Bayesian Analysis (2012) **7**, Number 2, pp. 477–502

Regularization in Regression: Comparing Bayesian and Frequentist Methods in a Poorly Informative Situation

Gilles Celeux*, Mohammed El Anbari†, Jean-Michel Marin‡ and Christian P. Robert§

Abstract. Using a collection of simulated and real benchmarks, we compare Bayesian and frequentist regularization approaches under a low informative constraint when the number of variables is almost equal to the number of observations on simulated and real datasets. This comparison includes new global noninformative approaches for Bayesian variable selection built on Zellner's g-priors that are similar to Liang et al. (2008). The interest of those calibration-free proposals is discussed. The numerical experiments we present highlight the appeal of Bayesian regularization methods, when compared with non-Bayesian alternatives. They dominate frequentist methods in the sense that they provide smaller prediction errors while selecting the most relevant variables in a parsimonious way.

Keywords: Model choice, regularization methods, noninformative priors, Zellner's g–prior, calibration, Lasso, elastic net, Dantzig selector

1 Introduction

Given a response variable, y and a collection of p associated potential predictor variables $x_1, \ldots, x_p$, the classical linear regression model imposes a linear dependence on the conditional expectation (Rao, 1973)

$$\mathbb{E}[y|x_1, \ldots, x_p] = \beta_0 + \beta_1 x_1 + \ldots \beta_P x_p \,.$$

A fundamental inferential direction for those models relates to the variable selection problem, namely that only variables of relevance should be kept within the regression while the others should be removed. While we cannot discuss at length the potential applications of this perspective, variable selection is particularly relevant when the number p of regressors is larger than the number n of observations (as in microarray and other genetic data analyses).

*Project SELECT, INRIA Saclay, Université Paris Sud, Orsay, France, Gilles.Celeux@inria.fr
†Université Caddi Ayyad, Marrakech, Maroc
‡Institut de Mathmatiques et Modlisation, Université de Montpellier 2, Montpellier, France, Jean-Michel.Marin@univ-montp2.fr
§Université Paris-Dauphine, CEREMADE, Institut Universitaire de France & CREST, Paris, France, Christian.Robert@ceremade.dauphine.fr

 DOI:10.1214/12-BA716

To deal with poorly or ill-posed regression problems, many regularization methods have been proposed, like ridge regression (Hoerl and Kennard, 1970) and Lasso (Tibshirani, 1996). Recently the interest for frequentist regularization methods has increased and this has produced a flurry of methods (see, among others, Candes and Tao, 2007, Zou and Hastie, 2005, Zou, 2006, Yuan and Lin, 2007).

However, a natural approach for regularization is to follow the Bayesian paradigm as demonstrated recently by the Bayesian Lasso of Park and Casella (2008). The amount of literature on Bayesian variable selection is quite enormous (a small subset of which is, for instance, Mitchell and Beauchamp, 1988, George and McCulloch, 1993, Chipman, 1996, Smith and Kohn, 1996, George and McCulloch, 1997, Dupuis and Robert, 2003, Brown and Vannucci, 1998, Philips and Guttman, 1998, George, 2000, Kohn et al., 2001, Nott and Green, 2004, Schneider and Corcoran, 2004, Casella and Moreno, 2006, Cui and George, 2008, Liang et al., 2008, Bottolo and Richardson, 2010). The number of approaches and scenarios that have been advanced to undertake the selection of the most relevant variables given a set of observations is quite large, presumably due to the vague decisional setting induced by the question *Which variables do matter?* Such a variety of resolutions signals a lack of agreement between the actors in the field. Most of the solutions, including Liang et al. (2008) and Bottolo and Richardson (2010), focus on the use of the g-prior, introduced by Zellner (1986). While this prior has a long history and while it reduces the prior input to a single integer, g, the influence of this remaining prior factor is long-lasting and large values of g are no guarantee of negligible effects, in connection with the Bartlett or Lindley–Jeffreys paradoxes (Bartlett, 1957, Lindley, 1957, Robert, 1993), as illustrated for instance in Celeux et al. (2006) or Marin and Robert (2007). In order to alleviate this influence, some empirical Bayes [Cui and George (2008)] and hierarchical Bayes [Zellner and Siow (1980), Celeux et al. (2006), Marin and Robert (2007), Liang et al. (2008) and Bottolo and Richardson (2010)] solutions have been proposed. In this paper, we pay special attention to two calibration-free hierarchical Zellner g-priors. The first one is the Jeffreys prior which is not location invariant. A second one avoids this problem by only considering models with at least one variable in the model.

The purpose of our paper is to compare the frequentist and the Bayesian points of view in regularization when n remains (slightly) greater than p, we limit our attention to full rank models. This comparison is considered from both the predictive and the explicative point of views. The outcome of this study is that Bayesian methods are quite similar while dominating their frequentist counterpart.

The plan of the paper is as follows: we recall the details of Zellner's (1986) original g-prior in Section 2, and discuss therein the potential choices of g. We present hierarchical noninformative alternatives in Section 3. Section 4 compares the results of Bayesian and frequentist methods on simulated and real datasets. Section 5 concludes the paper.

2 Zellner's g-priors

Following standard notations, we introduce a variable $\boldsymbol{\gamma} \in \boldsymbol{\Gamma} = \{0,1\}^p$ that indicates which variables are active in the regression, excluding the constant vector corresponding to the intercept that is assumed to be always present in the linear regression model.

We observe $\mathbf{y}, \mathbf{x}_1, \ldots, \mathbf{x}_p \in \mathbb{R}^n$, and the model $\mathcal{M}_{\boldsymbol{\gamma}}$ is defined as the conditional distribution

$$\mathbf{y}|\mathbf{X}, \boldsymbol{\gamma}, \boldsymbol{\beta}^{\boldsymbol{\gamma}}, \sigma^2 \sim \mathcal{N}_n\left(\mathbf{X}^{\boldsymbol{\gamma}}\boldsymbol{\beta}^{\boldsymbol{\gamma}}, \sigma^2 I_n\right), \tag{1}$$

where

- $p_{\boldsymbol{\gamma}} = \sum_{i=1}^{p} \gamma_i$,
- $\mathbf{X}^{\boldsymbol{\gamma}}$ is the $(n, p_{\boldsymbol{\gamma}} + 1)$ matrix with columns made of the vector $\mathbf{1}_n$ and of the variables $\mathbf{x}_i$ for which $\gamma_i = 1$,
- $\boldsymbol{\beta}^{\boldsymbol{\gamma}} \in \mathbb{R}^{p_{\boldsymbol{\gamma}}+1}$ and $\sigma^2 \in \mathbb{R}_+^*$ are unknown parameters.

The same symbol for the parameter σ^2 is used across all models. For model $\mathcal{M}_{\boldsymbol{\gamma}}$, Zellner's g-prior is given by

$$\boldsymbol{\beta}^{\boldsymbol{\gamma}}|\mathbf{X}, \boldsymbol{\gamma}, \sigma^2 \sim \mathcal{N}_{p_{\boldsymbol{\gamma}}+1}(\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}, g_{\boldsymbol{\gamma}}\sigma^2((\mathbf{X}^{\boldsymbol{\gamma}})'\mathbf{X}^{\boldsymbol{\gamma}})^{-1}),$$

$$\pi(\sigma^2|\mathbf{X}, \boldsymbol{\gamma}) \propto \sigma^{-2}.$$

The experimenter chooses the prior expectation $\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}$ and $g_{\boldsymbol{\gamma}}$. For such a prior, we obtain the classical average between prior and observed regressors,

$$\mathbb{E}(\boldsymbol{\beta}^{\boldsymbol{\gamma}}|\mathbf{X}, \boldsymbol{\gamma}, \mathbf{y}) = \frac{g_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}} + \tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}}{g_{\boldsymbol{\gamma}} + 1}.$$

This prior is traditionally called Zellner's g-prior in the Bayesian folklore because of the use of the constant $g_{\boldsymbol{\gamma}}$ by Zellner (1986) in front of Fisher's information matrix $((\mathbf{X}^{\boldsymbol{\gamma}})'\mathbf{X}^{\boldsymbol{\gamma}})^{-1}$. Its appeal is that, by using the information matrix as a global scale,

- it avoids the specification of a whole prior covariance matrix, which would be a tremendous task;
- it allows for a specification of the constant $g_{\boldsymbol{\gamma}}$ in terms of observational units, or virtual prior pseudo-observations in the sense of de Finetti (1972).

However, a fundamental feature of the g-prior is that this prior is improper, due to the use of an infinite mass on σ^2. From a theoretical point of view, this should jeopardize the use of posterior model probabilities since these probabilities are not uniquely scaled under improper priors, because there is no way of eliminating the residual constant factor in those priors (DeGroot, 1973, Kass and Raftery, 1995, Robert, 2001). However, under the assumption that σ^2 is a parameter that has a meaning common to all models $\mathcal{M}_{\boldsymbol{\gamma}}$,

Berger et al. (1998) develop a framework that allows us to work with a single improper prior that is common to all models (see also Marin and Robert, 2007). A fundamental appeal of Zellner's g-prior in model comparison and in particular in variable selection is its simplicity, since it reduces the prior input to the sole specification of a scale parameter g.

At this stage, we need to point out that an alternative g-prior is often used (Berger et al., 1998, Fernandez et al., 2001, Liang et al., 2008, Bottolo and Richardson, 2010), by singling out the intercept parameter in the linear regression. By first assuming a centering of the covariates, i.e. $\mathbf{1}_n'\mathbf{x}_i = 0$ for all i's, the intercept α is given a flat prior while the other parameters of $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$ are associated with a corresponding g-prior. Thus, this is an alternative to model $\mathcal{M}_{\boldsymbol{\gamma}}$, which we denote by model $\mathcal{M}_{\boldsymbol{\gamma}}^{\text{inv}}$ to stress the distinctions between both representations and which is such that

$$\mathbf{y}|\mathbf{X},\boldsymbol{\gamma},\alpha,\boldsymbol{\beta}_{\text{inv}}^{\boldsymbol{\gamma}},\sigma^2 \sim \mathcal{N}_n\left(\alpha\mathbf{1}_n + \mathbf{X}_{\text{inv}}^{\boldsymbol{\gamma}}\boldsymbol{\beta}_{\text{inv}}^{\boldsymbol{\gamma}},\sigma^2 I_n\right)\,, \tag{2}$$

where

- $\mathbf{X}_{\text{inv}}^{\boldsymbol{\gamma}}$ is the $(n, p_{\boldsymbol{\gamma}})$ matrix with columns made of the variables $\mathbf{x}_i$ for which $\gamma_i = 1$,
- $\alpha \in \mathbb{R}$, $\boldsymbol{\beta}_{\text{inv}}^{\boldsymbol{\gamma}} \in \mathbb{R}^{p_{\boldsymbol{\gamma}}}$ and $\sigma^2 \in \mathbb{R}_+^*$ are unknown parameters.

The parameters σ^2 and α are denoted the same way across all models and rely on the same prior. Namely, for model $\mathcal{M}_{\boldsymbol{\gamma}}^{\text{inv}}$, the corresponding Zellner's g-prior is given by

$$\boldsymbol{\beta}_{\text{inv}}^{\boldsymbol{\gamma}}|\mathbf{X},\boldsymbol{\gamma},\sigma^2 \sim \mathcal{N}_{p_{\boldsymbol{\gamma}}}(\tilde{\boldsymbol{\beta}}_{\text{inv}}^{\boldsymbol{\gamma}}, g_{\boldsymbol{\gamma}}\sigma^2((\mathbf{X}_{\text{inv}}^{\boldsymbol{\gamma}})'\mathbf{X}_{\text{inv}}^{\boldsymbol{\gamma}})^{-1})\,,$$

$$\pi(\alpha,\sigma^2|\mathbf{X},\boldsymbol{\gamma}) \propto \sigma^{-2}\,.$$

In that case, we obtain

$$\mathbb{E}(\boldsymbol{\beta}_{\text{inv}}^{\boldsymbol{\gamma}}|\mathbf{X},\boldsymbol{\gamma},\mathbf{y}) = \frac{g_{\boldsymbol{\gamma}}\hat{\boldsymbol{\beta}}_{\text{inv}}^{\boldsymbol{\gamma}} + \tilde{\boldsymbol{\beta}}_{\text{inv}}^{\boldsymbol{\gamma}}}{g_{\boldsymbol{\gamma}}+1}\,,$$

and

$$\mathbb{E}(\alpha|\mathbf{X},\boldsymbol{\gamma},\mathbf{y}) = \bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n} y_i\,.$$

For models $\mathcal{M}_{\boldsymbol{\gamma}}$ and $\mathcal{M}_{\boldsymbol{\gamma}}^{\text{inv}}$, in a noninformative setting, we can for instance choose $\tilde{\boldsymbol{\beta}}^{\boldsymbol{\gamma}} = 0_{p_{\boldsymbol{\gamma}}+1}$ or $\tilde{\boldsymbol{\beta}}_{\text{inv}}^{\boldsymbol{\gamma}} = 0_{p_{\boldsymbol{\gamma}}}$ and $g_{\boldsymbol{\gamma}}$ large. However, as pointed out in Marin and Robert (2007, Chapter 3) among others, there is a lasting influence of $g_{\boldsymbol{\gamma}}$ over the resulting inference and it is impossible to "let $g_{\boldsymbol{\gamma}}$ go to infinity" to eliminate this influence, because of the Bartlett and Lindley-Jeffreys (Bartlett, 1957, Lindley, 1957, Robert, 1993) paradoxes that an infinite value of $g_{\boldsymbol{\gamma}}$ ends up selecting the null model, regardless of the information brought by the data. For this reason, data-dependent versions of $g_{\boldsymbol{\gamma}}$ have been proposed with various degrees of justification:

- Kass and Wasserman (1995) use $g_{\boldsymbol{\gamma}} = n$ so that the amount of information about the parameters contained in the prior equals the amount of information brought by one observation. As shown by Foster and George (1994), for n large enough this perspective is very close to using the Schwarz (Kass and Wasserman, 1995) or BIC criterion in that the log-posterior corresponding to $g = n$ is equal to the penalized log-likelihood of this criterion.
- Foster and George (1994) and George and Foster (2000) propose $g_{\boldsymbol{\gamma}} = p_{\boldsymbol{\gamma}}^2$, in connection with the Risk Inflation Criterion (RIC) that penalizes the regression sum of squares.
- Fernandez et al. (2001) gather both perspectives in $g_{\boldsymbol{\gamma}} = \max(n, p_{\boldsymbol{\gamma}}^2)$ as a conservative bridge between BIC and RIC, a choice that they christened "benchmark prior".
- George and Foster (2000) and Cui and George (2008) resort to empirical Bayes techniques.

These solutions, while commendable since based on asymptotic properties (see in particular Fernandez et al., 2001 for consistency results), are nonetheless unsatisfactory in that they depend on the sample size and involve a degree of arbitrariness.

3 Mixtures of g-priors

The most natural Bayesian approach to solving the uncertainty on the parameter $g_{\boldsymbol{\gamma}} = g$ is to put a hyperprior on this parameter:

- This was implicitly proposed by Zellner and Siow (1980) since those authors introduced Cauchy priors on the $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$'s since this corresponds to a g-prior augmented by a Gamma $\mathcal{G}a(1/2, n/2)$ prior on g^{-1}.
- For model $\mathcal{M}_{\boldsymbol{\gamma}}^{\text{inv}}$, Liang et al. (2008), Cui and George (2008) and Bottolo and Richardson (2010) use

$$\boldsymbol{\beta}_{\text{inv}}^{\boldsymbol{\gamma}} | \mathbf{X}, \boldsymbol{\gamma}, \sigma^2 \sim \mathcal{N}_{p_{\boldsymbol{\gamma}}}(0_{p_{\boldsymbol{\gamma}}}, g\sigma^2((\mathbf{X}_{\text{inv}}^{\boldsymbol{\gamma}})'\mathbf{X}_{\text{inv}}^{\boldsymbol{\gamma}})^{-1})$$

  and a hyperprior of the form

$$\pi(\alpha, \sigma^2, g | \mathbf{X}, \boldsymbol{\gamma}) \propto (1+g)^{-a/2}\sigma^{-2}\,,$$

  with $a > 2$. This constraint on a is due to the fact that the hyperprior must be proper, in connection with the separate processing of the intercept α and the use of a Lebesgue measure as a prior on α. We note that a needs to be specified, $a = 3$ and $a = 4$ being the solutions favored by Liang et al. (2008).
- For model $\mathcal{M}_{\boldsymbol{\gamma}}$, Celeux et al. (2006) and Marin and Robert (2007) used

$$\boldsymbol{\beta}^{\boldsymbol{\gamma}} | \mathbf{X}, \boldsymbol{\gamma}, \sigma^2 \sim \mathcal{N}_{p_{\boldsymbol{\gamma}}+1}(0_{p_{\boldsymbol{\gamma}}+1}, g\sigma^2((\mathbf{X}^{\boldsymbol{\gamma}})'\mathbf{X}^{\boldsymbol{\gamma}})^{-1})$$

and a hyperprior of the form

$$\pi(\sigma^2, g|\mathbf{X}) \propto \sigma^{-2} g^{-1} \mathbb{I}_{\mathbb{N}^*}(g)\,.$$

The choice of the integer support is mostly computational, while the Jeffreys-like $1/g$ shape is not justified, but the authors claim that it is appropriate for a scale parameter.

For model $\mathcal{M}_{\boldsymbol{\gamma}}$ a more convincing modelling is possible since the Jeffreys prior is available. Indeed, if

$$\boldsymbol{\beta}^{\boldsymbol{\gamma}}|\mathbf{X}, \boldsymbol{\gamma}, \sigma^2 \sim \mathcal{N}_{p_{\boldsymbol{\gamma}}+1}(0_{p_{\boldsymbol{\gamma}}+1}, g\sigma^2((\mathbf{X}^{\boldsymbol{\gamma}})'\mathbf{X}^{\boldsymbol{\gamma}})^{-1})\,,$$

then

$$\mathbf{y}|\mathbf{X}, \boldsymbol{\gamma}, g, \sigma^2 \sim \mathcal{N}_{p_{\boldsymbol{\gamma}}+1}\left(0_n, \sigma^2\left[\mathbf{I}_n - \frac{g}{g+1}\mathbf{P}_{\boldsymbol{\gamma}}\right]^{-1}\right)\,,$$

where $\mathbf{I}_n$ is the (n, n) identity matrix and $\mathbf{P}_{\boldsymbol{\gamma}}$ is the orthogonal projector on the linear subspace spanned by the columns of $\mathbf{X}^{\boldsymbol{\gamma}}$. Since, the Fisher information matrix is

$$\mathfrak{I}(\sigma^2, g) = \left(\frac{1}{2}\right)\begin{bmatrix} n/\sigma^4 & (p_{\boldsymbol{\gamma}}+1)/(\sigma^2(g+1)) \\ (p_{\boldsymbol{\gamma}}+1)/(\sigma^2(g+1)) & (p_{\boldsymbol{\gamma}}+1)/(g+1)^2 \end{bmatrix},$$

the corresponding Jeffreys prior on (σ^2, g) is

$$\pi(\sigma^2, g|\mathbf{X}) \propto \sigma^{-2}(g+1)^{-1}\,.$$

Note that, for model $\mathcal{M}_{\boldsymbol{\gamma}}^{\text{inv}}$, Liang et al. (2008) discuss the choice of $a = 2$ and then $\pi(\alpha, \sigma^2, g|\mathbf{X}, \boldsymbol{\gamma}) \propto (1+g)^{-1}\sigma^{-2}$ as leading to the reference prior and Jeffreys prior, presumably also under the marginal model after integrating out $\boldsymbol{\beta}^{\boldsymbol{\gamma}}$, although details are not given.

For such a prior modelling, there exists a closed-form representation for posterior quantities in that

$$\pi(\boldsymbol{\gamma}, g|\mathbf{X}, \mathbf{y}) \propto (g+1)^{n/2-(p_{\boldsymbol{\gamma}}+1)/2-1}(1+g(1-\mathbf{y}'\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{y}/\mathbf{y}'\mathbf{y}))^{-n/2}$$

and

$$\pi(\boldsymbol{\gamma}|\mathbf{X}, \mathbf{y}) \propto \frac{{}_2F_1(n/2, 1; (p_{\boldsymbol{\gamma}}+3)/2; \mathbf{y}'\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{y}/\mathbf{y}'\mathbf{y})}{p_{\boldsymbol{\gamma}}+1}\,, \tag{3}$$

where ${}_2F_1$ is the Gaussian hypergeometric function (Butler and Wood, 2002). We can thus proceed to undertake Bayesian variable selection without resorting at all to numerical methods (Marin and Robert, 2007). Moreover, the shrinkage factor due to the Bayesian modelling can also be expressed in closed form as

$$\begin{aligned}\mathbb{E}(g/(g+1)|\mathbf{X}, \boldsymbol{\gamma}, \mathbf{y}) &= \frac{\displaystyle\int_0^\infty g(g+1)^{n/2-(p_{\boldsymbol{\gamma}}+1)/2-2}(1+g(1-\mathbf{y}'\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{y}/\mathbf{y}'\mathbf{y}))^{-n/2}\mathrm{d}g}{\displaystyle\int_0^\infty (g+1)^{n/2-(p_{\boldsymbol{\gamma}}+1)/2-1}(1+g(1-\mathbf{y}'\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{y}/\mathbf{y}'\mathbf{y}))^{-n/2}\mathrm{d}g} \\ &= \frac{2\,{}_2F_1(n/2, 2; (p_{\boldsymbol{\gamma}}+3)/2+1; \mathbf{y}'\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{y}/\mathbf{y}'\mathbf{y})}{(p_{\boldsymbol{\gamma}}+3)\,{}_2F_1(n/2, 1; (p_{\boldsymbol{\gamma}}+3)/2; \mathbf{y}'\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{y}/\mathbf{y}'\mathbf{y})}\,.\end{aligned}$$

This obviously leads to straightforward representations for Bayes estimates. If $\mathbf{X}_{\text{new}}$ is a $q \times p$ matrix containing q *new* values of the explanatory variables for which we would like to predict the corresponding response $\mathbf{y}_{\text{new}}$, the Bayesian predictor of $\mathbf{y}_{\text{new}}$ is given by

$$\begin{aligned}\hat{\mathbf{y}}^{\boldsymbol{\gamma}}_{\text{new}} &= \mathbb{E}\left[\mathbf{y}_{\text{new}}|\mathbf{X}_{\text{new}}, \mathbf{X}, \boldsymbol{\gamma}, \mathbf{y}\right] \\ &= 2\,\frac{{}_2F_1(n/2, 2; (p_{\boldsymbol{\gamma}}+3)/2+1; \mathbf{y}'\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{y}/\mathbf{y}'\mathbf{y})}{(p_{\boldsymbol{\gamma}}+3)\,{}_2F_1(n/2, 1; (p_{\boldsymbol{\gamma}}+3)/2; \mathbf{y}'\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{y}/\mathbf{y}'\mathbf{y})}\,\mathbf{X}_{\text{new}}\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}\,.\end{aligned}$$

Similarly, the Bayesian model averaging predictor of $\mathbf{y}_{\text{new}}$ is given by

$$\begin{aligned}\hat{\mathbf{y}}_{\text{new}} &= \mathbb{E}\left[\mathbf{y}_{\text{new}}|\mathbf{X}_{\text{new}}, \mathbf{X}, \mathbf{y}\right] \\ &= 2\,\frac{\sum_{\boldsymbol{\gamma}\in\Gamma}\,{}_2F_1(\frac{n}{2}, 2; (p_{\boldsymbol{\gamma}}+3)/2+1; \mathbf{y}'\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{y}/\mathbf{y}'\mathbf{y})/\left[(p_{\boldsymbol{\gamma}}+1)(p_{\boldsymbol{\gamma}}+3)\right]}{\sum_{\boldsymbol{\gamma}\in\Gamma}\,{}_2F_1(\frac{n}{2}, 1; (p_{\boldsymbol{\gamma}}+3)/2; \mathbf{y}'\mathbf{P}_{\boldsymbol{\gamma}}\mathbf{y}/\mathbf{y}'\mathbf{y})/(p_{\boldsymbol{\gamma}}+1)}\,\mathbf{X}_{\text{new}}\hat{\boldsymbol{\beta}}^{\boldsymbol{\gamma}}\,.\end{aligned} \tag{4}$$

This numerical simplification in the derivation of Bayesian estimates and predictors is found in Liang et al. (2008) and exploited further in Bottolo and Richardson (2010). Note also that Guo and Speckman (2009) have furthermore established the consistency of the Bayes factors based on such priors.

In contrast with this proposal, the prior of Liang et al. (2008) depends on a tuning parameter a. Despite that, there also exist arguments to support this prior modelling, including the important issue of invariance under location-scale transforms. As seen in the above formulae, the Jeffreys prior associated with model $\mathcal{M}_{\boldsymbol{\gamma}}$ ensures scale invariance but not location invariance. In order to ensure location invariance for model $\mathcal{M}_{\boldsymbol{\gamma}}$, it would be necessary to center the observation variable y as well as the dependent variables X. Obviously, this centering of the data is completely unjustified from a Bayesian perspective and further it creates artificial correlations between observations. However it could be argued that the lack of location invariance only pertains to quite specific and somehow artificial situations and that it is negligible in most situations. We will return to this point in the comparison section.

A location scale alternative consists in using the prior of Liang et al. (2008) with $a = 2$ and excluding the null model from the competitors. This prior leads to the model posterior probability

$$\pi(\boldsymbol{\gamma}|\mathbf{X}, \mathbf{y}) \propto \frac{{}_2F_1((n-1)/2, 1; (p_{\boldsymbol{\gamma}}+2)/2; (\mathbf{y}-\bar{\mathbf{y}})'\mathbf{P}_{\boldsymbol{\gamma}}(\mathbf{y}-\bar{\mathbf{y}})/(\mathbf{y}-\bar{\mathbf{y}})'(\mathbf{y}-\bar{\mathbf{y}}))}{p_{\boldsymbol{\gamma}}}\,. \tag{5}$$

Equations (3) and (5) are similar. However, in the last part of (5), $\mathbf{y}$ is centered, ensuring the location invariance of the selection procedure.

4 Numerical comparisons

We present here the results of numerical experiments aiming at comparing the behavior of Bayesian variable selection and of some (non-Bayesian) popular regularization methods in regression, when considered from a variable selection point of view: The regularization methods that we consider are the Lasso, the Dantzig selector, and elastic net, described in Section 4.1. The Bayesian variable selection procedures we consider adopt different strategies for selecting the hyperparameter g in Zellner's g-priors: We include in this comparison the intrinsic prior (Casella and Moreno, 2006) which is another default objective prior for the non informative setting that does not require any tuning parameters and is also invariant under location and scale changes. All procedures under comparison are described in Table 1. We have also included in this comparison the highly standard AIC and BIC penalized likelihood criteria. Moreover, we will refer to the performances of an ORACLE procedure that assumes the true model is known and that estimates the regression coefficients with the least squares method.

AIC	Akaike Information Criterion
BIC	Bayesian Information Criterion
BRIC	g prior with $g = \max(n, p^2)$ (Fernandez et al., 2001)
EB-L	Local EB estimate of g in g-prior (Cui and George, 2008)
EB-G	Global EB estimate of g in g-prior (Cui and George, 2008)
ZS-N	Base model in Bayes factor taken as the null model (Liang et al., 2008)
ZS-F	Base model in Bayes factor taken as the full model (Liang et al., 2008)
OVS	Objective variable selection using the intrinsic prior (Casella and Moreno, 2006)
HG-3	Hyper-g prior with $a = 3$ (Liang et al., 2008)
HG-4	Hyper-g prior with $a = 4$ (Liang et al., 2008)
HG-2	Hyper-g prior with $a = 2$ (Liang et al., 2008), null model excluded
NIMS	Jeffreys prior on the non-invariant model
LASSO	Lasso (Tibshirani, 1996)
DZ	The Dantzig Selector (Candes and Tao, 2007)
ENET	The elastic-net (Zou and Hastie, 2005)

Table 1: Acronyms and descriptions for the variable selection methods compared in the numerical experiment. The blocks separate the methods by their nature.

4.1 Regularization methods

***1) The Lasso*:** Introduced by Tibshirani (1996), the Lasso is a shrinkage method for linear regression. It is defined as the solution to the following ℓ_1 penalized least squares optimization problem

$$\hat{\boldsymbol{\beta}}_{\text{Lasso}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - X\boldsymbol{\beta}||_2^2 + \lambda \sum_{j=1}^{p} |\beta_j|,$$

where λ is a positive tuning parameter.

***2) The Dantzig Selector*:** Candes and Tao (2007) introduced the Dantzig Selector as an alternative to the Lasso. The Dantzig Selector is the solution to the optimization problem

$$\min_{\beta\in\mathbb{R}^p} \|\beta\|_1 \quad \text{subject to} \quad \|\mathbf{X}^t(\mathbf{y}-\mathbf{X}\beta)\|_\infty \leq \lambda,$$

where λ is a positive tuning parameter. The constraint $\|\mathbf{X}^t(\mathbf{y}-\mathbf{X}\beta)\|_\infty \leq \lambda$ can be viewed as a relaxation of the normal equation in the classical linear regression.

***3) The Elastic Net (Enet)*:** The Lasso has at least two limitations: a) Lasso does not encourage grouped selection in the presence of high correlated covariates and b) for the $p > n$ case Lasso can select at most n covariates. To overcome these limitations, Zou and Hastie (2005) proposed an elastic net that combines both ridge ℓ_2 and Lasso ℓ_1 penalties, i.e.

$$\hat{\boldsymbol{\beta}}_{\text{Enet}} = \arg\min_{\boldsymbol{\beta}} ||\mathbf{y} - X\boldsymbol{\beta}||_2^2 + \lambda\sum_{j=1}^{p}|\beta_j| + \mu\sum_{j=1}^{p}\beta_j^2,$$

where λ and μ are two positive tuning parameters.

4.2 Numerical experiments on simulated datasets

We have designed six different simulated datasets as benchmarks chosen as follows:

1. Example 1 (**sparse uncorrelated design**) corresponds to an uncorrelated covariate setting ($\rho = 0$), with $p = 10$ predictors and where the components of $\mathbf{x}_i$ ($i = 1, \ldots, 10$) are iid $\mathcal{N}_1(0, 1)$ realizations. The response is simulated as
$$\mathbf{y} \sim \mathcal{N}_n(2 + \mathbf{x}_2 + 2\mathbf{x}_3 - 2\mathbf{x}_6 - 1.5\mathbf{x}_7, \mathbf{I}_n)\,.$$
2. Example 2 (**sparse correlated design**) corresponds to a correlated case ($\rho = 0.9$), with $p = 10$ predictors and $\mathbf{x}_i = (\mathbf{z}_i + 3\mathbf{z}_{11})/\sqrt{10}$, for $i = 1, 2$, $\mathbf{x}_i = (\mathbf{z}_i + 3\mathbf{z}_{12})/\sqrt{10}$, for $i = 3, 4, 5$, and $\mathbf{x}_i = (\mathbf{z}_i + 3\mathbf{z}_{13})/\sqrt{10}$ for $i = 6, \ldots, 10$, the components of $\mathbf{z}_i$ ($i = 1, \ldots, 13$) being iid $\mathcal{N}_1(0, 1)$ realizations. The use of common terms in the $\mathbf{x}_i$'s obviously induces a correlation among those $\mathbf{x}_i$'s: the correlation between variables $\mathbf{x}_1$ and $\mathbf{x}_2$ is 0.9, as for the variables ($\mathbf{x}_3$, $\mathbf{x}_4$ and $\mathbf{x}_5$), and for the variables ($\mathbf{x}_6$, $\mathbf{x}_7$, $\mathbf{x}_8$, $\mathbf{x}_9$ and $\mathbf{x}_{10}$). There is no correlation between those three groups of variables. The response is simulated as
$$\mathbf{y} \sim \mathcal{N}_n(2 + \mathbf{x}_2 + 2\mathbf{x}_3 - 2\mathbf{x}_6 - 1.5\mathbf{x}_7, \mathbf{I}_n)\,.$$
3. Example 3 (**sparse noisy correlated design**) involves $p = 8$ predictors. Those variables are generated using a multivariate Gaussian distribution with correlations
$$\rho(\mathbf{x}_i, \mathbf{x}_j) = 0.5^{|i-j|}\,.$$
The response is simulated as
$$\mathbf{y} \sim \mathcal{N}_n(3\mathbf{x}_1 + 1.5\mathbf{x}_2 + 2\mathbf{x}_5, 9\mathbf{I}_n)\,.$$

4. Example 4 (**saturated correlated design**) is the same as Example 4, except that the response is simulated as

$$\mathbf{y} \sim \mathcal{N}_n\left(0.85\sum_{i=1}^{8}\mathbf{x}_i, \mathbf{I}_n\right).$$

5. Example 5 involves $p = 9$ predictors. Those variables are generated using a multivariate Gaussian distribution with correlations

$$\rho(\mathbf{x}_i, \mathbf{x}_j) = 0.7^{|i-j|}.$$

   The response is simulated as

$$\mathbf{y} \sim \mathcal{N}_n(2\mathbf{x}_2 - 3\mathbf{x}_4, \mathbf{I}_n).$$

6. Example 6 (**null model**) involves $p = 8$ predictors. Those variables are generated using a multivariate Gaussian distribution with correlations

$$\rho(\mathbf{x}_i, \mathbf{x}_j) = 0.5^{|i-j|}.$$

   The response is simulated as

$$\mathbf{y} \sim \mathcal{N}_n(2, 4\mathbf{I}_n).$$

Each dataset consists of a training set of size $n = 15$, on which the regression model has been fitted and a test set T of size $n_T = 200$ for assessing performances. Tuning parameters in the Lasso, the Dantzig selector (DZ), and the elastic net (ENET) have been selected by minimizing the cross-validation prediction error through leave-one-out. For each example, 100 independent datasets have been simulated. We use three measures of performances:

1. The root mean squared error (MSE)
$$\text{MSE}_y = \sqrt{\textstyle\sum_{i=1}^{n_T}(y_i - \hat{y}_i)^2/n_T},$$
   $\hat{y}_i$ being the prediction of y_i in the test set;
2. HITS: the number of correctly identified influential variables;
3. FP (False Positives): the number of non-influential variables declared as influential.

Using those six different datasets as benchmarks, we compare the variable selection methods listed in Table 1. The performances of the above selection methods are summarized in Tables 2–13, presented in the Appendix. In the Bayesian approaches, the set of variables is naturally selected according to the maximum posterior probability $\pi(\boldsymbol{\gamma}|\mathbf{X}, \mathbf{y})$ and the predictive is obtained via the Bayesian model averaging predictors.

In this numerical experiment, the Bayesian procedures are clearly much more parsimonious than the regularization procedures in that they almost always avoid overfitting. In all examples, the false positive rate FP is smaller for the Bayesian solutions than for the regularization methods. Except for the ZS-F and OVS scenarios which behave slightly worse than the others, all the Bayesian procedures tested here produce the same selection of predictors. It seems that ZS-F has a slight tendency to select too many variables. The performances of OVS are somewhat disappointing and this procedure seems to have a tendency to be too parsimonious. From a predictive viewpoint, computing the MSE by model averaging, Bayesian approaches also perform better than regularization approaches except for the saturated correlated example (Example 4). We further note that the classical selection procedures based on AIC and BIC do not easily reject variables and are thus slightly worse than Bayesian and regularization procedures (a fact not surprising for AIC). In all examples, the NIMS and HG-2 approaches lead to optimal performances in that they select the right covariates and only the right covariates, while achieving close to the minimal root mean squared error compared with all the other Bayesian solutions we considered. They also do almost systematically better than BIC and AIC.

A global remark about this comparison is that all Bayesian procedures have a very similar MSE and thus that they all correspond to the same regularization effect, except for OVS which does systematically worse. However it is important to notice that the MSE for OVS has not been computed by model averaging, but by using the best model. Otherwise, it would be hazardous to recommend one of the priors from those simulations since there is no sensitive difference between them from both selection and prediction points of view.

Translating the data Since NIMS is not location invariant, it is important to measure the impact of adding a constant to all observations. As stressed by a reviewer, when this constant goes to infinity, keeping n fixed, the last argument of ${}_2F_1$ in (3) goes to one for all models. Thus if the empirical mean is large relative to the regression sum of squares, the data end up having little input in distinguishing between models. In order to measure this possible negative impact of adding a large constant, we replace in Example 1 $\mathbf{y}$ by $\mathbf{y} = \mathbf{y} + 10^k$ RSS (Regression Sum of Squares) for $k \in \{1, 2, 3\}$. The results derived from NIMS criterion are summarized in Tables 14 and 15, presented in the Appendix. As predicted, the NIMS criterion tends to choose the null model as k increases and the null model with no variable is always selected when $k = 3$. Therefore some prior assumption must be made about the magnitude of the intercept when using NIMS. Otherwise, the criterion is over-parsimonious. If this is a possible case, we suggest using instead the HG-2 approach.

4.3 Real datasets

Two datasets considered in this section are associated with a moderate number of variables compared to the number of observations.

Body fat dataset The body fat dataset has been first used by Penrose et al. (1985). The corresponding study aims at estimating the percentage of body fat from various body circumference measurements observed on 252 men. The thirteen regressor variables are:

1. age,
2. weight (lbs),
3. height (inches),
4. neck circumference,
5. chest circumference,
6. abdomen to circumference,
7. hip circumference,
8. thigh circumference,
9. knee circumference,
10. ankle circumference,
11. biceps (extended) circumference,
12. forearm circumference,
13. wrist circumference.

In order to investigate the performances of the different methods, a dataset from Penrose et al. (1985) has been split 25 times into a training set of 151 observations and a test set of 101 observations. Tuning parameters for the frequentist regularization methods have been chosen by minimizing the (ten fold) cross-validated prediction error.

For this dataset, the Bayesian procedures we investigated are much more parsimonious than the standard regularization procedures, as shown in Table 16 in the Appendix. There is no variability in the prediction MSE. (We stress that MSEs are computed by model averaging for the Bayesian procedures.) As in the simulation experiment, all Bayesian approaches are highly similar, except for ZS-F which remains more open to incorporating the last two covariates (see Table 17 in the Appendix).

Ozone data This second benchmark dataset is taken from Breiman and Friedman (1985) and consists of daily measurements of the maximum ozone concentration and of eight meteorological variables near Los Angeles. Those variables are:

1. the daily ozone concentration (maximum one hour average, parts per million) at Upland, California (CA), which is the response variable;
2. the Vandenburg 500 millibar pressure height (m);
3. the wind speed (mph) at Los Angeles International Airport (LAX);
4. the humidity (percent) at LAX;
5. the Sandburg Air Force Base temperature (F^o);
6. the inversion base height at LAX;
7. the inversion base temperature at LAX;

8. the Daggett Pressure gradient (mm Hg) from LAX to Daggett, CA;
9. the visibility (miles) at LAX.

The original Ozone database contains 366 observations, of which 203 are complete. Our study is made just on the complete observations. We split this dataset 25 times into a training set of 101 observations and a test set of 102 observations.

For this dataset, as shown by Tables 18 and 19 in the Appendix, all Bayesian approaches, as well as AIC and BIC, select about three variables, while the regularization methods opt for five. The MSE differences between all procedures are negligible. (This lack of significant differences in the MSEs is also exhibited through the boxplots of Figure 1.)

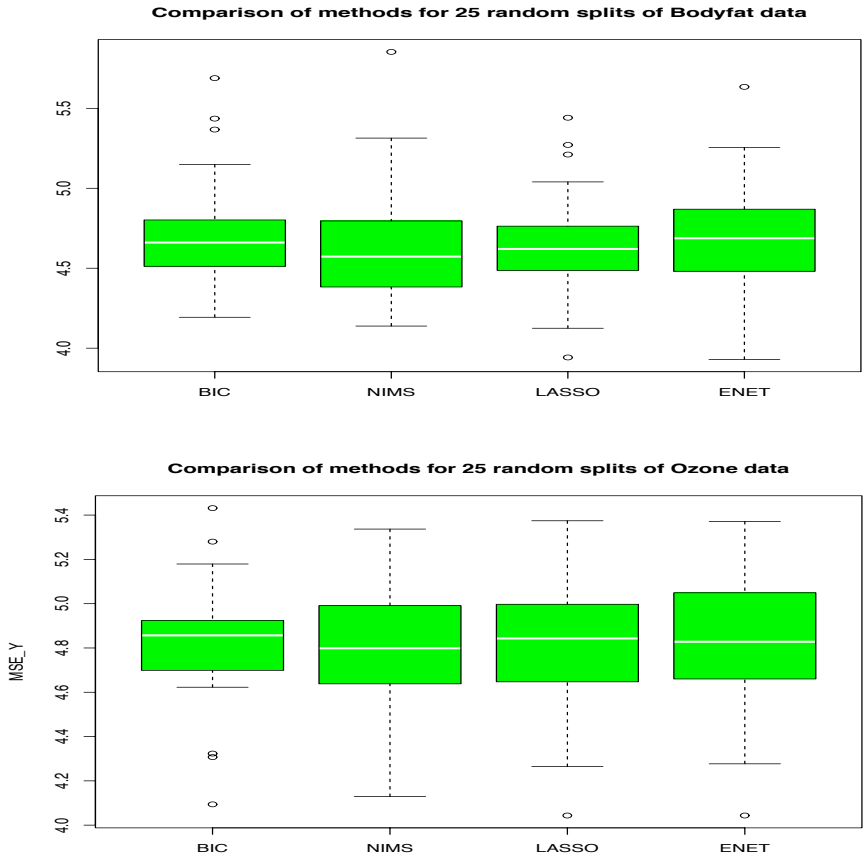

Figure 1: Body fat and Ozone datasets: variability of the root mean squared errors over 25 random splits for BIC, NIMS, LASSO and ENET methods.

5 Conclusion

In this numerical study, we have compared Bayesian variable selection methods with regularization methods in a poorly informative setting. From a variable selection point of view, it appears that the Bayesian methods are more parsimonious and more relevant than the regularization methods. From a predictive point of view, there is no significant difference between both approaches. Regularization methods could however be expected to perform better from this latter point of view since they minimize a cross-validated prediction error. But, owing to model averaging, efficiency, Bayesian methods provide competitive MSE's.

An additional appeal of this study is to single-out and to assess two calibration-free prior models (NIMS and HG-2). They both appear as valuable competitors when compared with earlier Bayesian approaches. However, both methods have a clear drawback (NIMS is not location invariant and HG-2 excludes the null model). Nonetheless our series of examples shows that they provide an acceptable objective Bayesian solution for Bayesian variable selection and regularization in linear models.

A limitation of this study on our objective Bayesian approach is that we do not consider large dimensions as in Bottolo and Richardson (2010), which require different computational tools to face the enormous number of potential models. This difficulty is obviously faced by all Bayesian solutions considered in this paper and is not an issue in terms of the validity of the prior modelling.

References

Bartlett, M. (1957). A comment on D.V. Lindley's statistical paradox. *Biometrika*, 44:533–534. 478, 480

Berger, J., Pericchi, L., and Varshavsky, I. (1998). Bayes factors and marginal distributions in invariant situations. *Sankhya A*, 60:307–321. 480

Bottolo, L. and Richardson, S. (2010). Evolutionary stochastic search for Bayesian model exploration. *Bayesian Analysis*, 5(3):583–618. 478, 480, 481, 483, 490

Breiman, L. and Friedman, J.H. (1985). Estimating optimal transformations for multiple regression and correlation. *Journal of the American Statistical Association*, 85(391):580–598. 488

Brown, J. and Vannucci, M. (1998). Multivariate Bayesian variable selection and prediction. *Journal of the Royal Statistical Society Series B*, 60(3):627–641. 478

Butler, R. and Wood, A. (2002). Laplace approximations for hypergeometric functions with matrix arguments. *Annals of Statistics*, 30:1155–1177. 482

Candes, E. and Tao, T. (2007). The Dantzig Selector: statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351. 478, 484

Casella, G. and Moreno, E. (2006). Objective Bayesian variable selection. *Journal of the American Statistical Association*, 101(473):157–167. 478, 484

Celeux, G., Marin, J.-M., and Robert, C. (2006). Sélection bayésienne de variables en régression linéaire. *Journal de la Société Française de Statistique*, 147(1):59–79. 478, 481

Chipman, H. (1996). Bayesian variable selection with related predictors. *Canadian Journal of Statistics*, 1:17–36. 478

Cui, W. and George, E. (2008). Empirical Bayes vs. fully Bayes variable selection. *Journal of Statistical Planning and Inference*, 138:888–900. 478, 481, 484

de Finetti, B. (1972). *Probability, Induction and Statistics*. John Wiley, New York. 479

DeGroot, M. (1973). Doing what comes naturally: Interpreting a tail area as a posterior probability or as a likelihood ratio. *Journal of the American Statistical Association*, 68:966–969. 479

Dupuis, J. and Robert, C. (2003). Bayesian variable selection in qualitative models by Kullback-Leibler projections. *Journal of Statistical Planning and Inference*, 111:77–94. 478

Fernandez, C., Ley, E., and Steel, M. (2001). Benchmark priors for Bayesian model averaging. *Journal of Econometrics*, 100:381–427. 480, 481, 484

Foster, D. and George, E. (1994). The risk inflation criterion for multiple regression. *Annals of Statistics*, 22:1947–1975. 481

George, E. (2000). The variable selection problem. *Journal of the American Statistical Association*, 95:1304–1308. 478

George, E. and Foster, D. (2000). Calibration and empirical Bayes variable selection. *Biometrika*, 87(4):731–747. 481

George, E. and McCulloch, R. (1993). Variable selection via Gibbbs sampling. *Journal of the American Statistical Association*, 88:881–889. 478

George, E. and McCulloch, R. (1997). Approaches to Bayesian variable selection. *Statistica Sinica*, 7:339–373. 478

Guo, R. and Speckman, P. (2009). Bayes factor consistency in linear models. In *The 2009 International Workshop on Objective Bayes Methodology, Philadelphia, June 5-9, 2009*. URL http://www-stat.wharton.upenn.edu/statweb/Conference/OBayes09/AbstractPapers/speckman.pdf. 483

Hoerl, A. and Kennard, R. (1970). Ridge regression: biased estimation for non orthogonal problems. *Technometrics*, 12:55–67. 478

Kass, R. and Raftery, A. (1995). Bayes factor and model uncertainty. *Journal of the American Statistical Association*, 90:773–795. 479

Kass, R. and Wasserman, L. (1995). A reference Bayesian test for nested hypotheses and its relationship to the Schwarz criterion. *Journal of the American Statistical Association*, 90:928–934. 481

Kohn, R., Smith, M., and Chan, D. (2001). Nonparametric regression using linear combinations of basis functions. *Statistics and Computing*, 11:313–322. 478

Liang, F., Paulo, R., Molina, G., Clyde, M., and Berger, J. (2008). Mixtures of g-priors for Bayesian variable selection. *Journal of the American Statistical Association*, 103(481):410–423. 477, 478, 480, 481, 482, 483, 484

Lindley, D. (1957). A statistical paradox. *Biometrika*, 44:187–192. 478, 480

Marin, J. and Robert, C. (2007). *Bayesian Core: A Practical Approach to Computational Bayesian Statistics*. Springer-Verlag, New York. 478, 480, 481, 482

Mitchell, T. and Beauchamp, J. (1988). Bayesian variable selection in linear regression. *Journal of the American Statistical Association*, 83:1023–1032. 478

Nott, D. J. and Green, P. J. (2004). Bayesian variable selection and the Swendsen-Wang algorithm. *Journal of Computational and Graphical Statistics*, 13:1–17. 478

Park, T. and Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, 103(473):681–686. 478

Penrose, K., Nelson, A., and Fisher, A. (1985). Generalized body composition prediction equation for men using simple measurement techniques. *Medicine and Science in Sports and Exercise*, 17(2):189. 488

Philips, R. and Guttman, I. (1998). A new criterion for variable selection. *Statistics and Probability Letters*, 38:11–19. 478

Rao, C. (1973). *Linear Statistical Inference and its Applications*. John Wiley, New York. 477

Robert, C. (1993). A note on the Jeffreys-Lindley paradox. *Statistica Sinica*, 3:601–608. 478, 480

Robert, C. (2001). *The Bayesian Choice*. Springer-Verlag, 2 edition. 479

Schneider, U. and Corcoran, J. (2004). Perfect sampling for Bayesian variable selection in a linear regression model. *Journal of Statistical Planning and Inference*, 126:153–171. 478

Smith, M. and Kohn, R. (1996). Nonparametric regression using Bayesian variable selection. *Journal of Econometrics*, 75:317–343. 478

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B*, 58(1):267–288. 478, 484

Yuan, M. and Lin, Y. (2007). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society Series B*, 68(1):49–67. 478

Zellner, A. (1986). On assessing prior distributions and Bayesian regression analysis with g-prior distribution regression using Bayesian variable selection. In *Bayesian inference and decision techniques: Essays in Honor of Bruno De Finetti*, pages 233–243. North-Holland / Elsevier. 478, 479

Zellner, A. and Siow, A. (1980). Posterior odds ratios for selected regression hypotheses. In *Bayesian Statistics*, pages 585–603. Valencia: University Press. (Proceedings of the first Valencia meeting). 478, 481

Zou, H. (2006). The adaptive Lasso and its oracle properties. *Journal of the American Statistical Association*, 101:1418–1429. 478

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B*, 67(2):301–320. 478, 484, 485

Acknowledgments

We are grateful to the Associate Editor and one reviewer for their very valuable comments and suggestions on a previous version of this paper. They greatly contributed in improving the quality and presentation of this comparative study.

Appendix: Tables

	MSE_y	HITS	FP
ORACLE	1.24(0.02)	4.00(0.00)	0.00(0.00)
AIC	1.75(0.08)	3.94(0.02)	2.78(0.17)
BIC	1.69(0.08)	3.90(0.03)	2.29(0.17)
BRIC	1.43(0.04)	3.75(0.05)	0.65(0.09)
EB-L	1.46(0.04)	3.80(0.04)	0.66(0.09)
EB-G	1.45(0.04)	3.78(0.04)	0.65(0.09)
ZS-N	1.44(0.03)	3.78(0.04)	0.65(0.09)
ZS-F	1.49(0.03)	3.90(0.03)	1.73(0.14)
OVS	1.52(0.06)	3.63(0.06)	0.54(0.09)
HG-3	1.49(0.04)	3.75(0.05)	0.55(0.09)
HG-4	1.57(0.04)	3.65(0.05)	0.54(0.08)
HG-2	1.50(0.04)	3.75(0.05)	0.59(0.09)
NIMS	1.45(0.03)	3.75(0.05)	0.57(0.08)
LASSO	1.67(0.05)	3.89(0.03)	2.68(0.20)
DZ	1.66(0.06)	3.72(0.07)	2.41(0.15)
ENET	1.72(0.05)	3.89(0.04)	2.79(0.29)

Table 2: Example 1: Mean of MSE, HITS and FP. The numbers between parentheses are the corresponding standard errors.

Variables	1	2	3	4	5	6	7	8	9	10
AIC	0.47	0.95	1.00	0.45	0.44	0.99	1.00	0.46	0.52	0.44
BIC	0.41	0.91	1.00	0.38	0.40	0.99	1.00	0.32	0.44	0.34
BRIC	0.18	0.77	1.00	0.10	0.11	0.99	0.99	0.07	0.10	0.09
EB-L	0.17	0.81	1.00	0.11	0.11	0.99	1.00	0.07	0.11	0.09
EB-G	0.17	0.79	1.00	0.11	0.11	0.99	1.00	0.07	0.10	0.09
ZS-N	0.17	0.79	1.00	0.11	0.11	0.99	1.00	0.07	0.10	0.09
ZS-F	0.34	0.90	1.00	0.29	0.33	1.00	1.00	0.20	0.33	0.24
OVS	0.14	0.72	0.98	0.07	0.08	0.97	0.96	0.08	0.10	0.07
HG-3	0.17	0.77	1.00	0.11	0.10	0.99	0.99	0.07	0.09	0.08
HG-4	0.15	0.77	1.00	0.10	0.08	0.99	0.99	0.07	0.08	0.07
HG-2	0.10	0.83	0.99	0.07	0.16	0.98	0.95	0.13	0.06	0.07
NIMS	0.15	0.77	1.00	0.09	0.09	0.99	0.99	0.06	0.10	0.08
LASSO	0.49	0.91	1.00	0.41	0.45	0.98	1.00	0.49	0.47	0.37
DZ	0.42	0.84	0.96	0.41	0.47	0.97	0.95	0.38	0.37	0.36
ENET	0.45	0.93	1.00	0.45	0.43	0.99	0.97	0.52	0.44	0.50

Table 3: Example 1: Relative frequencies of the selected variables for methods under comparison.

	MSE_y	HITS	FP
ORACLE	1.19(0.01)	4.00(0.00)	0.00(0.00)
AIC	1.81(0.06)	3.12(0.08)	2.75(0.16)
BIC	1.76(0.05)	2.97(0.09)	2.39(0.16)
BRIC	1.46(0.02)	2.44(0.10)	0.99(0.10)
EB-L	1.45(0.02)	2.43(0.10)	1.03(0.10)
EB-G	1.45(0.02)	2.42(0.10)	0.95(0.10)
ZS-N	1.45(0.02)	2.43(0.10)	1.03(0.10)
ZS-F	1.42(0.02)	2.97(0.08)	2.18(0.10)
OVS	1.71(0.04)	2.16(0.11)	1.09(0.09)
HG-3	1.45(0.02)	2.32(0.11)	0.96(0.10)
HG-4	1.45(0.02)	2.35(0.10)	0.86(0.09)
HG-2	1.52(0.04)	2.35(0.10)	0.81(0.09)
NIMS	1.45(0.02)	2.42(0.10)	0.96(0.09)
LASSO	1.66(0.05)	3.35(0.09)	2.95(0.15)
DZ	1.59(0.03)	2.83(0.09)	2.23(0.10)
ENET	1.50(0.03)	3.70(0.07)	4.36(0.17)

Table 4: Example 2: Mean of MSE, HITS and FP. The numbers between parentheses are the corresponding standard errors.

Variables	1	2	3	4	5	6	7	8	9	10
AIC	0.46	0.79	0.88	0.44	0.46	0.78	0.67	0.52	0.48	0.39
BIC	0.41	0.71	0.86	0.43	0.33	0.77	0.63	0.42	0.45	0.35
BRIC	0.21	0.60	0.80	0.17	0.13	0.65	0.39	0.18	0.18	0.12
EB-L	0.22	0.59	0.80	0.17	0.14	0.66	0.38	0.19	0.19	0.12
EB-G	0.21	0.59	0.81	0.16	0.13	0.65	0.37	0.19	0.16	0.10
ZS-N	0.22	0.59	0.80	0.17	0.14	0.66	0.38	0.19	0.19	0.12
ZS-F	0.40	0.72	0.84	0.37	0.31	0.79	0.62	0.38	0.41	0.31
OVS	0.23	0.44	0.74	0.17	0.23	0.62	0.36	0.19	0.18	0.09
HG-3	0.21	0.54	0.80	0.16	0.13	0.63	0.35	0.18	0.18	0.10
HG-4	0.18	0.56	0.81	0.15	0.11	0.63	0.35	0.17	0.17	0.08
HG-2	0.22	0.60	0.78	0.16	0.13	0.59	0.42	0.10	0.15	0.11
NIMS	0.19	0.59	0.80	0.16	0.14	0.66	0.37	0.19	0.18	0.10
LASSO	0.47	0.77	0.90	0.53	0.40	0.89	0.79	0.57	0.55	0.43
DZ	0.40	0.65	0.79	0.46	0.37	0.76	0.63	0.32	0.36	0.32
ENET	0.68	0.85	0.97	0.74	0.74	0.96	0.92	0.76	0.75	0.69

Table 5: Example 2: Relative frequencies of the selected variables for methods under comparison.

	MSE_y	HITS	FP
ORACLE	3.31(0.03)	3.00(0.00)	0.00(0.00)
AIC	4.32(0.09)	2.11(0.07)	2.06(0.14)
BIC	4.24(0.08)	1.97(0.07)	1.68(0.14)
BRIC	4.07(0.07)	1.66(0.07)	0.53(0.08)
EB-L	4.06(0.06)	1.84(0.07)	0.79(0.09)
EB-G	4.07(0.07)	1.88(0.07)	0.83(0.09)
ZS-N	4.01(0.06)	1.81(0.07)	0.76(0.09)
ZS-F	4.04(0.07)	2.10(0.07)	1.26(0.11)
OVS	4.27(0.09)	1.78(0.07)	0.64(0.09)
HG-3	4.05(0.06)	1.81(0.07)	0.77(0.09)
HG-4	4.08(0.06)	1.84(0.07)	0.78(0.09)
HG-2	3.98(0.05)	1.80(0.08)	0.73(0.10)
NIMS	3.99(0.06)	1.83(0.07)	0.77(0.09)
LASSO	4.03(0.06)	2.33(0.07)	1.61(0.16)
DZ	4.32(0.10)	2.20(0.11)	2.06(0.16)
ENET	4.13(0.06)	2.38(0.06)	2.04(0.16)

Table 6: Example 3: Mean of MSE, HITS and FP. The numbers between parentheses are the corresponding standard errors.

Variables	1	2	3	4	5	6	7	8
AIC	0.89	0.52	0.45	0.43	0.70	0.36	0.42	0.40
BIC	0.89	0.44	0.39	0.36	0.64	0.30	0.33	0.30
BRIC	0.82	0.35	0.09	0.13	0.49	0.12	0.08	0.11
EB-L	0.87	0.38	0.13	0.19	0.59	0.18	0.14	0.15
EB-G	0.89	0.39	0.15	0.20	0.60	0.18	0.14	0.16
ZS-N	0.87	0.37	0.13	0.19	0.57	0.16	0.13	0.15
ZS-F	0.92	0.51	0.23	0.34	0.67	0.22	0.24	0.23
OVS	0.86	0.37	0.12	0.14	0.55	0.16	0.08	0.14
HG-3	0.87	0.38	0.13	0.19	0.56	0.16	0.14	0.15
HG-4	0.88	0.38	0.13	0.19	0.58	0.17	0.14	0.15
HG-2	0.80	0.46	0.19	0.17	0.60	0.21	0.12	0.15
NIMS	0.87	0.38	0.12	0.19	0.58	0.17	0.14	0.15
LASSO	0.96	0.70	0.32	0.40	0.67	0.29	0.23	0.37
DZ	0.82	0.71	0.42	0.47	0.67	0.47	0.31	0.39
ENET	0.97	0.71	0.49	0.50	0.70	0.40	0.30	0.35

Table 7: Example 3: Relative frequencies of the selected variables for methods under comparison.

	MSE_y	HITS	FP
ORACLE	1.43(0.03)	8.00(0.00)	0.00(0.00)
AIC	1.60(0.03)	6.32(0.11)	0.00(0.00)
BIC	1.64(0.03)	5.99(0.12)	0.00(0.00)
BRIC	1.79(0.04)	4.35(0.11)	0.00(0.00)
EB-L	1.75(0.04)	4.39(0.10)	0.00(0.00)
EB-G	1.76(0.04)	4.34(0.10)	0.00(0.00)
ZS-N	1.74(0.04)	4.38(0.10)	0.00(0.00)
ZS-F	1.62(0.04)	5.37(0.10)	0.00(0.00)
OVS	2.22(0.04)	3.82(0.10)	0.00(0.00)
HG-3	1.76(0.04)	4.32(0.10)	0.00(0.00)
HG-4	1.78(0.03)	4.19(0.09)	0.00(0.00)
HG-2	1.77(0.04)	4.18(0.11)	0.00(0.00)
NIMS	1.75(0.04)	4.39(0.10)	0.00(0.00)
LASSO	1.59(0.04)	7.13(0.12)	0.00(0.00)
DZ	1.56(0.03)	6.82(0.11)	0.00(0.00)
ENET	1.54(0.03)	7.53(0.08)	0.00(0.00)

Table 8: Example 4: Mean of MSE, HITS and FP. The numbers between parentheses are the corresponding standard errors.

Variables	1	2	3	4	5	6	7	8
AIC	0.80	0.81	0.78	0.75	0.76	0.86	0.77	0.79
BIC	0.76	0.76	0.75	0.72	0.68	0.83	0.71	0.78
BRIC	0.45	0.58	0.50	0.65	0.54	0.55	0.48	0.60
EB-L	0.46	0.57	0.52	0.67	0.54	0.54	0.50	0.59
EB-G	0.45	0.57	0.52	0.66	0.54	0.53	0.48	0.59
ZS-N	0.46	0.57	0.52	0.67	0.54	0.54	0.49	0.59
ZS-F	0.62	0.69	0.60	0.78	0.65	0.67	0.62	0.74
OVS	0.38	0.57	0.45	0.64	0.40	0.49	0.44	0.45
HG-3	0.45	0.57	0.51	0.67	0.54	0.53	0.48	0.57
HG-4	0.44	0.57	0.48	0.66	0.51	0.52	0.45	0.56
HG-2	0.53	0.56	0.50	0.50	0.54	0.55	0.53	0.47
NIMS	0.46	0.58	0.51	0.67	0.54	0.54	0.50	0.59
LASSO	0.82	0.90	0.96	0.92	0.85	0.91	0.87	0.90
DZ	0.84	0.85	0.84	0.82	0.83	0.91	0.89	0.84
ENET	0.89	0.93	0.96	0.97	0.96	0.93	0.96	0.93

Table 9: Example 4: Relative frequencies of the selected variables for methods under comparison.

	MSE_y	HITS	FP
ORACLE	1.07(0.09)	2.00(0.00)	0.00(0.00)
AIC	1.48(0.05)	1.93(0.02)	2.88(0.19)
BIC	1.39(0.04)	1.94(0.02)	2.04(0.18)
BRIC	1.24(0.02)	1.93(0.02)	0.50(0.09)
EB-L	1.27(0.02)	1.93(0.02)	0.58(0.10)
EB-G	1.27(0.02)	1.93(0.02)	0.60(0.10)
ZS-N	1.26(0.02)	1.93(0.02)	0.57(0.10)
ZS-F	1.33(0.03)	1.94(0.02)	1.84(0.14)
OVS	1.32(0.04)	1.89(0.03)	0.76(0.08)
HG-3	1.28(0.02)	1.93(0.02)	0.53(0.09)
HG-4	1.30(0.02)	1.93(0.02)	0.54(0.09)
HG-2	1.25(0.02)	1.93(0.02)	0.36(0.09)
NIMS	1.22(0.02)	1.93(0.02)	0.57(0.10)
LASSO	1.39(0.03)	1.99(0.01)	2.93(0.21)
DZ	1.36(0.04)	1.91(0.03)	2.70(0.18)
ENET	1.43(0.03)	1.96(0.02)	3.25(0.20)

Table 10: Example 5: Mean of MSE, HITS and FP. The numbers between parentheses are the corresponding standard errors.

Variables	1	2	3	4	5	6	7	8	9
AIC	0.36	0.94	0.47	0.99	0.35	0.36	0.34	0.53	0.47
BIC	0.30	0.94	0.38	1.00	0.26	0.24	0.22	0.35	0.29
BRIC	0.10	0.94	0.09	1.00	0.10	0.03	0.05	0.08	0.05
EB-L	0.10	0.93	0.14	1.00	0.11	0.04	0.05	0.08	0.06
EB-G	0.11	0.93	0.14	1.00	0.11	0.04	0.05	0.08	0.07
ZS-N	0.10	0.93	0.13	1.00	0.11	0.04	0.05	0.08	0.06
ZS-F	0.29	0.94	0.32	1.00	0.23	0.22	0.19	0.31	0.28
OVS	0.16	0.92	0.10	0.97	0.15	0.07	0.09	0.11	0.08
HG-3	0.10	0.93	0.11	1.00	0.11	0.03	0.04	0.08	0.06
HG-4	0.10	0.93	0.12	1.00	0.11	0.03	0.04	0.08	0.06
HG-2	0.08	0.95	0.07	1.00	0.04	0.03	0.02	0.06	0.06
NIMS	0.06	0.97	0.10	1.00	0.11	0.08	0.05	0.08	0.08
LASSO	0.51	0.99	0.35	1.00	0.47	0.38	0.37	0.41	0.44
DZ	0.50	0.93	0.32	0.98	0.42	0.45	0.26	0.32	0.43
ENET	0.52	0.96	0.37	1.00	0.55	0.44	0.43	0.50	0.44

Table 11: Example 5: Relative frequencies of the selected variables for methods under comparison.

	MSE_y	FP
ORACLE	1.99(0.01)	0.00(0.00)
AIC	2.80(0.07)	3.16(0.21)
BIC	2.62(0.06)	2.24(0.19)
BRIC	2.19(0.02)	0.59(0.11)
EB-L	2.12(0.02)	2.87(0.15)
EB-G	2.11(0.02)	1.54(0.19)
ZS-N	2.26(0.02)	1.02(0.17)
ZS-F	2.31(0.03)	2.51(0.17)
OVS	2.57(0.06)	2.10(0.17)
HG-3	2.13(0.02)	2.18(0.18)
HG-4	2.10(0.01)	2.54(0.17)
HG-2	2.16(0.02)	2.17(0.15)
NIMS	2.24(0.02)	0.99(0.13)
LASSO	2.19(0.04)	1.79(0.22)
DZ	2.57(0.05)	2.49(0.20)
ENET	2.20(0.04)	2.23(0.23)

Table 12: Example 6: Mean of MSE and FP. The numbers between parentheses are the corresponding standard errors.

Variables	1	2	3	4	5	6	7	8
AIC	0.38	0.36	0.31	0.37	0.49	0.42	0.41	0.42
BIC	0.26	0.22	0.23	0.26	0.31	0.36	0.33	0.27
BRIC	0.09	0.04	0.07	0.08	0.08	0.09	0.09	0.05
EB-L	0.37	0.27	0.28	0.30	0.43	0.43	0.38	0.41
EB-G	0.19	0.12	0.16	0.16	0.21	0.27	0.25	0.18
ZS-N	0.14	0.07	0.11	0.10	0.16	0.16	0.18	0.10
ZS-F	0.29	0.27	0.23	0.28	0.41	0.38	0.34	0.31
OVS	0.26	0.26	0.36	0.23	0.28	0.26	0.28	0.17
HG-3	0.27	0.21	0.20	0.26	0.32	0.35	0.30	0.27
HG-4	0.32	0.25	0.23	0.29	0.40	0.38	0.35	0.32
HG-2	0.25	0.19	0.23	0.25	0.31	0.35	0.32	0.27
NIMS	0.12	0.06	0.10	0.11	0.14	0.17	0.18	0.11
LASSO	0.22	0.17	0.23	0.22	0.24	0.25	0.29	0.17
DZ	0.23	0.30	0.17	0.20	0.30	0.27	0.25	0.23
ENET	0.30	0.26	0.27	0.25	0.28	0.28	0.33	0.26

Table 13: Example 6: Relative frequencies of the selected variables for methods under comparison.

	MSE_y	HITS	FP
$\mathbf{y} = \mathbf{y} + 10 \times RSS$	3.41(0.03)	0.15(0.04)	0.00(0.00)
$\mathbf{y} = \mathbf{y} + 10^2 \times RSS$	3.59(0.03)	0.01(0.01)	0.00(0.00)
$\mathbf{y} = \mathbf{y} + 10^3 \times RSS$	3.59(0.02)	0.00(0.00)	0.00(0.00)

Table 14: Example 1: Mean of MSE, HITS and FP after replacing $\mathbf{y}$ by $\mathbf{y} = \mathbf{y}+10^k RSS$ for $k \in \{1, 2, 3\}$. The numbers between parentheses are the corresponding standard errors for the NIMS selection procedure.

Variables	1	2	3	4	5	6	7	8	9	10
$\mathbf{y} = \mathbf{y} + 10 \times RSS$	0.00	0.00	0.09	0.00	0.00	0.05	0.01	0.00	0.00	0.00
$\mathbf{y} = \mathbf{y} + 10^2 \times RSS$	0.00	0.00	0.01	0.00	0.00	0.01	0.00	0.00	0.00	0.00
$\mathbf{y} = \mathbf{y} + 10^3 \times RSS$	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 15: Example 1: Relative frequencies of the selected variables after replacing $\mathbf{y}$ by $\mathbf{y} = \mathbf{y} + 10^k RSS$ for $k \in \{1, 2, 3\}$.

	MSE_y	Mean of selected variables
AIC	4.58(0.05)	5.56(0.20)
BIC	4.60(0.05)	4.20(0.18)
BRIC	4.51(0.05)	2.84(0.15)
EB-L	4.52(0.05)	3.00(0.18)
EB-G	4.52(0.05)	3.28(0.17)
ZS-N	4.52(0.05)	2.96(0.18)
ZS-F	4.49(0.05)	4.28(0.20)
OVS	4.65(0.07)	2.96(0.18)
HG-3	4.54(0.05)	3.00(0.18)
HG-4	4.56(0.05)	3.24(0.17)
HG-2	4.50(0.05)	2.48(0.14)
NIMS	4.50(0.05)	2.44(0.14)
LASSO	4.54(0.05)	8.17(0.52)
DZ	4.51(0.06)	11.03(0.11)
ENET	4.54(0.05)	9.04(0.56)

Table 16: Body fat dataset: Mean of the MSE_y and of the selected variables.

Variables	1	2	3	4	5	6	7	8	9	10	11	12	13
AIC	0.44	0.84	0.16	0.64	0.04	1.00	0.20	0.16	0.08	0.16	0.44	0.80	0.88
BIC	0.08	0.84	0.08	0.32	0.00	1.00	0.12	0.08	0.04	0.00	0.16	0.28	0.40
BRIC	0.08	0.84	0.08	0.32	0.00	1.00	0.12	0.08	0.04	0.00	0.16	0.24	0.40
EB-L	0.08	0.84	0.08	0.32	0.00	1.00	0.12	0.08	0.04	0.00	0.16	0.28	0.40
EB-G	0.08	0.88	0.08	0.36	0.00	1.00	0.08	0.08	0.04	0.00	0.20	0.36	0.40
ZS-N	0.08	0.84	0.08	0.32	0.00	1.00	0.12	0.08	0.04	0.00	0.16	0.24	0.40
ZS-F	0.20	0.84	0.12	0.40	0.00	1.00	0.12	0.12	0.08	0.04	0.24	0.60	0.68
OVS	0.12	0.68	0.08	0.16	0.04	1.00	0.08	0.00	0.00	0.00	0.04	0.24	0.52
HG-3	0.08	0.84	0.08	0.32	0.00	1.00	0.12	0.08	0.04	0.00	0.16	0.28	0.40
HG-4	0.08	0.88	0.08	0.32	0.00	1.00	0.08	0.08	0.04	0.00	0.16	0.36	0.40
HG-2	0.04	0.88	0.00	0.08	0.00	1.00	0.08	0.04	0.00	0.04	0.16	0.28	0.60
NIMS	0.04	0.88	0.04	0.08	0.00	1.00	0.04	0.08	0.04	0.00	0.04	0.04	0.12
LASSO	1.00	0.28	1.00	0.88	0.24	1.00	0.44	0.52	0.28	0.56	0.68	0.84	1.00
DZ	1.00	0.80	1.00	0.88	0.60	1.00	0.80	0.72	0.40	0.88	0.92	0.88	0.96
ENET	1.00	0.40	1.00	0.80	0.28	1.00	0.40	0.64	0.44	0.64	0.68	0.84	1.00

Table 17: Body fat dataset: relative frequencies of selections of the variables over the 25 random splits.

	MSE_y	Mean number of selected variables
AIC	4.79(0.05)	3.52(0.14)
BIC	4.77(0.05)	2.88(0.07)
BRIC	4.78(0.05)	2.88(0.07)
EB-L	4.78(0.05)	2.88(0.07)
EB-G	4.78(0.05)	2.92(0.05)
ZS-N	4.78(0.05)	2.88(0.07)
ZS-F	4.77(0.05)	3.12(0.07)
OVS	4.81(0.05)	2.88(0.10)
HG-3	4.78(0.05)	2.88(0.07)
HG-4	4.78(0.05)	2.92(0.05)
HG-2	4.80(0.05)	2.68(0.10)
NIMS	4.79(0.05)	2.68(0.10)
LASSO	4.78(0.05)	5.24(0.21)
DZ	4.80(0.05)	5.12(0.13)
ENET	4.79(0.05)	5.32(0.16)

Table 18: Ozone dataset: Mean of the MSE_y and of the selected variables.

Variables	1	2	3	4	5	6	7	8
AIC	0.20	0.12	0.96	1.00	0.56	0.08	0.44	0.16
BIC	0.04	0.00	0.96	1.00	0.60	0.00	0.36	0.04
BRIC	0.04	0.00	0.96	1.00	0.60	0.00	0.40	0.04
EB-L	0.04	0.00	0.96	1.00	0.60	0.40	0.36	0.04
EB-G	0.04	0.00	0.96	1.00	0.60	0.00	0.36	0.04
ZS-N	0.04	0.00	0.96	1.00	0.60	0.00	0.36	0.04
ZS-F	0.04	0.08	0.92	1.00	0.60	0.00	0.40	0.08
OVS	0.00	0.00	1.00	0.92	0.00	0.00	0.80	0.08
HG-3	0.04	0.00	0.96	1.00	0.60	0.00	0.36	0.04
HG-4	0.04	0.00	0.96	1.00	0.60	0.00	0.36	0.04
HG-2	0.04	0.00	0.96	1.00	0.60	0.00	0.32	0.04
NIMS	0.04	0.00	0.96	1.00	0.60	0.00	0.32	0.04
LASSO	0.00	0.00	1.00	1.00	1.00	0.00	1.00	1.00
DZ	0.00	0.00	1.00	1.00	1.00	0.00	1.00	1.00
ENET	0.00	0.00	1.00	1.00	1.00	0.00	1.00	1.00

Table 19: Ozone dataset: relative frequencies of selections of the variables over the 25 random splits.