Combining a Relaxed EM Algorithm with Occam's Razor for Bayesian Variable Selection in High-Dimensional Regression

Pierre Latouche, Pierre-Alexandre Mattei, Charles Bouveyron, Julien Chiquet

Combining a Relaxed EM Algorithm with Occam's Razor for Bayesian Variable Selection in High-Dimensional Regression

Pierre Latouche[1], Pierre-Alexandre Mattei*[2], Charles Bouveyron[2], and Julien Chiquet[3]

[1]*Laboratoire SAMM, EA 4543, Université Paris 1 Panthéon-Sorbonne*
[2]*Laboratoire MAP5, UMR CNRS 8145, Université Paris Descartes*
[3]*Laboratoire LaMME, UMR CNRS 8071/UEVE, USC INRA, Évry, France*

Abstract

We address the problem of Bayesian variable selection for high-dimensional linear regression. We consider a generative model that uses a spike-and-slab-like prior distribution obtained by multiplying a deterministic binary vector, which traduces the sparsity of the problem, with a random Gaussian parameter vector. The originality of the work is to consider inference through relaxing the model and using a type-II log-likelihood maximization based on an EM algorithm. Model selection is performed afterwards relying on Occam's razor and on a path of models found by the EM algorithm. Numerical comparisons between our method, called spinyReg, and state-of-the-art high-dimensional variable selection algorithms (such as lasso, adaptive lasso, stability selection or spike-and-slab procedures) are reported. Competitive variable selection results and predictive performances are achieved on both simulated and real benchmark data sets. An original regression data set involving the prediction of the number of visitors of the Orsay museum in Paris using bike-sharing system data is also introduced, illustrating the efficiency of the proposed approach. An R package implementing the spinyReg method is currently under development and is available at `https://r-forge.r-project.org/projects/spinyreg`.

Keywords: EM algorithm, high-dimensional data, linear regression, Occam's razor, spike-and-slab, variable selection.

*Corresponding author, Laboratoire MAP5, Université Paris Descartes, 45 rue des Saints Pères, 75270 Paris Cedex 06, France, tel. +33183945891, `pierrealex.mattei@gmail.com`.

1 Introduction

Over the past decades, parsimony has emerged as a very natural way to deal with high-dimensional data spaces (Candès, 2014). In the context of linear regression, finding a parsimonious parameter vector can both prevent overfitting, make an ill-posed problem (such as a "large p, small n" situation) tractable, and allow to interpret easily the data by finding which predictors are relevant. The problem of finding such predictors is referred to as *sparse regression* or *variable selection* and has mainly been considered either by likelihood penalization of the data, or by using Bayesian models.

1.1 Penalized likelihood

The most natural sparsity-inducing penalty, the ℓ_0-pseudonorm, is linked to the Akaike information criterion (Akaike, 1973) and to optimal subset selection. As proven by Natarajan (1995), it unfortunately leads to an NP-hard optimization problem that is intractable as soon as the number of predictors exceeds a few dozens. To overcome this restriction, convex relaxation of the ℓ_0-pseudonorm, that is, ℓ_1-regularization, have become a basic tool in modern statistics. The most spread formulation of the ℓ_1-penalized linear regression was introduced by Tibshirani (1996) as the "least absolute shrinkage and selection operator" (lasso) and by Chen et al. (1998) as "basis pursuit" in a signal processing framework. Several algorithms allow fast computations of the lasso, even when the number of predictors largely exceeds the number of observations. Among them is the popular least angle angle regression algorithm (LARS) (Efron et al., 2004). The Dantzig selector, introduced by Candès and Tao (2007) as a refined ℓ_1-regularization problem, gives good variable selection performances while simply involving the resolution of a linear program. However, as proved by Zhao and Yu (2006), the crude lasso is not model-consistent unless some cumbersome conditions on the design matrix. Moreover, Zou and Hastie (2005) showed that it can be sensitive to highly correlated predictors and Pötscher and Leeb (2009) warned that its distributional properties can be surprisingly complex. A large number of proposals have been made to enhance the lasso as a selection operator. The adaptive lasso of Zou (2006) is a weighted version enjoying nice oracle properties that works extremely well in practice. "Bolasso", introduced by Bach (2008), achieves model consistency by combining

the lasso with a bootstrap step. In a similar fashion, the stability selection of Meinshausen and Bühlmann (2010) applies many lasso procedures with randomized weights on subsamples of the original data. This technique leads to an effective model selection, even in the presence of correlated predictors.

1.2 Bayesian modelling

Bayesian models have also been widely studied in a variable selection context (see O'Hara and Sillanpää (2009) for a recent review). However, most Bayesian techniques have difficulties in treating the case where the number of observations is smaller than the number of predictors (the so called "large p, small n" situation), mostly because of the exponential growth of the number of possible models (p predictors lead to 2^p models). Another drawback is the fact that the most classical linear regression prior, Zellner's g-prior (for example reviewed and improved by Liang et al. (2008)), involves to invert the Fisher information matrix which is impossible in a "large p, small n" situation. Even though some regularization attempts of the g-prior have been made by Baragatti and Pommeret (2012), the most efficient high-dimensional Bayesian techniques essentially rest on spike-and-slab procedures. Spike-and-slab models, first introduced by Mitchell and Beauchamp (1988), use mixtures of two distributions as priors for the regression coefficients: a thin one, corresponding to irrelevant predictors (the *spike*, typically a Dirac law or a Gaussian distribution with small variance) and a thick one, corresponding to the relevant variables (the *slab*, typically a uniform or Gaussian distribution of large variance). Notably, the refined spike-and-slab model of Ishwaran and Rao (2005a) or the PAC-Bayesian approach of Rigollet and Tsybakov (2011) have been particularly efficient even in very high-dimensional settings. Markov chain Monte Carlo (MCMC) methods have been usually chosen to select models with the highest posterior distributions. MCMC techniques, reviewed for example by Robert and Casella (2004), have an important computational cost and may suffer, as underlined by O'Hara and Sillanpää (2009), from poor mixing properties in the case of spike-and-slab-like priors. A few deterministic methods have also recently been proposed to tackle this issue. The expectation propagation (EP) algorithm was applied to perform approximate inference for group feature selection with a spike-and-slab model by Hernández-Lobato

et al. (2013). The expectation maximization (EM) algorithm was used by Ročková and George (2013) in the case of a hierarchical Bayesian model or by Yengo et al. (2014a) in the case of a multi-slab empirical Bayes framework.

1.3 Our approach

As an alternative, our approach uses spike-and-slab-like priors induced by a binary vector which segregates the relevant from the irrelevant predictors. Such vectors, introduced by George and McCulloch (1993) have been widely used in the Bayesian literature, but have always been considered as random parameters. In most Bayesian contexts like the (hierarchical) ones of George and McCulloch (1993) and Ishwaran and Rao (2005b) or the (empirical Bayes) one of George and Foster (2000), such a binary vector would be classically endowed with a product of Bernoulli prior distributions. In a PAC-Bayesian perspective, more complex prior distributions used for example by Alquier and Lounici (2011) or Rigollet and Tsybakov (2011) led to precise oracle inequalities and competitive predictive performances. In our work, the originality is to consider a deterministic binary vector, and to relax it in order to rely on an EM algorithm. This relaxed procedure allows us to find a family of p models, ordered by sparsity. Model selection is performed afterwards by maximizing the marginal likelihood over this family of models. This way to treat some parameters in a Bayesian way, and others in a frequentist one, is particularly motivated by the unifying multi-level inference approach advocated by Guyon et al. (2010) and by recent advances in Bayesian theory on the merging partly frequentist empirical Bayes methods and classical hierarchical Bayesian approaches (Scott and Berger, 2010; Petrone et al., 2014).

The remainder of this document is organized as follows. In Section 2, a sparse generative model is defined and the general properties of its posterior distribution are exhibited. Section 3 shows how a relaxation of this model is considered in order to perform inference through an EM algorithm. Section 4 explains the model selection procedure of our approach and gives details about Occam's razor automatic selection as well as a link with classical frequentist penalized estimators. In Section 5, a new algorithm, called "spinyReg", for variable selection in high-dimensional regression is introduced. Section 6 presents a benchmark comparison between spinyReg and classical frequentist and Bayesian variable

selection procedures, real and simulated data sets are considered. In Section 7, an original high-dimensional regression database, called "OrsayVelib", is introduced and is used to demonstrate the efficiency of our approach.

1.4 Notation

Vectors and matrices are denoted by bold cases. Given a vector $\mathbf{x} \in \mathbb{R}^p$, we define its Euclidean norm as $||\mathbf{x}||_2 = (\sum_{i=1}^p |x_i|^2)^{1/2}$, his support as $\text{Supp}(\mathbf{x}) = \{i \in \{1, ..., p\}, x_i \neq 0\}$, and its ℓ_0-pseudonorm as $||\mathbf{x}||_0 = \#\text{Supp}(\mathbf{x})$, where x_i denotes the i-th coordinate of $\mathbf{x}$. We write $\mathcal{M}_{n,p}$ the set of real matrices of dimension $n \times p$. Given a vector $\mathbf{x} \in \mathbb{R}^n$, we denote $\text{diag}(\mathbf{x})$ the matrix of $\mathcal{M}_{n,n}$ with diagonal $\mathbf{x}$. For two matrices A and B of $\mathcal{M}_{n,p}$, we define their Hadamard product as $A \odot B = (a_{ij}b_{ij})_{i \leq n, j \leq p}$ where a_{ij} and b_{ij} respectively denote the (i, j)-th coordinate of A and B. The identity matrix of dimension n is denoted by I_n. Given a binary vector $\mathbf{z} \in \{0, 1\}^p$, we denote $\bar{\mathbf{z}}$ the binary vector of $\{0, 1\}^p$ whose support is exactly the complement of $\text{Supp}(\mathbf{z})$. Given a binary vector $\mathbf{z} \in \{0, 1\}^p$ and a matrix $\mathbf{A} \in \mathcal{M}_{n,p}$, we denote $\mathbf{A}_\mathbf{z}$ the extracted matrix of $\mathbf{A}$ where only the columns corresponding to the nonzero indexes of $\mathbf{z}$ have been kept. Given a mean vector $\boldsymbol{\mu} \in \mathbb{R}^n$ and a positive definite covariance matrix $\mathbf{S} \in \mathcal{M}_n$, the density of the normal distribution is denoted $\mathcal{N}(\,\cdot\,; \boldsymbol{\mu}, \mathbf{S})$. Given a real number y, δ_y denotes the Dirac function with mass at y.

2 A sparse generative model

This section introduces a sparse generative model based on a spike-and-slab-like prior, and describes the general properties of its posterior distribution. Links with related models are also discussed.

2.1 The model

Let us consider the following regression model

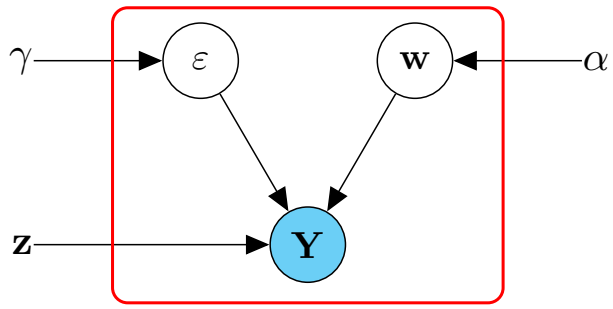


Figure 1: Graphical representation of the sparse generative model.

$$\begin{cases} \mathbf{Y} &= \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \\ \boldsymbol{\beta} &= \mathbf{z} \odot \mathbf{w}, \end{cases} \tag{1}$$

where $\mathbf{Y} \in \mathbb{R}^n$ is the set of n observed responses, $\mathbf{X} \in \mathcal{M}_{n,p}(\mathbb{R})$ is the design matrix with p input variables. The vector $\boldsymbol{\varepsilon}$ is a noise term with $p(\boldsymbol{\varepsilon}|\gamma) = \mathcal{N}(\boldsymbol{\varepsilon}; 0, I_n/\gamma)$. A prior distribution $p(\mathbf{w}|\alpha) = \mathcal{N}(\mathbf{w}; 0, I_p/\alpha)$ with an isotropic covariance matrix is further assumed. Moreover, we denote by $\mathbf{z} \in \{0,1\}^p$ a binary deterministic parameter vector, whose nonzero entries correspond to the active variables of the regression model. It is worth noticing that such modeling induces a spike-and-slab-like prior distribution for $\boldsymbol{\beta}$:

$$\begin{aligned} p(\boldsymbol{\beta}|\mathbf{z}, \alpha) &= \prod_{j=1}^{p} p(\beta_j|z_j, \alpha) \\ &= \prod_{j=1}^{p} \delta_0(\beta_j)^{1-z_j} \mathcal{N}(\beta_j; 0, 1/\alpha)^{z_j}. \end{aligned} \tag{2}$$

However, we emphasize that, contrary to standard spike-and-slab models (Mitchell and Beauchamp, 1988) which assume a Bernoulli prior distribution over $\mathbf{z}$, we see $\mathbf{z}$ here as a deterministic parameter to be inferred from the data. As we shall see in Section 3, this allows us to work with a marginal log-likelihood which involves an Occam's razor term, allowing model selection afterwards. In the same spirit, we do not put any prior distribution on γ nor α. Finally, the graphical model is presented in Figure 1 and we denote by $q = \sum_{j=1}^{p} z_j$ the number of relevant variables and $\mathbf{Z} = \mathrm{diag}(\mathbf{z})$.

2.2 Posterior distribution

From now on, to simplify notations, the dependency on $\mathbf{X}$ in conditional distributions will be omitted.

Proposition 1. *The posterior distribution of* $\mathbf{w}$ *given the data is given by*

$$p(\mathbf{w}|\mathbf{Y}, \mathbf{Z}, \alpha, \gamma) = \mathcal{N}(\mathbf{w}; \mathbf{m}, \mathbf{S}), \tag{3}$$

where $\mathbf{S} = (\gamma \mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z} + \alpha I_p)^{-1}$ *and* $\mathbf{m} = \gamma \mathbf{S}\mathbf{Z}\mathbf{X}^T\mathbf{Y}$.

Proof. Using Bayes' rule, we have

$$\begin{aligned}
\log p(\mathbf{w}|\mathbf{Y}, \mathbf{Z}, \alpha, \gamma) &= \log p(\mathbf{Y}|\mathbf{w}, \mathbf{Z}, \gamma) + \log p(\mathbf{w}|\alpha) + K_1 \\
&= -\frac{\gamma}{2} \|\mathbf{Y} - \mathbf{X}\mathbf{Z}\mathbf{w}\|_2^2 - \frac{\alpha}{2} \|\mathbf{w}\|_2^2 + K_2 \\
&= -\frac{\gamma}{2} \mathbf{w}^T \mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z}\mathbf{w} + \gamma \mathbf{w}^T \mathbf{Z}\mathbf{X}^T\mathbf{Y} - \frac{\alpha}{2} \|\mathbf{w}\|_2^2 + K_3 \\
&= -\frac{1}{2} \mathbf{w}^T \mathbf{S}^{-1} \mathbf{w} + \mathbf{w}^T \mathbf{S}^{-1} \mathbf{m} + K_3.
\end{aligned}$$

where K_1, K_2 and K_3 are quantities that do not depend on $\mathbf{w}$. Therefore $p(\mathbf{w}|\mathbf{Y}, \mathbf{Z}, \alpha, \gamma) = \mathcal{N}(\mathbf{w}, \mathbf{m}, \mathbf{S})$. □

The vector $\mathbf{m}$ is the maximum a posteriori (MAP) estimate of $\boldsymbol{\beta}$. Next proposition assures that it recovers the support of the parameter vector. Moreover, its nonzero coefficients correspond to ridge estimates with regularization parameter α/γ of the model where only the q predictors corresponding to the support of $\mathbf{z}$ have been kept.

Proposition 2. *We have* $\mathrm{Supp}(\mathbf{m}) = \mathrm{Supp}(\mathbf{z})$ *almost surely and*

$$\mathbf{m}_\mathbf{z} = (\mathbf{X}_\mathbf{z}^T \mathbf{X}_\mathbf{z} + \frac{\alpha}{\gamma} I_p)^{-1} \mathbf{X}_\mathbf{z}^T \mathbf{Y}. \tag{4}$$

Proof. Using (3), one can write

$$\mathbf{S}^{-1}\mathbf{m} = \gamma \mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z}\mathbf{m} + \alpha \mathbf{m} = \gamma \mathbf{Z}\mathbf{X}^T\mathbf{Y},$$

which leads, by separating the lines corresponding to zero and nonzero coefficients of $\mathbf{z}$, to $\mathbf{m}_{\bar{\mathbf{z}}} = 0$ and to (4). Notice that $\mathbf{m}_{\bar{\mathbf{z}}} = 0$ implies $\mathrm{Supp}(\mathbf{m}) \subset \mathrm{Supp}(\mathbf{z})$.

The vector $\mathbf{m_z}$ therefore corresponds to the ridge estimator of the model where only the q predictors corresponding to the support of $\mathbf{z}$ have been kept. As a particular instance of a strictly convex bridge estimator, the coefficients of $\mathbf{m_z}$ are almost surely nonzero (Fu, 1998, Theorem 1), therefore $\text{Supp}(\mathbf{m}) \subset \text{Supp}(\mathbf{z})$ implies that $\mathbf{m}$ and $\mathbf{z}$ have almost surely same support. □

2.3 Links with spike-and-slab models

Let us briefly link the proposed model to typical spike-and-slab models. The corresponding frameworks (Mitchell and Beauchamp, 1988; Hernández-Lobato et al., 2013) would add a hierarchical layer above the model of Figure 1 by using a multivariate Bernoulli prior of the form

$$p(\mathbf{z}) = \prod_{j=1}^{p} \tau_j^{z_j} (1-\tau_j)^{1-z_j},$$

where $\boldsymbol{\tau} = (\tau_1, ..., \tau_p) \in [0,1]^p$. However, as emphasized by Scott and Berger (2010), the estimation of $\boldsymbol{\tau}$ using empirical Bayes techniques can be extremely delicate and is likely to lead to poor variable selection performances. For instance, Hernández-Lobato et al. (2013) underline the fact that, in the case of their spike-and-slab model, the maximization of the evidence led to a sub-optimal choice of the hyper-parameter $\boldsymbol{\tau}$, and therefore to poor variable selection. To avoid such drawbacks, the use of Bernoulli priors are not considered in this paper.

3 Inference

This section now focuses on inferring the model proposed above. To this end, $\mathbf{w}$ is seen as a latent variable while $\mathbf{Z} = \text{diag}(\mathbf{z})$, α, γ are parameters to be estimated from the data $(\mathbf{X}, \mathbf{Y})$ using an empirical Bayes framework.

3.1 Inference strategy and relaxation

The estimators of $\mathbf{z}$, α and γ will be the ones that maximize the *evidence* (or *type-II likelihood*) of the data:

$$p(\mathbf{Y}|\mathbf{X}, \mathbf{z}, \alpha, \gamma) = \int_{\mathbb{R}^p} p(\mathbf{Y}|\mathbf{X}, \mathbf{w}, \mathbf{z}, \alpha, \gamma) p(\mathbf{w}|\alpha) d\mathbf{w}. \tag{5}$$

Seing $\mathbf{w}$ as a latent variable, a natural optimization procedure is the expectation-maximization (EM) algorithm introduced by Dempster et al. (1977). However, the maximization of (5) would be problematic for two reasons – both linked to the discreteness of the model parameter. First, because the optimization problem in $\mathbf{z}$ is combinatorial and 2^p values of $\mathbf{z}$ are possible. Then, because in this case, the parameter space is partly discrete and all theoretical convergence properties of the EM algorithm require a continuous parameter space (Wu, 1983; McLachlan and Krishnan, 2008).

To overcome these issues, we propose to use a simple relaxation by replacing the model parameter by a vector $\mathbf{z}^{\text{relaxed}}$ in $[0, 1]^p$. This relaxation allows us to efficiently maximize the new, relaxed version of (5) using an EM approach.

From now on, and until the end of this section, we will only consider the relaxed model with $\mathbf{z}^{\text{relaxed}} \in [0, 1]^p$. In order to simplify notations, we denote $\mathbf{Z} = \text{diag}(\mathbf{z}^{\text{relaxed}})$.

3.2 E-step

At the E-step of the relaxed EM algorithm, one has to compute the expectation of the complete data log-likelihood $\mathbb{E}_{\mathbf{w}}(\log p(\mathbf{Y}, \mathbf{w}, |\mathbf{Z}, \alpha, \gamma))$ with respect to the posterior distribution $p(\mathbf{w}|\mathbf{Y}, \mathbf{Z}, \alpha, \gamma)$. Consequently, the parameters $\mathbf{S}$ and $\mathbf{m}$ of the Gaussian posterior (3) have to be computed at each step. Notice that these two parameters also allow us to compute a convenient expression of the evidence.

Proposition 3. *The type-II log-likelihood is given by*

$$\log p(\mathbf{Y}|\mathbf{Z}, \alpha, \gamma) = -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\gamma) + \frac{p}{2}\log(\alpha) - \frac{\gamma}{2}\|\mathbf{Y}\|_2^2 + \frac{1}{2}\log\det\mathbf{S} + \frac{1}{2}\mathbf{m}^T\mathbf{S}^{-1}\mathbf{m}. \tag{6}$$

Proof. By directly computing the integrand of (5), we find

$$\log p(\mathbf{Y}|\mathbf{Z},\alpha,\gamma) = -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\gamma) + \frac{p}{2}\log(\alpha) \\ + \log \int_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{\sqrt{(2\pi)^p}} \exp\left(-\frac{\gamma}{2}\mathbf{Y}^T\mathbf{Y} + \gamma\mathbf{Y}^T\mathbf{X}\mathbf{Z}\mathbf{w} - \frac{\gamma}{2}\mathbf{w}^T\mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z}\mathbf{w} - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}\right) d\mathbf{w},$$

which leads to

$$\log p(\mathbf{Y}|\mathbf{Z},\alpha,\gamma) = -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\gamma) + \frac{p}{2}\log(\alpha) - \frac{\gamma}{2}\|\mathbf{Y}\|_2^2 \\ + \log \int_{\mathbf{w}\in\mathbb{R}^p} \frac{1}{\sqrt{(2\pi)^p}} \exp\left(-\frac{1}{2}\mathbf{w}^T\mathbf{S}^{-1}\mathbf{w} + \mathbf{w}^T\mathbf{S}^{-1}\mathbf{m}\right) d\mathbf{w},$$

which allows us to conclude. □

Notice that, by replacing $\mathbf{Z}$ by $\mathrm{diag}(\mathbf{z})$, the expression (6) remains valid in the non-relaxed binary case.

3.3 M-step

At the M-step, the expectation of the complete data log-likelihood $\mathbb{E}_{\mathbf{w}}(\log p(\mathbf{Y},\mathbf{w}|\mathbf{Z},\alpha,\gamma))$ with respect to $p(\mathbf{w}|\mathbf{Y},\mathbf{Z},\alpha,\gamma)$, is maximized over $\mathbf{Z},\alpha,\gamma$.

Proposition 4. *Denoting $\mathbf{\Sigma} = \mathbf{S} + \mathbf{m}\mathbf{m}^T$, the expected complete data log-likelihood is given by*

$$\mathbb{E}_{\mathbf{w}}(\log p(\mathbf{Y},\mathbf{w}|\mathbf{Z},\alpha,\gamma)) = \frac{n}{2}\log(\gamma) - \frac{\gamma}{2}\mathbf{Y}^T\mathbf{Y} - \frac{\alpha}{2}\mathrm{Tr}(\mathbf{\Sigma}) + \frac{p}{2}\log(\alpha) - \frac{p+n}{2}\log(2\pi) \\ + \gamma\mathbf{z}^{\mathrm{relaxed}\,T}(\mathbf{m}\odot(\mathbf{X}^T\mathbf{Y})) - \frac{\gamma}{2}\mathbf{z}^{\mathrm{relaxed}\,T}(\mathbf{X}^T\mathbf{X}\odot\mathbf{\Sigma})\mathbf{z}^{\mathrm{relaxed}}. \quad (7)$$

Proof. We have $\log p(\mathbf{Y},\mathbf{w}|\mathbf{Z},\alpha,\gamma) = \log p(\mathbf{Y}|\mathbf{w},\mathbf{Z},\alpha,\gamma) + \log p(\mathbf{w}|\alpha)$. Thus, since both the prior on $\mathbf{w}$ and the noise are Gaussian, we can write

$$\log p(\mathbf{Y},\mathbf{w}|\mathbf{Z},\alpha,\gamma) = \frac{n}{2}\log\gamma + \frac{p}{2}\log(\alpha) - \frac{p+n}{2}\log(2\pi) - \frac{\gamma}{2}(\mathbf{Y}-\mathbf{X}\mathbf{Z}\mathbf{w})^T(\mathbf{Y}-\mathbf{X}\mathbf{Z}\mathbf{w}) - \frac{\alpha}{2}\mathbf{w}^T\mathbf{w}.$$

Therefore, by expanding and computing the expectation of the expression, we find :

$$\mathbb{E}_{\mathbf{w}}(\log p(\mathbf{Y},\mathbf{w}|\mathbf{Z},\alpha,\gamma)) = \frac{n}{2}\log(\gamma) + \frac{p}{2}\log(\alpha) - \frac{p+n}{2}\log(2\pi) - \frac{\gamma}{2}\mathbf{Y}^T\mathbf{Y} \\ - \frac{\gamma}{2}\mathbb{E}_{\mathbf{w}}(\mathbf{w}^T\mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z}\mathbf{w}) + \gamma\mathbf{Y}^T\mathbf{X}\mathbf{Z}\mathbb{E}_{\mathbf{w}}(\mathbf{w}) - \frac{\alpha}{2}\mathbb{E}_{\mathbf{w}}(\mathbf{w}^T\mathbf{w}).$$

From (4), we have $\mathbb{E}_{\mathbf{w}}(\mathbf{w}) = \mathbf{m}$ and, by using the properties of the trace operator,

$$\mathbb{E}_{\mathbf{w}}(\mathbf{w}^T\mathbf{w}) = \mathbb{E}_{\mathbf{w}}(\mathrm{Tr}(\mathbf{w}\mathbf{w}^T)) = \mathrm{Tr}(\mathbb{E}_{\mathbf{w}}(\mathbf{w}\mathbf{w}^T)) = \mathrm{Tr}(\mathbf{S} + \mathbf{m}\mathbf{m}^T) = \mathrm{Tr}(\boldsymbol{\Sigma}).$$

Thus, we will also have

$$\mathbb{E}_{\mathbf{w}}(\mathbf{w}^T\mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z}\mathbf{w}) = \mathbb{E}_{\mathbf{w}}(\mathrm{Tr}(\mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z}\mathbf{w}\mathbf{w}^T)) = \mathrm{Tr}(\mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z}\boldsymbol{\Sigma}).$$

Moreover, since $\mathbf{Z} = \mathrm{diag}(\mathbf{z}^{\mathrm{relaxed}})$, we can compute

$$\mathbf{Y}^T\mathbf{X}\mathbf{Z}\mathbf{m} = \mathbf{z}^{\mathrm{relaxed}^T}(\mathbf{m} \odot (\mathbf{X}^T\mathbf{Y}))$$

and

$$\mathrm{Tr}(\mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z}\boldsymbol{\Sigma}) = \mathbf{z}^{\mathrm{relaxed}^T}(\mathbf{X}^T\mathbf{X} \odot \boldsymbol{\Sigma})\mathbf{z}^{\mathrm{relaxed}}.$$

By replacing the values of the terms we have just computed, we eventually find the appropriate value of the evidence. □

Maximizing the expectation of the complete data log-likelihood (7) with respect to the parameter γ, α, $\mathbf{z}^{\mathrm{relaxed}}$ leads to the following M-step updates.

Proposition 5. *The values of* γ, α, $\mathbf{z}^{\mathrm{relaxed}}$ *maximizing* (7) *are*

$$\hat{\gamma}^{-1} = \frac{1}{n}\left\{\mathbf{Y}^T\mathbf{Y} + \mathbf{z}^T(\mathbf{X}^T\mathbf{X} \odot \boldsymbol{\Sigma})\mathbf{z} - 2\mathbf{z}^T(\mathbf{m} \odot (\mathbf{X}^T\mathbf{Y}))\right\} \tag{8}$$

$$\hat{\alpha} = \frac{p}{\mathrm{Tr}(\boldsymbol{\Sigma})} \tag{9}$$

$$\hat{\mathbf{z}}^{\mathrm{relaxed}} = \underset{\mathbf{u}\in[0,1]^p}{\mathrm{argmax}} \left\{-\frac{1}{2}\mathbf{u}^T(\mathbf{X}^T\mathbf{X} \odot \boldsymbol{\Sigma})\mathbf{u} + \mathbf{u}^T(\mathbf{m} \odot (\mathbf{X}^T\mathbf{Y}))\right\} \tag{10}$$

Notice that the $\mathbf{z}^{\mathrm{relaxed}}$ update (10) is a quadratic program (QP) which is strictly convex if, and only if $\boldsymbol{\Sigma} \odot \mathbf{X}^T\mathbf{X}$ is positive definite. In fact, the next proposition assures that it is the case if and only if $\mathbf{X}$ has no null column. Therefore, in all practical cases, the objective function of this program is strictly convex and fast convex optimization procedures such as the L-BFGS-B method of Byrd et al. (1995) can be used.

Proposition 6. *The matrix* $\mathbf{X}^T\mathbf{X} \odot \boldsymbol{\Sigma}$ *is positive definite if and only if* $\mathbf{X}$ *has no null column.*

Proof. According to the Schur product theorem (Bapat and Raghavan, 1997, chap. 3), since $\mathbf{X}^T\mathbf{X}$ and $\mathbf{\Sigma}$ are positive semidefinite, $\mathbf{X}^T\mathbf{X} \odot \mathbf{\Sigma}$ is also positive semidefinite. Therefore, $\mathbf{X}^T\mathbf{X} \odot \mathbf{\Sigma}$ is positive definite if and only if its determinant is different from zero.

If one of the columns of $\mathbf{X}$ is null, then the same column of $\mathbf{\Sigma} \odot \mathbf{X}^T\mathbf{X}$ is also null and $\det(\mathbf{\Sigma} \odot \mathbf{X}^T\mathbf{X}) = 0$. The proposed condition is therefore necessary.

If none of the columns $\mathbf{x_1}, ...\mathbf{x_p}$ of $\mathbf{X}$ are null, then Oppenheim's inequality (Oppenheim, 1930; Markham, 1986) leads to

$$\det(\mathbf{\Sigma} \odot \mathbf{X}^T\mathbf{X}) \geq ||\mathbf{x_1}||_2^2 ... ||\mathbf{x_p}||_2^2 \det(\mathbf{\Sigma}). \tag{11}$$

Since $\mathbf{\Sigma} = \mathbf{S} + \mathbf{m}^T\mathbf{m}$, the determinant matrix lemma assures that

$$\det(\mathbf{\Sigma}) = (1 + \mathbf{m}^T\mathbf{S}^{-1}\mathbf{m}) \det(\mathbf{S}),$$

and, since $\mathbf{S}$ and $\mathbf{S}^{-1}$ are positive definite, $\det(\mathbf{S}) > 0$ and $\mathbf{m}^T\mathbf{S}^{-1}\mathbf{m} \geq 0$. Therefore, we find

$$\det(\mathbf{\Sigma}) = (1 + \mathbf{m}^T\mathbf{S}^{-1}\mathbf{m}) \det(\mathbf{S}) \geq \det(\mathbf{S}) > 0,$$

which, combined to (11), leads to $\det(\mathbf{\Sigma} \odot \mathbf{X}^T\mathbf{X}) > 0$. The condition is therefore also sufficient. □

3.4 Pseudo-code

Algorithm 1 presents a pseudo-code for the EM algorithm of the relaxed model.

3.5 Links with automatic relevance determination

Interestingly, this relaxed model is somehow related to the automatic relevance determination (ARD) which uses a prior of the form $p(\boldsymbol{\beta}|\mathbf{a}) = \mathcal{N}(0; \text{diag}(\mathbf{a}))$ and for which the most classical way of inference is also an EM algorithm (MacKay, 1999; Tipping, 2001).

However, our method avoids several drawbacks of this technique. First, we do not assume any hyperprior on $\mathbf{z}^{\text{relaxed}}$ while Tipping (2001) uses a product of flat Gamma priors. More importantly, as pointed out by Wipf and Nagarajan (2008), the convergence of the EM algorithm is extremely slow and not theoretically guaranteed in the case of the ARD model. However, with our approach, since we only need the *order* of the coefficients

Algorithm 1: EM algorithm for the relaxed model

Input: $\mathbf{X}$,$\mathbf{Y}$

Output: $\mathbf{z}^{\text{relaxed}}$

Initialize $\gamma = 1$, $\alpha = 1$, $\mathbf{z}^{\text{relaxed}} = (1, ..., 1)$;

repeat

// E-step

$\mathbf{S} = \gamma(\mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z} + \alpha I_p)^{-1}$;

$\mathbf{m} = \gamma\mathbf{S}\mathbf{Z}\mathbf{X}^T\mathbf{Y}$; $\mathbf{\Sigma} = \mathbf{S} + \mathbf{m}\mathbf{m}^T$;

// M-step

Compute $\hat{\alpha}$ and $\hat{\gamma}$ using (8) and (9);

Compute $\hat{\mathbf{z}}^{\text{relaxed}}$ using (10) and the L-BFGS-B method;

until *convergence of the evidence*;

of $\mathbf{z}^{\text{relaxed}}$ (see Section 4), we do not have to wait for the full convergence of this parameter. In practice, in all the experiments that we carried out, we only had to perform less than a few hundreds of iterations of the algorithm to obtain convergence of the evidence in order to perform variable selection. Notice that the fact that the evidence converges faster than the parameters of the model is a quite general property of EM algorithms (Xu and Jordan, 1996). Moreover, conversely to ARD-like models, our model additionally includes a "ridge parameter" α which, according to Occam's razor (see Section 4), also controls the sparsity. This also leads to an objective function different from the classical ARD one.

4 Model selection

In practice, the vector $\mathbf{z}^{\text{relaxed}}$ has to be binarized in order to select the relevant input variables. A common choice would consist in relying on a threshold τ such that z_j is set to 1 if $z_j \geq \tau$, and to 0 otherwise. However, numerical experiments showed that such a procedure would lead to poor estimates of $\mathbf{z}$. In order to perform an efficient variable selection, we will use the outputs of the relaxed EM algorithm to create a path of models and, relying on Occam's razor, we will afterward maximize the type-II likelihood over this

path to finally select the relevant variables.

4.1 Occam's Razor

One of the key advantages of the approach proposed is that it maximizes a marginal log-likelihood, which automatically penalizes the model complexity by adding a term to the sum of squared errors.

Proposition 7. *Up to unnecessary additive constants, the negative type-II log-likelihood can be written as*

$$\begin{aligned} -\log p(\mathbf{Y}|\mathbf{z},\alpha,\gamma) &= -\log p(\mathbf{Y}|\mathbf{m},\mathbf{z},\gamma) + \text{pen}(\mathbf{z},\alpha,\gamma) \\ &= \frac{\gamma}{2}||\mathbf{Y}-\mathbf{X}_\mathbf{z}\mathbf{m}_\mathbf{z}||_2^2 + \text{pen}(\mathbf{z},\alpha,\gamma) \end{aligned} \tag{12}$$

where

$$\text{pen}(\mathbf{z},\alpha,\gamma) = -\log p(\mathbf{m}|\alpha) - \frac{1}{2}\log\det\mathbf{S} \tag{13}$$

$$= \frac{\alpha}{2}\left\|\mathbf{m}\right\|_2^2 - \frac{\log\alpha}{2}\left\|\mathbf{m}\right\|_0 - \frac{1}{2}\log\det(\gamma\mathbf{X}_\mathbf{z}^T\mathbf{X}_\mathbf{z} + \alpha I_q) \text{ a.s} \tag{14}$$

is the Occam factor.

Proof. First, replacing $\mathbf{w}$ by $\mathbf{m}$ in the log-likelihood leads to

$$\log p(\mathbf{Y}|\mathbf{m},\mathbf{Z},\alpha,\gamma) = -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\gamma) - \frac{\gamma}{2}\left\|\mathbf{Y}\right\|_2^2 - \frac{\gamma}{2}\mathbf{m}^T\mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z}\mathbf{m} + \gamma\mathbf{Y}^T\mathbf{X}\mathbf{Z}\mathbf{m}$$

therefore, since $\mathbf{m}^T\mathbf{S}^{-1}\mathbf{m} = \gamma\mathbf{m}^T\mathbf{Z}\mathbf{X}^T\mathbf{Y} = \gamma\mathbf{Y}^T\mathbf{X}\mathbf{Z}\mathbf{m}$, we have

$$\log p(\mathbf{Y}|\mathbf{m},\mathbf{Z},\alpha,\gamma) = -\frac{n}{2}\log(2\pi) + \frac{n}{2}\log(\gamma) - \frac{\gamma}{2}\left\|\mathbf{Y}\right\|_2^2 - \frac{\gamma}{2}\mathbf{m}^T\mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z}\mathbf{m} + \mathbf{m}^T\mathbf{S}^{-1}\mathbf{m}.$$

Furthermore, $\log p(\mathbf{m}|\alpha) = -\frac{p}{2}\log(2\pi) + \frac{p}{2}\log(\alpha) - \frac{\alpha}{2}\mathbf{m}^T\mathbf{m}$. By summing the terms of the right-hand side of (12), we find the same expression of the type-II log-likelihood as in (6), which proves (12). To prove (14), let us note that

$$-\frac{1}{2}\log\det\mathbf{S} = \frac{1}{2}\log\det(\gamma\mathbf{Z}\mathbf{X}^T\mathbf{X}\mathbf{Z} + \alpha I_p) = \frac{\log\alpha}{2}(p - \left\|\mathbf{z}\right\|_0) - \frac{1}{2}\log\det(\gamma\mathbf{X}_\mathbf{z}^T\mathbf{X}_\mathbf{z} + \alpha I_q).$$

Then, since $\left\|\mathbf{z}\right\|_0 = \left\|\mathbf{m}\right\|_0$ almost surely (see Proposition 2), we find

$$-\frac{1}{2}\log\det\mathbf{S} = \frac{\log\alpha}{2}(p - \left\|\mathbf{m}\right\|_0) - \frac{1}{2}\log\det(\gamma\mathbf{X}_\mathbf{z}^T\mathbf{X}_\mathbf{z} + \alpha I_q) \text{ a.s.}$$

which leads to (14). □

The sparse generative model therefore automatically adds a ℓ_0-ℓ_2 penalty to the likelihood of the model at the MAP value of $\mathbf{w}$. This is somehow similar to the "elastic net" penalty of Zou and Hastie (2005), combined with a penalty linked to the volume of the gaussian posterior $\mathcal{N}(\mathbf{w}; \mathbf{m}, \mathbf{S})$. Notice that, when α is small, the Occam factor will be extremely sparsity-inducing but the coefficients will have a large variance. When α is close to 1, this penalty will lead to moderately sparse but notably shrinked solution. Moreover, if we write $\lambda = (\alpha - \log \alpha)/2$ and $\kappa = \alpha/(\alpha - \log \alpha)$, we obtain almost surely the expression

$$\text{pen}(\mathbf{z}, \alpha, \gamma) = \lambda \left((1-\kappa) \left\| \mathbf{m} \right\|_0 + \kappa \left\| \mathbf{m} \right\|_2^2 \right) - \frac{1}{2} \log \det(\gamma \mathbf{X}_\mathbf{z}^T \mathbf{X}_\mathbf{z} + \alpha I_q),$$

involving a convex combination of the ℓ_0 and ℓ_2 penalties in an elastic net fashion. The elastic net can therefore be seen as some kind of strictly convex approximation of Occam's automatic penalty.

Interestingly, the term $\text{pen}(\mathbf{z}, \alpha, \gamma)$ exactly corresponds to Occam's razor described by MacKay (1992) and detailed by Bishop (2006, chap. 4). Such a term has been widely used for model selection purposes. Let us emphasize that $\text{pen}(\mathbf{z}, \alpha, \gamma)$ is related to the penalization term of the Bayesian information criterion (BIC). Indeed, if a broad Gaussian prior distribution for the vector $\mathbf{w}$ is considered and if the corresponding matrix $\mathbf{S}$ is assumed to have full rank, then Occam's razor is approximately $(-1/2)q \log n$. Contrary to BIC which relies on an asymptotic Laplace approximation, we obtained here an analytical expression of the evidence.

In our model, the minimization of Equation (12) assures that the selected model realizes a tradeoff between the log-likelihood and an automatic penalty term. Note that a PAC-Bayesian study of the performance of Occam's penalty – similarly to the study of BIC-like penalties by Bunea et al. (2007) for instance – would be particularly interesting. However, to the best of our knowledge, such a work has yet to be done.

5 SpinyReg: an algorithm for sparse regression

We called our algorithm, which successively runs the Algorithm 1 and performs model selection over the path of models using Algorithm 2, spinyReg.

5.1 Prediction

The spinyReg algorithm is essentially a variable selection algorithm. In order to perform prediction, the natural estimator of the model is $\hat{\mathbf{z}}$ where

$$\hat{\mathbf{m}} = \gamma(\gamma \text{diag}(\hat{\mathbf{z}})\mathbf{X}^T\mathbf{X}\text{diag}(\hat{\mathbf{z}}) + \alpha I_p)^{-1}\text{diag}(\hat{\mathbf{z}})\mathbf{X}^T\mathbf{Y}.$$

However, as it was stated at the end of Subsection 3.1, this estimator is exactly the ridge estimator performed on a small model where only the predictors corresponding to nonzero coefficients of $\hat{\mathbf{z}}$ are kept. Since we do not wait for the full convergence of the parameters in the EM algorithm, we would rather recommend to perform an ordinary least squares (OLS) estimation or a ridge regression with only a small amount of regularization on the same small model. This is the choice we made in the numerical simulations hereafter.

5.2 Initialization

The choice of initialization $\mathbf{z}^{\text{relaxed}} = (1, ..., 1)$ appears particularly natural because it helps to avoid the unwanted apparition of true zero coefficients in $\mathbf{z}^{\text{relaxed}}$. Indeed, if a coefficient of $\mathbf{z}^{\text{relaxed}}$ by the M-step update (10), then it can not go back to a positive value. This behavior is typical of ARD-like iterative procedures (MacKay, 1999; Tipping, 2001).

Contrary to ARD models, we do not need true zeros in the vector $\mathbf{z}^{\text{relaxed}}$. Therefore, another solution to avoid their apparition would be to perform the quadratic program (10) over $[\eta_n, 1 - \eta_n]$ were $(\eta_n)_{n\leq 1}$ is a vanishing real sequence. The resulting algorithm would be a generalized EM (GEM) algorithm satisfying Wu's convergence conditions (Wu, 1983), contrary to the classical EM algorithm for ARD (Tipping, 2001; Wipf and Nagarajan, 2008). However, because we do not wait for the convergence of $\mathbf{z}^{\text{relaxed}}$, setting the initial coefficients at 1 is sufficient in practice to avoid true zeros. Regarding the parameter α, the form of the Occam factor suggests that using a small value such as $\alpha = 10^{-3}$ will lead to sparse solutions. This is the choice we made in the numerical simulations hereafter.

5.3 Computational cost

At each iteration, the most expensive step is the inversion of the $p \times p$ matrix $\mathbf{S}$ during the E-step. It would imply a $O(p^3)$ complexity, not allowing us to deal with high-dimentional

data. However, using the Woodbury identity, one can write when $p > n$,

$$\mathbf{S} = \frac{1}{\alpha} I_p + \frac{1}{\alpha^2} \left(\mathbf{Z}\mathbf{X}^T\right) \left(\frac{1}{\gamma} I_n + \frac{1}{\alpha} \mathbf{X}\mathbf{Z}^2\mathbf{X}^T\right)^{-1} (\mathbf{X}\mathbf{Z}) .$$

Thus, the final computational cost has therefore a $O(p^2 \min(n, p))$ complexity, which is more suitable for high-dimensional problems.

Overall, MCMC-based Bayesian variable selection methods for regression have a very large computational cost. To the best of our knowledge, the fastest efficient spike-and-slab algorithm for linear regression is the EP procedure of Hernández-Lobato et al. (2013). Each iteration of the EP algorithm costs $O(n^2 p)$ operations, and in practice it needs more iterations than our relaxed EM algorithm to converge. complexity of the LARS algorithm is $O(pqn + pq^2 + q^3)$ (Bach et al., 2012). SpinyReg therefore realizes a complexity tradeoff between slow MCMC Bayesian techniques and fast ℓ_1-based methods.

Let us also emphasize that, whereas frequentist methods use cross-validation to optimize the prediction performance, spinyReg automatically estimates its hyper-parameters. In particular, its inference procedure includes the estimation of the penalty term α which is linked to the sparsity level. Therefore, the computational cost of spinyReg has to be compared to the one of ℓ_1-based methods with the cross-validation included.

5.4 Path of Models

We rely on $\hat{\mathbf{z}}^{\text{relaxed}}$ to find a path of models which are likely to have a high evidence. We build a path by assuming that the larger the coefficients of $\hat{\mathbf{z}}^{\text{relaxed}}$ are, the more likely they are to correspond to relevant variables.

We define the set of vectors $(\hat{\mathbf{z}}^{(k)})_{k \leq p}$ as the binary vectors such that, for each k, the k top coefficients of $\hat{\mathbf{z}}^{\text{relaxed}}$ are set to 1 and the others to 0. For example, $\hat{\mathbf{z}}^{(1)}$ contains only zeros and a single 1 at the position of the highest coefficient of $\hat{\mathbf{z}}^{\text{relaxed}}$. The set of vectors $(\hat{\mathbf{z}}^{(k)})_{k \leq p}$ defines a path of models to look at for model selection. Note that this path allows us to deal with a family of p models (ordered by sparsity) instead of 2^p, allowing our approach to deal with a large number of input variables. Thus, the evidence is evaluated for all $\hat{\mathbf{z}}^{(k)}$ and the number $\hat{q}$ of relevant variables is chosen such that the evidence is maximized:

$$\hat{q} = \text{argmax}_{1 \leq k \leq p} p(\mathbf{Y} | \hat{\mathbf{z}}^{(k)}, \hat{\alpha}, \hat{\gamma}) \quad \text{and} \quad \hat{\mathbf{z}} = \hat{\mathbf{z}}^{(\hat{q})}. \tag{15}$$

Algorithm 2: Model selection algorithm

Input: $\mathbf{X}$,$\mathbf{Y}$,$\hat{\alpha}$,$\hat{\gamma}$,$\hat{\mathbf{z}}^{\text{relaxed}}$

Output: $\mathbf{z}$

for $k = 1..p$ **do**

  Compute $\hat{\mathbf{z}}^{(k)}$;

$\hat{q} = \text{argmax}_{1 \leq k \leq p} p(\mathbf{Y}|\hat{\mathbf{z}}^{(k)}, \hat{\alpha}, \hat{\gamma})$;

$\hat{\mathbf{z}} = \hat{\mathbf{z}}^{(\hat{q})}$;

6 Numerical comparisons

In this section, we illustrate the behavior of spinyReg on simulated and real data sets, and compare it to the most efficient state-of-the-art methods.

6.1 Simulation setup

In order to consider a wide range of scenarios, we use three different simulation scenarios: "uniform", "Toeplitz" and "blockwise". The simulation of the parameter $\mathbf{w}$ and of the noise $\boldsymbol{\varepsilon}$ is common for the three schemes: $\mathbf{w} \sim \mathcal{N}(0, I_p/\alpha)$ and $\boldsymbol{\varepsilon} \sim \mathcal{N}(0, I_n/\gamma)$. The design matrix $\mathbf{X}$ is simulated according to a Gaussian distribution with zero mean and a covariance matrix R depending on the chosen scheme. The correlation structure of $R = (r_{ij})_{i,j=1,...,p}$ is as follows:

- "uniform": $r_{ii} = 1$ for all $i = 1, ...p$ and $r_{ij} = \rho$ for $i, j = 1, \ldots, p$ and $i \neq j$,
- "Toeplitz": $r_{ii} = 1$ for all $i = 1, ...p$ and $r_{ij} = \rho^{|i-j|}$ for $i, j = 1, \ldots, p$ and $i \neq j$,
- "blockwise": $R = \text{diag}(R_1, ..., R_4)$ is a 4-blocks diagonal matrix where R_ℓ is such that $r_{\ell ii} = 1$ and $r_{\ell ij} = \rho$ for $i, j = 1, \ldots, p/4$ and $i \neq j$.

These three correlation structures are represented on Figure 2.

Then, $\mathbf{Z}$ is simulated by randomly picking q active variables among p. The predictive vector Y is finally computed according to Equation (1).

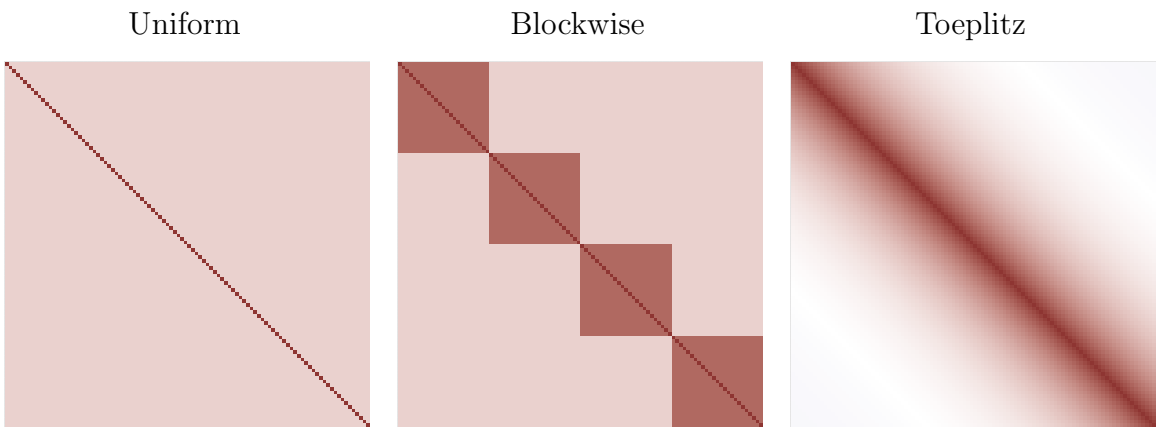


Figure 2: Covariance structures for the simulation setup.

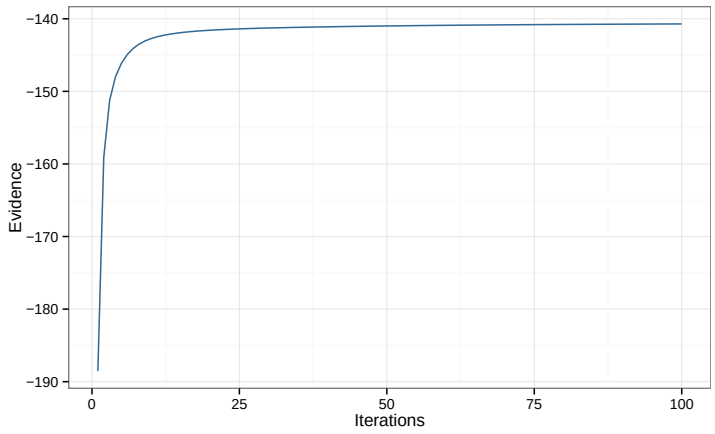
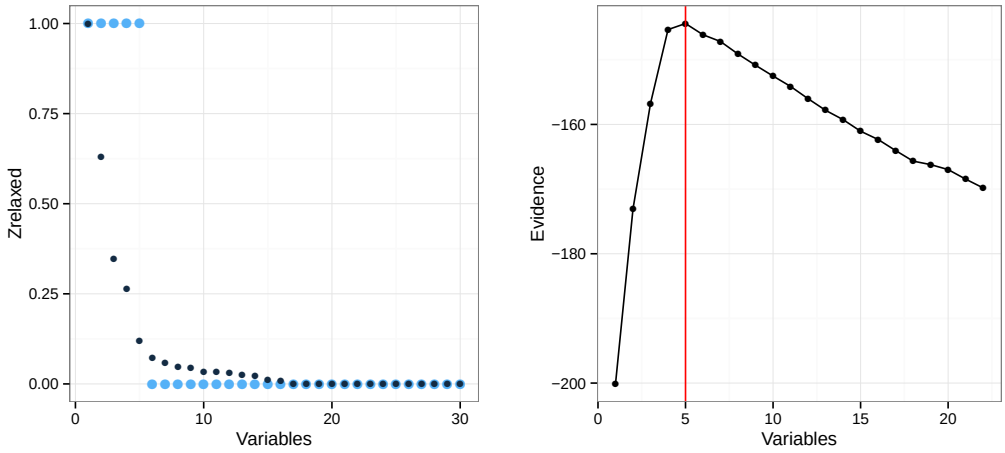


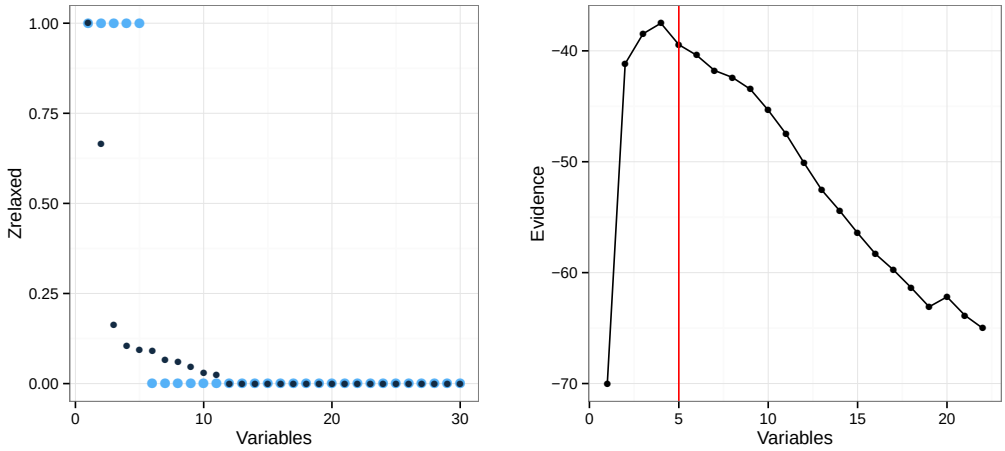
Figure 3: Evolution of the evidence of the relaxed model along the iterations of the EM algorithm.

6.2 An introductory example

We consider here an introductory example which aims at highlighting the main features of the proposed approach. For this experiment, the Toeplitz simulation setup is used with $p = 30$, $q = 5$, $\rho = 0.25$, $\alpha = 1$ and $\gamma = 1$. From this setup, two data sets were simulated with respectively $n = 100$ and $n = 30$ observations. The second setting corresponds to a difficult scenario where $n = p$ whereas the first one should be easier to fit. Notice that the dimensionality is kept relatively low mainly for visualization purpose. Figure 3 shows the evolution of the evidence of the relaxed model along the iterations of the EM algorithm.



Toeplitz setup with $\rho = 0.25$, $p = 30$ and $n = 100$



Toeplitz setup with $\rho = 0.25$, $p = 30$ and $n = 30$

Figure 4: Variable selection with spinyReg on the two introductory examples ($p = 30$ and $n = 150$ or $n = 30$). The left panels present the values of $\hat{\mathbf{z}}^{\text{relaxed}}$ (dark blue) and the actual binary values of $\mathbf{z}$ (pale blue). The right panels show the values of evidence computed on the path of models.

Figure 4 presents the results of the application of spinyReg on those two data sets. The left panels present in dark blue the values of $\hat{\mathbf{z}}^{\text{relaxed}}$ (sorted in decreasing order) and the corresponding true values of $\mathbf{z}$ (pale blue points) used in the simulations. The right panels show the values of evidence computed on the path of models.

Regarding the first example, one can see that the five largest values of $\hat{\mathbf{z}}^{\text{relaxed}}$ actually

correspond to the five active variables. This confirms that spinyReg succeeds here in finding the relevant variables in the regression model. The second panel confirms that spinyReg would select five variables among the 30 original ones. On this quite simple example, spinyReg yields a true positive rate (TPR) equals to 1 and a false positive rate (FPR) equals to 0.

For the second and much more difficult situation (bottom row of Figure 4), the estimated values for $\mathbf{z}^{\text{relaxed}}$ are less discriminative. Indeed, the values of $\hat{\mathbf{z}}^{\text{relaxed}}$ are smaller than in the simpler case. However, even though the ranking of variables induced by $\hat{\mathbf{z}}^{\text{relaxed}}$ respects the partition between active and inactive variables, Occam's razor leads to a too conservative choice and misses one active variable. On this more difficult data set, spinyReg yields a true positive rate (TPR) equals to 0.8 and a false positive rate (FPR) equals to 0.

6.3 Benchmark study on simulated data

We now compare the performance of spinyReg with three of the most recent and popular variable selection methods based on ℓ_1 regularization: the lasso of Tibshirani (1996), the adaptive lasso of Zou (2006) and the stability selection of Meinshausen and Bühlmann (2010). We also added two very recent spike-and-slab approaches: the multi-slab framework of CLERE (Yengo et al., 2014a) and the EP procedure of Hernández-Lobato et al. (2013). To this end, we simulated 100 data sets for each of the three simulations schemes (uniform, Toeplitz and blockwise), for three data set sizes ($n = p/2$, $n = p$, $n = 2p$) and two values for the correlation parameter ($\rho = 0.25$ and $\rho = 0.75$). The other simulation parameters were $p = 100$, $q = 40$, $\alpha = 1$ and $\gamma = 1$. The measures used to evaluate the method performances are the prediction mean square error on test data (MSE, hereafter), the F-score (the harmonic mean of precision and recall, which provides a good summary of variable selection performances) and the estimated value of q (number of relevant predictors).

Lasso and Stability selection were trained using the R package `quadrupen` (Grandvalet et al., 2012). We used the package `parcor` (Kraemer et al., 2009) to train the adaptive lasso and the package `clere` (Yengo et al., 2014b) to train CLERE. The spike-and-slab approach of Hernández-Lobato et al. (2013), which uses expectation propagation, will be subsequently denoted SSEP and was trained using the code available on the authors' web

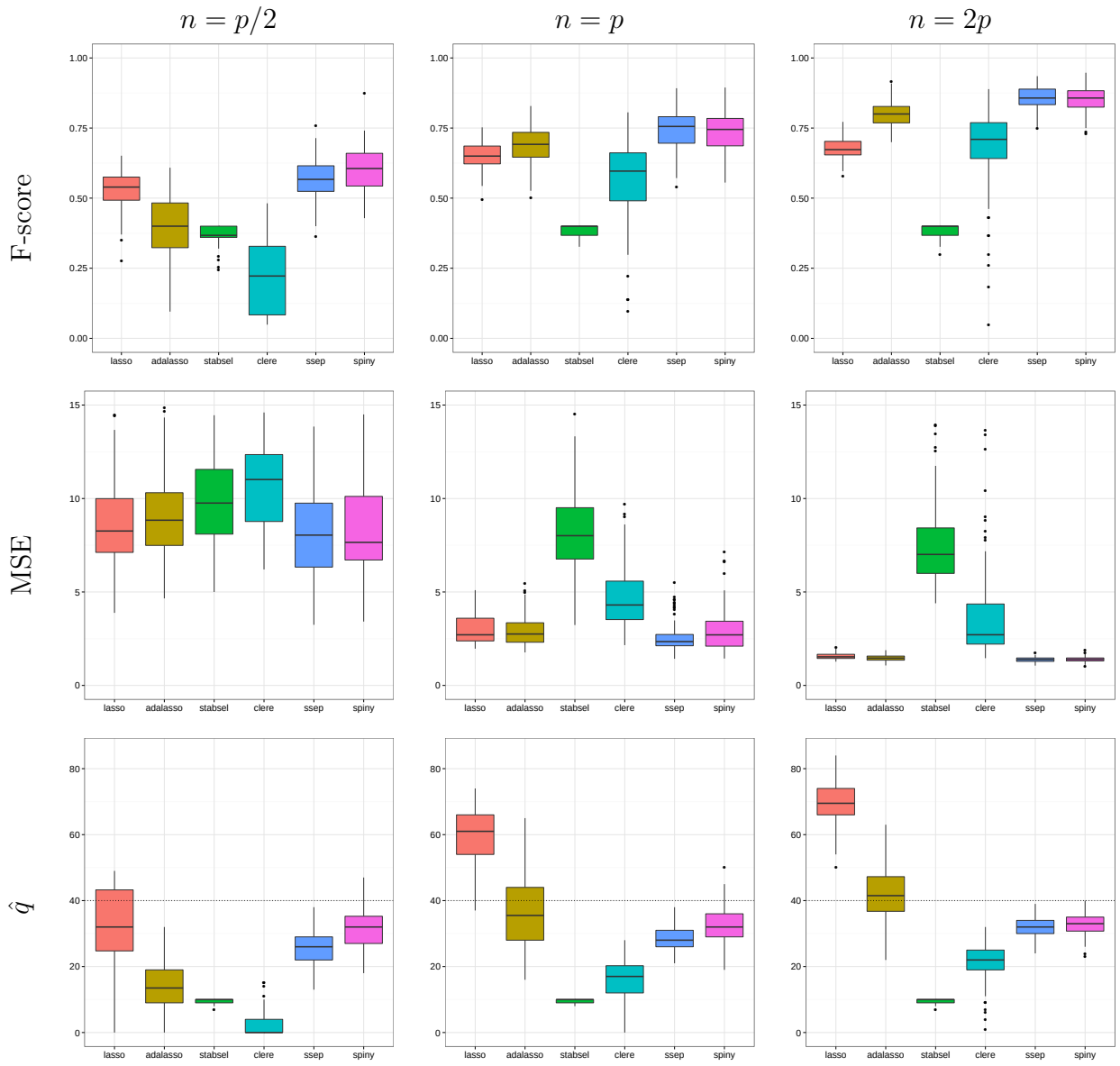


Figure 5: Scenario "blockwise" with $\rho = 0.75$.

pages.

We present here only the results for two simulation setups: the "blockwise" one with $\rho = 0.75$ and the "Toeplitz" one with $\rho = 0.25$. All the other results are available as supplementary material. Note that similar conclusions can be drawn on these other scenarios. Figure 5 presents the F-score, MSE and $\hat{q}$ of the 6 studied methods for the blockwise simulation setup with $\rho = 0.75$ and for the three data set sizes, while Figure 6 presents these measures for the Toeplitz simulation setup with $\rho = 0.25$ and for the three data set sizes.

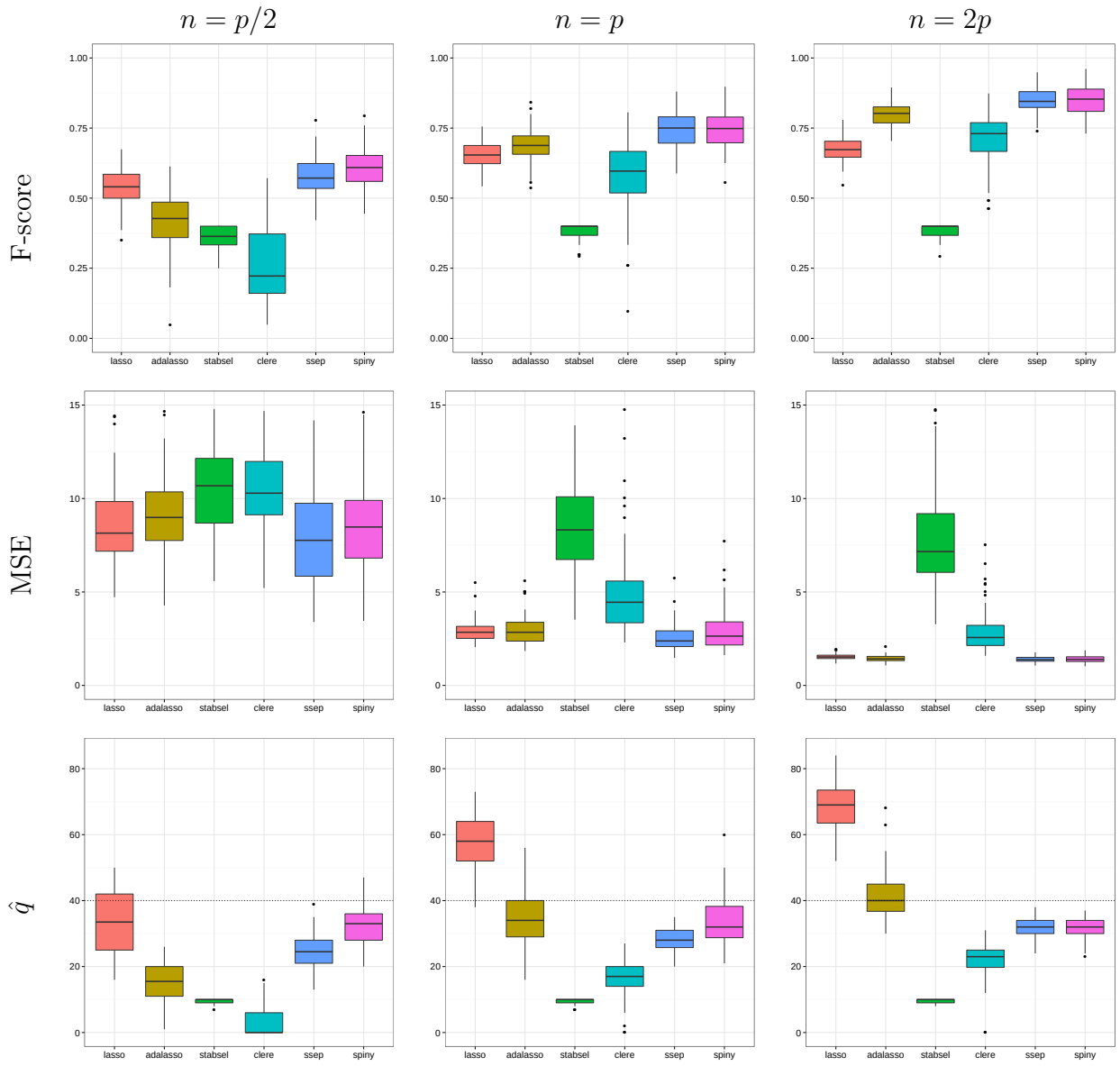


Figure 6: Scenario "Toeplitz" with $\rho = 0.25$.

The first row of Figure 5 and Figure 6 gives the F-score. This measure allows us to figure out how the methods behave in terms of detection of the relevant variables. We can see that spinyReg and SSEP outperform other methods and have close variable selection performances. SpinyReg appears to be at his best in the "$n = p/2$" case on these runs.

The second row of Figure 5 and Figure 6 provides the MSE values for the studied methods. Most of the methods perform well except stability selection and CLERE when $n \leq p$. In particular, spinyReg has the best prediction performance for $n = p/2$ with the

highly correlated blockwise case.

The last row of Figure 5 and Figure 6 gives the number q of active variables estimated by the 6 methods. We remind that the actual number of active variables is $q = 40$ for these simulations (represented by the dashed lines on Figure 5). It is worth noticing that lasso has a clear tendency to overestimate the number of active variables, particularly when n becomes large. Conversely, stability selection has the opposite behavior and underestimates q. Its very conservative behavior has the advantage that it avoids false-positives. It turns out that spinyReg provides consistently a good estimate of the actual value of q.

6.4 Study on classical regression data sets

We now consider four real-world data sets: the classical `prostate` data set used for example by Tibshirani (1996), the `eyedata` data set of Scheetz et al. (2006), which contains gene expression data of mammalian eye tissue samples, the `OzoneI` data set included in the `spikeslab` package (Ishwaran et al., 2010) and which uses the ozone data set of Breiman and Friedman (1985) with some additional interactions and the `DiabetesI` data set which is also available in the `spikeslab` package and uses the diabetes data set of Efron et al. (2004) with some additional interactions. Applying the same methods as before, we trained our data randomly using 80% of the observations and computed the test error on the remaining data. Repeating this procedure 100 times, we computed the mean and the standard deviation of the test error and of the number of variables selected. Results are reported in Table 1. We did not compute the test error for methods which did not succeed in selecting variables.

We can see that spinyReg obtains competitive predictive results on all data sets. Moreover, we can note that it is less conservative than most other algorithms. On the challenging `eyedata` data set for example, while the two other Bayesian methods fail to select at least one variable, spinyReg selects three quarters of the predictors and has the lowest MSE. The three ℓ_1 based methods select only a few variables and have higher MSE. It is worth noticing that we tried to apply the elastic net of Zou and Hastie (2005) (which, using a ℓ_1-ℓ_2 regularization, is able to select more variables than most classical ℓ_1 procedures) to this data set. Elastic net selected all variables. This behavior is close to the one of spinyReg

	Prostate ($n = 77$, $p = 8$)		Eyedata ($n = 96$, $p = 200$)	
	MSE×100	Selected variables	MSE×100	Selected variables
Lasso	63.6 ± 21.8	3.33 ± 0.877	1.26 ± 0.964	16.7 ± 5.56
Adalasso	58.4 ± 15.9	4.42 ± 1.57	1.50 ± 1.248	2.4 ± 0.700
Stability Selection	61.6 ± 14.4	1.94 ± 0.239	1.58 ± 0.850	1.7 ± 0.823
Clere	59.8 ± 19.7	2.87 ± 0.825	-	-
SSEP	56.6 ± 15.0	2.76 ± 0.474	-	-
SpinyReg	58.3 ± 15.4	3.34 ± 0.607	1.25 ± 0.920	143 ± 9
	OzoneI ($n = 162$, $p = 134$)		DiabetesI ($n = 353$, $p = 64$)	
	MSE	Selected variables	MSE/1000	Selected variables
Lasso	18.9 ± 4.96	10.3 ± 2.27	3.22 ± 0.407	7.43 ± 2.41
Adalasso	16.84 ± 4.48	8.32 ± 3.16	3.02 ± 0.395	9.31 ± 2.25
Stability Selection	17.9 ± 5.25	9.68 ± 1.10	2.97 ± 0.387	7.77 ± 0.423
Clere	19.6 ± 5.48	5.43 ± 2.55	3.15 ± 0.384	2.33 ± 0.587
SSEP	29.6 ± 10.2	74.8 ± 5.45	3.70 ± 0.647	62.0 ± 1.36
SpinyReg	18.9 ± 5.46	10.79 ± 2.69	3.13 ± 0.376	8.5 ± 1.45

Table 1: Results on real-world data sets

and reminds the interesting analogy between the Occam factor (12) used in spinyReg and the elastic net penalty.

Let us finally highlight that the medium prediction rank of spinyReg is the second best, behind the adaptive lasso. Let us also emphasize that all frequentist methods were trained using cross-validation which optimizes prediction performance. Conversely, SSEP, CLERE and spinyReg automatically estimate their hyper-parameters. In particular, the inference procedure of spinyReg includes the estimations of the penalty term α which is linked to the sparsity level.

7 Prediction of the frequentation of the Orsay museum using bike-sharing data

In this section, we introduce a new regression problem, which aims at predicting the number of visitors of the Orsay museum (Paris) using the activity of the Paris bike-sharing system (*Vélib'*).

7.1 Predicting a touristic index using open data

The emergence of open data systems has brought about a surge of complex data illustrating various social behaviors. In this challenging context, the analysis of bike-sharing systems (BSSs) provides a new insight into the touristic patterns of a city. We therefore wanted to see how well, in a city like Paris, bike-sharing data could predict a touristic index, such as the number of visitors of an important museum.

With nearly three million annual visitors, the Orsay museum is one of the ten most visited museums in the world (Skeggs, 2014). Known for having the vastest collection of impressionist paintings in the world, it holds for example Manet's *Le Déjeuner sur l'herbe* or Van Gogh's *Nuit étoilée sur le Rhône*. The frequentation of the museum at each hour was given as a courtesy by the museum services.

The Paris bike-sharing system, called *Vélib'*, was launched by JCDecaux and the city of Paris in 2007 and is nowadays certainly the most active BSS in Europe. Statistical studies of the *Vélib'* system have been for example conducted by (Bouveyron et al., 2014; Njato Randriamanamihaga et al., 2014). The predictive variables that will interest us for our regression problem are the percentages of parked bikes (or *loadings*) for all the *Vélib'* stations of Paris. These percentages are available through the open data API provided by JCDecaux[1].

7.2 The "OrsayVelib" database

At each hour, the number of visitors present in the museum constitutes the response variable of our regression problem. The predictors are the loadings at each hour of the $p = 1158$

[1]The real time data are available at https://developer.jcdecaux.com/ with an API key.

Vélib' stations in Paris. Only the hours corresponding to opening days (from 8am to 6pm, except Mondays) of the museum are kept. The month of September 2014 constitutes the learning set (with $n = 316$ observations), and the first two weeks of October 2014 the test set (see Figure 7).

This data set, thereafter called the "OrsayVelib" database, has several interesting aspects:

- While most "large p, small n" regression problems inherit their dimensionality from genomics or signal processing, this data set is purely related to social sciences. This illustrates the fact that modern social data can also lead to high-dimensional challenging statistical problems.

- Since the variables are the *Vélib'* stations, a sparse solution can be easily interpretable and visualizable. We would expect the relevant predictors to correspond – at least to some extend – to stations used by the visitors of the Orsay museum. In particular, the behavior of the stations closest to the museum are expected to be of important interest. For visualization purposes, one can plot on a map the location of the selected variables, being able to efficiently interpret the selection.

- The learning/test segregation of the data harshly punishes overfitting. Indeed, while September 2014 (the learning month) corresponded to exceptionally good weather conditions in Paris, whereas October had some rainy days. Since BSS data are naturally heavily linked to the weather, this means that overfitting algorithms will struggle with predicting the number of predictors on rainy days (such as October 8th). This interesting behavior is exhibited in the next subsection.

To illustrate the behavior of the data, Figure 7 provides the curve of the number of visitors during the learning and test phases and Figure 8 shows the loadings of four *Vélib'* stations during the first week of September. Two of these stations correspond to touristic areas with different behaviors: one is the closest one to the Orsay museum and one is one of the closest ones to the Eiffel tower. The other two correspond to large railway stations (which also happen to be large subway stations). We will show in the next subsection that

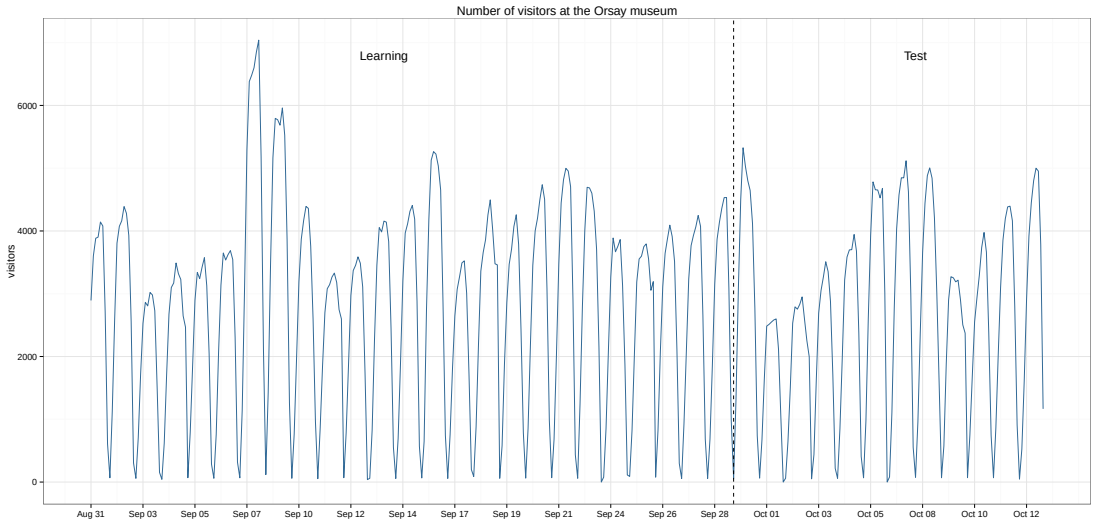


Figure 7: Number of visitors during the learning and test phases. Only opening hours of the museum (8am to 7pm, from Tuesday to Sunday) are shown.

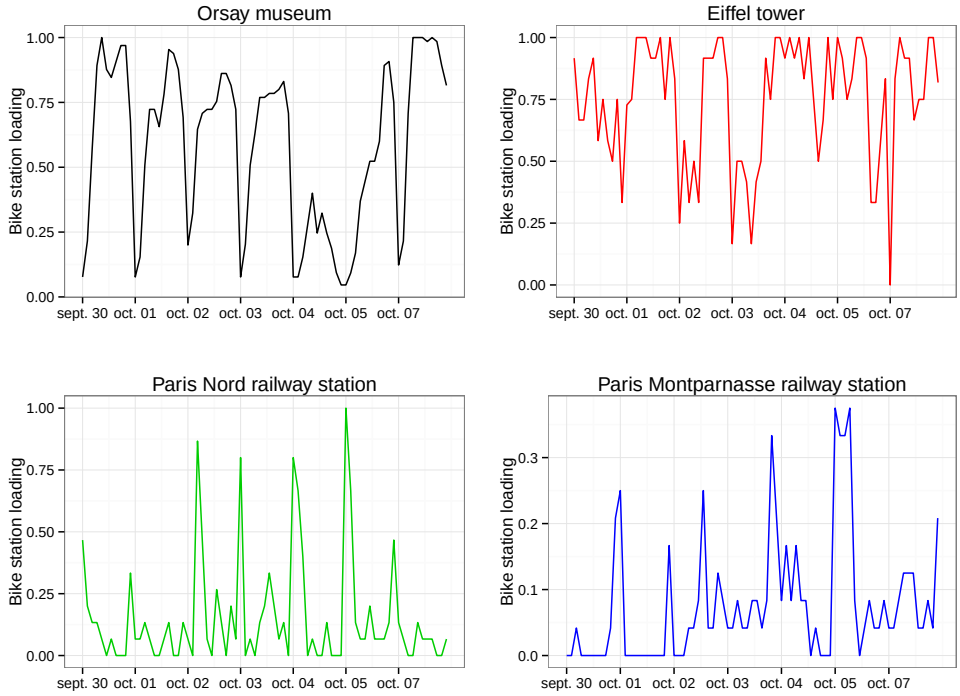


Figure 8: Loadings of four *Vélib'* stations during the first week of September. Only opening hours of the museum (8am to 7pm, from Tuesday to Sunday) are shown.

	Ridge	SSEP	Lasso	Adalasso	SpinyReg
MSE$\times 10^4$	145.66	144.38	132.08	159.17	127.36
Selected variables	1158	1146	167	155	45

Table 2: Test error and number of selected predictors for each method.

these stations are of particular interest if we aim at predicting the number of visitors of the museum.

7.3 Results

We applied the algorithms of Section 6 to the OrsayVelib database. Since the sparsity of this regression problem is not absolutely certain, we also added a non-sparse method to the benchmark: ridge regression with a cross-validated regularization parameter. The test errors and sparsity patterns obtained are detailed in Table 2 (for the sake of clarity, only the five best methods are displayed). One can notice that spinyReg has the lowest generalization error and that it selects fewer variables than its competitors.

Figure 9 allows to compare the true number of visitors during the test phase with the predicted values of the four methods. We can notice that, as expected, all algorithms struggle with October 8th, which was a rainy day. On this specific day, spinyReg is (especially in the afternoon) the closest one to the truth. In a similar fashion, spinyReg is the only method that accurately predicts the small augmentation of the first three days of October.

Eventually, one can plot the location of the selected variables on the map of Paris. For the sake of clarity, we only did it for the two best methods: lasso and spinyReg. Figure 10 presents the maps of selected stations by both methods. Green dots correspond to positive coefficients and red dots to negative coefficients. The dot size indicates the magnitude of the coefficient (the larger the dot, the larger the absolute value of the coefficient). The black dot corresponds to the location of the Orsay museum.

The lasso selection appears to be very broad and difficult to interpret. In particular, the lasso does not select the closest station to the museum. Conversely, the spinyReg selection is more interpretable: one can see that it does select the closest stations to the museum, and that their regression coefficients are positive (which means that these stations are

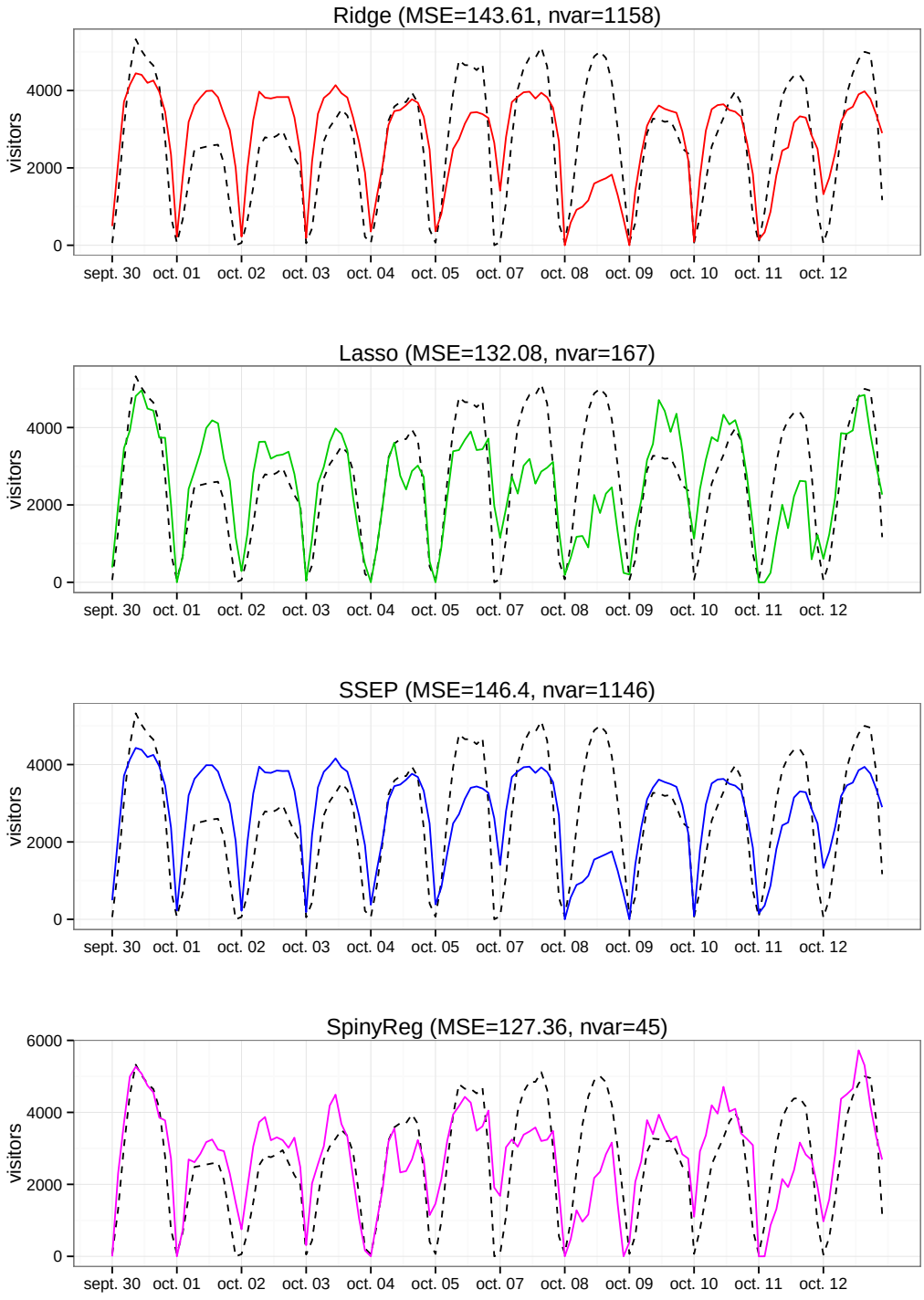
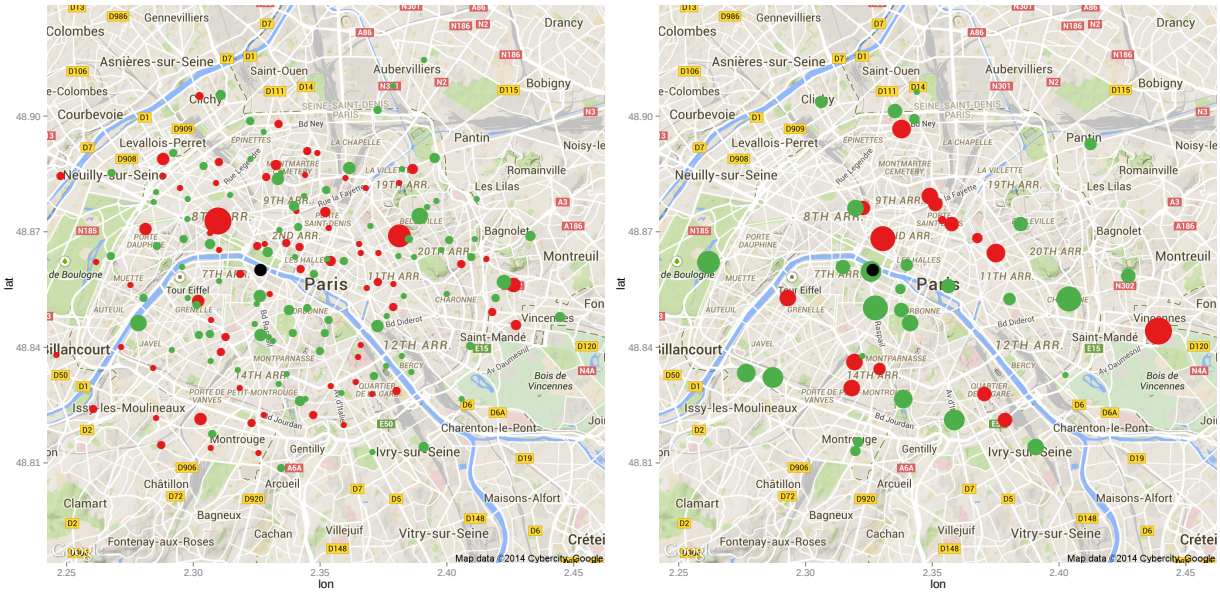


Figure 9: Ground truth (dashed line) and predicted values for the number of visitors at each hour.



Lasso (MSE=132.08, 167 var.) SpinyReg (MSE=127.36, 45 var.)

Figure 10: Stations selected by the lasso (left) and by spinyReg (right). Green dots correspond to positive coefficients and red dots to negative coefficients (the larger the dot, the larger the absolute value of the coefficient). The black dot corresponds to the location of the Orsay museum.

likely to be full when the museum is crowded). Around the neighborhood of the museum, there is a ring of stations with almost exclusively negative coefficients (Eiffel tower, Paris Nord and Montparnasse railway stations, place de la Bastille) which can be interpreted as stations from where the visitors of the museum rent their bikes. Beyond this ring, the selected stations essentially correspond to popular public parks (bois de Vincennes, parc Montsouris, parc André Citroën, bois de Boulogne). This is not surprising since their frequentation is also linked to the touristic activity of the city.

As a summary, spinyReg both succeeds in providing an interpretable selection of *Vélib'* stations while having the most effective prediction performance.

8 Conclusion

We considered the problem of Bayesian variable selection for high-dimensional linear regression through a sparse generative model. The sparsity is induced by a deterministic binary

vector which multiplies with the Gaussian regressor vector. The originality of the work was to consider its inference through relaxing the model and using a type-II log-likelihood maximization based on an EM algorithm. Model selection can be performed relying on Occam's razor and on a path of models found by the EM algorithm. Numerical experiments on simulated data have shown that spinyReg performs well compared to the most recent competitors both in terms of prediction and of selection, especially in moderately sparse cases and with highly correlated predictors. SpinyReg was finally applied for the prediction of a touristic index from open data. The OrsayVelib, a new high-dimensional regression database, was introduced to this end and allowed us to illustrate the powerful aspects of the proposed method.

SUPPLEMENTARY MATERIALS

Additional benchmark results: Boxplots corresponding to all scenarios described in Section 6 (.pdf file)

References

H. Akaike. Information theory and an extension of the maximum likelihood principle. In *2nd International Symposium on Information Theory.*, pages 267–281. Budapest: Akademia Kiado. B. N. Petrov and F. Csaki, editors., 1973.

P. Alquier and K. Lounici. Pac-bayesian bounds for sparse regression estimation with exponential weights. *Electronic Journal of Statistics*, 5:127–145, 2011.

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Optimization with sparsity-inducing penalties. *Foundations and Trends® in Machine Learning*, 4(1):1–106, 2012.

F. R. Bach. Bolasso: model consistent lasso estimation through the bootstrap. In *Proceedings of the 25th international conference on Machine learning*, pages 33–40. ACM, 2008.

R. B. Bapat and T. E. S. Raghavan. *Nonnegative matrices and applications*, volume 64. Cambridge University Press, 1997.

M. Baragatti and D. Pommeret. A study of variable selection using g-prior distribution with ridge parameter. *Computational Statistics & Data Analysis*, 56(6):1920–1934, 2012.

C.M. Bishop. *Pattern recognition and machine learning*. Springer-Verlag, 2006.

C. Bouveyron, E. Côme, and J. Jacques. The discriminative functional mixture model for the analysis of bike sharing systems. *Preprint HAL n°01024186, Laboratoire MAP5, Université Paris Descartes*, 2014.

L. Breiman and J. H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598, 1985.

F. Bunea, A. B. Tsybakov, and M. H. Wegkamp. Aggregation for gaussian regression. *The Annals of Statistics*, 35(4):1674–1697, 2007.

R.H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *Journal on Scientific and Statistical Computing*, 16:1190–1208, 1995.

E. Candès. Mathematics of sparsity (and a few other things). In *Proceedings of the International Congress of Mathematicians, Seoul, South Korea*, 2014.

E. Candès and T. Tao. The Dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

S. S. Chen, D. L. Donoho, Michael, and A. Saunders. Atomic decomposition by basis pursuit. *SIAM Journal on Scientific Computing*, 20:33–61, 1998.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood for incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, B39:1–38, 1977.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2):407–499, 2004.

W. J. Fu. Penalized regressions: the bridge versus the lasso. *Journal of computational and graphical statistics*, 7(3):397–416, 1998.

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

E. I. George and R. E. McCulloch. Variable selection via gibbs sampling. *Journal of the American Statistical Association*, 88(423):881–889, 1993.

Y. Grandvalet, J. Chiquet, and C. Ambroise. Sparsity by worst-case quadratic penalties. Technical report, arXiv preprint, 2012. URL `http://arxiv.org/abs/1210.2077`.

I. Guyon, A. Saffari, G. Dror, and G. Cawley. Model selection: Beyond the bayesian/frequentist divide. *The Journal of Machine Learning Research*, 11:61–87, 2010.

D. Hernández-Lobato, J. M. Hernández-Lobato, and P. Dupont. Generalized spike-and-slab priors for bayesian group feature selection using expectation propagation. *The Journal of Machine Learning Research*, 14(1):1891–1945, 2013.

H. Ishwaran and J. S. Rao. Spike and slab gene selection for multigroup microarray data. *Journal of the American Statistical Association*, 100(471):764–780, 2005a.

H. Ishwaran and J. S. Rao. Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*, pages 730–773, 2005b.

H. Ishwaran, U. B Kogalur, and J. S. Rao. spikeslab: Prediction and variable selection using spike and slab regression. *R Journal*, 2(2), 2010.

N. Kraemer, J. Schaefer, and A.-L. Boulesteix. Regularized estimation of large-scale gene regulatory networks using gaussian graphical models. *BMC Bioinformatics*, 10(384), 2009.

F. Liang, R. Paulo, G. Molina, M. A. Clyde, and J. O. Berger. Mixtures of g-priors for bayesian variable selection. *Journal of the American Statistical Association*, 103(481), 2008.

D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992.

D. J. C. MacKay. Comparison of approximate methods for handling hyperparameters. *Neural computation*, 11(5):1035–1068, 1999.

T. L. Markham. Oppenheim's inequality for positive definite matrices. *American Mathematical Monthly*, pages 642–644, 1986.

G.J. McLachlan and T. Krishnan. *The EM Algorithm and Extensions. Second Edition.* John Wiley & Sons, New York, 2008.

N. Meinshausen and P. Bühlmann. Stability selection. *Journal of the Royal Statistical Society: Series B*, 27, 2010.

T. J. Mitchell and J. J. Beauchamp. Bayesian variable selection in linear regression (with discussion). *Journal of the American Statistical Association*, 83:1023–1036, 1988.

B. K. Natarajan. Sparse approximate solutions to linear systems. *SIAM journal on computing*, 24(2):227–234, 1995.

A. Njato Randriamanamihaga, E. Côme, L. Oukhellou, and G. Govaert. Clustering the vélib'dynamic origin/destination flows using a family of poisson mixture models. *Neurocomputing*, 2014.

R. B. O'Hara and M. J. Sillanpää. A review of bayesian variable selection methods: what, how and which. *Bayesian analysis*, 4(1):85–117, 2009.

A. Oppenheim. Inequalities connected with definite hermitian forms. *Journal of the London Mathematical Society*, 1(2):114–119, 1930.

S. Petrone, J. Rousseau, and C. Scricciolo. Bayes and empirical bayes: do they merge? *Biometrika*, 2014.

B. M. Pötscher and H. Leeb. On the distribution of penalized maximum likelihood estimators: The lasso, scad, and thresholding. *Journal of Multivariate Analysis*, 100(9): 2065–2082, 2009.

P. Rigollet and A. Tsybakov. Exponential screening and optimal rates of sparse estimation. *The Annals of Statistics*, 39(2):731–771, 2011.

C. P. Robert and G. Casella. *Monte Carlo statistical methods*, volume 319. Springer, 2004.

V. Ročková and E. I. George. Emvs: The em approach to bayesian variable selection. *Journal of the American Statistical Association*, just-accepted, 2013.

T. E. Scheetz, K.-Y. A. Kim, R. E. Swiderski, A. R. Philp, T. A Braun, K. L. Knudtson, A. M. Dorrance, G. F. DiBona, J. Huang, and T. L. Casavant. Regulation of gene expression in the mammalian eye and its relevance to eye disease. *Proceedings of the National Academy of Sciences*, 103(39):14429–14434, 2006.

J. G. Scott and J. O. Berger. Bayes and empirical-bayes multiplicity adjustment in the variable-selection problem. *The Annals of Statistics*, 38(5):2587–2619, 2010.

T. Skeggs. Special report, visitor figures 2013. *The Art Newspaper*, 23(256), April 2014.

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

M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *The Journal of Machine Learning Research*, 1:211–244, 2001.

D. P. Wipf and S. S. Nagarajan. A new view of automatic relevance determination. In *Advances in neural information processing systems*, pages 1625–1632, 2008.

C. F. J. Wu. On the convergence properties of the em algorithm. *The Annals of statistics*, pages 95–103, 1983.

L. Xu and M. I. Jordan. On convergence properties of the em algorithm for gaussian mixtures. *Neural computation*, 8(1):129–151, 1996.

L. Yengo, J. Jacques, and C. Biernacki. Variable clustering in high dimensional linear regression models. *Journal de la Société Française de Statistique*, 155(2):38–56, 2014a.

L. Yengo, J. Jacques, C. Biernacki, and M. Canouil. Variable clustering in high-dimensional linear regression: The r package clere. *Preprint HAL n°00940929*, 2014b.

P. Zhao and B. Yu. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.

H. Zou. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429, 2006.

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