–191. MR1436107. doi: https://doi.org/10.2307/2291462. 2

Rossell, D. and Rubio, F. J. (2017). "Tractable Bayesian variable selection: beyond normality." arXiv:1609.01708. MR3902243. doi: https://doi.org/10.1080/01621459.2017.1371025.

Rossell, D. and Telesca, D. (2017). "Non-local priors for high-dimensional estimation." *Journal of the American Statistical Association* **112**, 254–265. MR3646569. doi: https://doi.org/10.1080/01621459.2015.1130634. 21

Shively, T. S., Kohn, R. and Wood, S. (1999). "Variable selection and function estimation in additive nonparametric regression using a data-based prior." *Journal of the American Statistical Association* **447**, 777–794. MR1723272. doi: https://doi.org/10.2307/2669990.

Scott, J. C. and Berger, J. O. (2010). "Bayes and empirical-Bayes multiplicity adjustment in variable-selection problems." *Annals of Statistics* **38**, 2587–2619. MR2722450. doi: https://doi.org/10.1214/10-AOS792. 1, 4, 5, 8, 9, 23

Villa, C. and Lee, J. E. (2019). "A loss-based prior for variable selection in linear regression methods. Supplementary Material." *Bayesian Analysis*. doi: https://doi.org/10.1214/19-BA1162SUPP. 7

Villa, C. and Walker, S. G. (2015). "An objective Bayesian criterion to determine model prior probabilities." *Scandinavian Journal of Statistics* **42**, 947–966. MR3426304. doi: https://doi.org/10.1111/sjos.12145. 2, 5

Woods, H., Steinour, H. and Starke, H. (1932). "Effect of Composition of Porland Cement on Heat Evolved During Hardening." *Industrial and Engineering Chemistry Research* **24**, 1207–1214. 1, 19

Zellner, A. (1986). "On assessing prior distributions and Bayesian regression analysis with g-prior distributions." In *Bayesian inference and Decision Techniques: Essays in Honour of Bruno de Finetti*, Goel, P. K., Zellner, A. (eds). North-Holland: Amsterdam, 233–243. MR0881437. 4

Zellner, A. and Siow, A. (1980). "Posterior odds ratios for selected regression hypotheses." In *Bayesian Statistics*, Bernardo, J. M., DeGroot, M. H., Lindley, D. V., Smith, A. F. M. (eds). University Press: Valencia, 585–603. 4

Acknowledgments

The authors are grateful for the two referees, the Associate Editor and the Editor for constructive comments on earlier versions of the paper which helped in improving the quality of this work. We are also very thankful to Fabrizio Leisen for his valuable feedback and comments during the drafting of the paper.