ENSAE ParisTech
Bayesian Statistics
Professor : Judith Rousseau

EMVS: the EM approach to Bayesian Variable Selection

Ročková (2013); Ročková and George (2014)

Jérémy L'Hour[1], **Nicolas Saleille**[2]

Contents

[1] jeremy.l.hour@ensae-paristech.fr
[2] nicolas.saleille@ensae-paristech.fr

I Introduction

I.1 The high-dimension variable selection problem

Model selection and parsimony amongst explanatory variables are traditional scientific problems that date back at least to the fourteenth century and Occam's razor (*Pluralitas non est ponenda sine necessitate*). They have a particular echo in modern Statistics and have received growing attention over the past decade, following two factors: on one hand the growing computational power we can harness, on the other high-dimensional datasets have become increasingly available to statisticians in various fields. Consequently, the statistician is more frequently than ever faced with a dataset that contains a very large number of explanatory variables that he has to chose from.

Let us first clarify the problem in the framework of the Gaussian homoscedastic linear model:

$$y = X\beta + \epsilon \tag{1}$$

y is n-dimensional vector of the outcome of interest, X the $n \times p$ design matrix and β the p-dimensional vector of coefficients. Let X_i the i-th row of the design matrix. We assume that each copy (y_i, X_i) of elements from y and X is i.i.d. and that ϵ, the error term is a Gaussian vector, uncorrelated with X, of mean zero and diagonal variance-covariance matrix with diagonal elements all equal to σ^2.

The high-dimensional case arises when $n < p$, in which case the OLS estimator cannot be computed. For the applied statistician, this problem may arise even if we have $p \leq n$ but p relatively large compared to n because $X^T X$ may be close to singular. We notice that it encompasses two situations: i) when the dataset possess a large number of raw variables to select from, ii) when instead of assuming a linear regression function, one wants to consider a non-parametric regression function and approximate it using sieve estimation by expending it in a convenient basis.

This high-dimensional curse occurs in a large number of scientific realms:

- ♢ A classical example in genomics is the detection of genes responsible for obesity, in which case n the number of patients can be a few hundreds as DNA microarrays are costly, while we may consider $p = 300,000$ genes.
- ♢ In policy evaluation, variable selection mistakes may prevent identification of the causal effect of a treatment when using the Conditional Independence Assumption (CIA) (Imbens and Wooldridge, 2008; Givord, 2010). This assumption amounts to say that conditionalonto a set of well-identified observables, the outcome of the treatment is independent of taking or not the treatment. Considering a large number of variables and a flexible functional form strengthens the robustness of the results. Several authors have documented the problem and proposed Lasso-type estimators to select covariates, for example: Belloni *et al.* (2012, 2013); Farrell (2013)
- ♢ In economics, when one wants to consider a large number of regressors but assume that only a few of them are actually significantly related to the outcome, as a way to have an agnostic look at the data. It is often the role of economic theory to justify the use of a regressor, but the researcher may want to try all the available variables rather than relying on prior beliefs given by the economic theory (see for example Sala-i Martin (1997) in the context of finding growth determinants). Considering a high-dimensional X is a way to take a step back and recover the sparsity pattern instead of simply assuming it.

The high-dimensional literature is concerned with two main tasks which objectives are somewhat contradictory. The first one is *prediction*: how to best extract information from the variables contained in the design matrix to predict the outcome ? The second one is *estimation*: how to accurately estimate a model's parameters in a way which facilitate interpretability and hereby the understanding of a phenomenon ? The first task often requires the use of methods that are well-understood in their functioning

but that look like a black box when one tries to explain it to a non-statistician (*e.g.* model averaging, SVM). The second task calls for a simple model precisely because we want to gather a few features that our brain will be able to retain in order to get a better understanding of a problem.

I.2 Sparisty and penalization as a way to overcome the curse of dimensionality

A convenient way to deal with high-dimensional models is to assume the *sparsity* in β, *i.e.* only $s << n$ elements of β are non-zero. We call *sparsity pattern* the set of non-zero components of β: $J(\beta) := \{j : \beta_j \neq 0\}$. In the detection of the obesity genes problem, it means that we expect only two or three genes to be responsible for obesity and all others to have an insignificant impact (*i.e.* to be outside the sparsity pattern). Consequently, we want to penalize models that are not parsimonious in their use of the variables.

The traditional, combinatorial, way of selecting such a model would be to estimate all the possible models that include only a given number of variables s ($s = 1, ..., \min(n, p)$) by OLS and take the one that gives the best fit (denoted by $M_s^\star$). Then compare models $M_1^\star, ..., M_{\min(n,p)}^\star$ using a criterion that penalizes the number of variables included in the model such as the BIC of Schwarz (1978). However, in the obesity genes example, this would requires to estimate $2^{\min(n,p)}$ models by least squares, which isn't technically feasible. This technical limitation comes from the fact that non-zero elements are penalized by mean of the ℓ_0-norm which isn't convex. Fairly recently, a lot of attention has been devoted to estimators that penalize non-zero elements of β in a way that makes computation easier, namely with a convex function. The most famous of these estimators is the Lasso of Tibshirani (1996) that uses the closest convex function - the ℓ_1-norm. The Lasso is defined as the result of the following minimization program:

$$\min_{\beta} n^{-1} \sum_{i=1}^{n} \left(y_i - x_i^T \beta\right)^2 + \lambda ||\beta||_{\ell_1} \tag{2}$$

When $\lambda = 0$, we obtain the OLS, while β is null when $\lambda = +\infty$. Properties of the Lasso and optimal choice for the penalty level λ have been studied in Bickel *et al.* (2009); Meinshausen and Yu (2009); Lounici (2008); Belloni and Chernozhukov (2010); Zhao and Yu (2006); Zhang and Huang (2008). However, these penalty functions are far from the only available choices and it is easy to be lost.

According to Fan and Li (2001), a good penalty should display the following three properties that translate into mathematical properties of the penalty functions:

- ♢ *sparsity*: unimportant coefficients should be set to zero so as to operate variable selection.
- ♢ *unbiasedness*: large coefficients should not be shrunk unnecessarily to zero so as to avoid biasing the estimation.
- ♢ *continuity in data*: small perturbation in data should be discounted so as to have a robust estimation and a stable prediction.

It is to be noted that no simple penalty function fulfills all these conditions simultaneously.

I.3 From frequentist penalization to Bayesian prior

So far, our implicit framework was frequentist. But we can also cast the problem in a Bayesian framework. Recall that the posterior distribution of a parameter of interest is proportional to the product of the likelihood and the prior:

$$\pi(\beta|y, X) \propto \pi(\beta) \times L(y, X; \beta) \tag{3}$$

When one considers this equation in log, it can easily be seen that maximizing the posterior distribution (to obtain what is called the MAP) is equivalent to perform a penalized likelihood estimation where

the prior plays the role of the penalty term. Prior distributions that induce sparsity are to be centered around zero. The lower the variance of the prior distribution, the higher the number of coefficients to be set to zero. As an example, the Lasso estimator can be recovered as a MAP with a Laplace prior distribution. The regularization parameter λ corresponds to the inverse of the prior variance.

A popular prior is the spike-and-slab distribution which is a mixture between a Dirac mass at zero and a Gaussian distribution. The first component drives the coefficients to be exactly zero, while the other allows for nonzero entries. The mixing proportion between the two distributions plays the same role as the regularization parameter. However, the mass at zero poses significant computational difficulties. Several solutions propose replacing the Dirac mass by a Normal distribution centered at zero with a very small but positive variance, such as the Stochastic Search Variable Selection (SSVS) prior of George and McCulloch (1993).

Bayesian methods to handle high-dimensional problems are very popular for prediction in economics and finance. In macroeconomics in particular, Bayesian methods that favor a parsimonious model have given great results since the first works of Litterman (1986) and Bayesian hierarchical models are now one of the main tools for prediction in several Central Banks. The most popular papers regarding the subject are: Giannone *et al.* (2010); De Mol *et al.* (2008); Giannone *et al.* (2012). However, for example in Giannone *et al.* (2012) as it is a hierarchical model that requires to run a Metropolis-Hastings algorithm, computation of the posterior distribution takes a long time which is a burden when one wants to explore different specifications and different datasets. The work of Ročková and George (2014) is very interesting in this respect as it is computationally very fast while not giving away the complexity of the model.

In Section II we expose the main features of their EMVS approach. In Section III we apply it to real data and compare it with other popular penalized estimators.

II EMVS PRESENTATION

The EMVS (Expectation - Maximisation applied to Variable Selection) algorithm proposed in Ročková and George (2014) is based on a deterministic approach; its main purpose is to lower the computational burden of Markov-Chain Monte Carlo methods when estimating posterior distributions over subsets of potential predictors. This algorithm is thus particularly appropriate in high-dimensional settings with $n < p$.

The EMVS is based on the continuous **spike-and-slab** normal mixture model and on a closed form expression of the EM algorithm. Variable selection is achieved through two important assumptions on the spike distribution, namely continuity and a positive variance parameter that introduces sparsity in the selection process. Once posterior modes have been discovered, model evaluation is carried out in a second step using a point mass spike distribution. In the following subsections we present (1) the conjugate spike-and-slab formulation, (2) the closed form EM algorithm, and (3) the resulting variable selection procedure. Finally, we briefly introduce extensions suggested by the authors.

II.1 Conjugate Spike-and-Slab formulation

Suppose we have a set of p potential predictors stacked in a $(n \times p)$ matrix $X = (x_1, \ldots, x_p)$. We put no restriction on p and n and allow for $p > n$. We want to select the best model to predict the $(n \times 1)$ response vector y under the Gaussian Linear model assumption:

$$y|\alpha, \beta, \sigma \sim \mathcal{N}\big(\alpha + X\beta, \sigma^2 I_n\big) \tag{4}$$

For that purpose, define a vector of latent variables $\gamma = (\gamma_1 \ldots, \gamma_p)'$ such that

$$\gamma_i = \mathbb{1}\big\{\beta_i \neq 0\big\} \quad \forall i \in \{1, \ldots, p\} \tag{5}$$

i.e. $\gamma_i = 1$ indicates that variable x_i should be included in the model.

The EMVS approach is based on a Bayesian hierarchical approach and takes part of prior distributions on the model parameters α, β, σ to estimate the posterior distribution of interest, $\pi(\gamma_i|y)$. This distribution is particularly interesting when performing variable selection because it gives the posterior probability that the variable x_i is relevant in the model. For a sufficiently low posterior probability of the event $\{\gamma_i = 1\}$, regressor x_i will be discarded. We now describe the hierarchical structure of the model.

Prior on β. The general set-up underlying the EMVS algorithm is the "Spike-and-Slab" Gaussian mixture prior on β presented in George and McCulloch (1997). This specification is the stepping stone of Bayesian variable selection methods. The prior distribution of β is set to be Gaussian and centered with a mixture component arising from the covariance matrix:

$$\pi(\beta|\sigma, \gamma, v_0, v_1) = \mathcal{N}_p(0, D_{\sigma,\gamma}) \tag{6}$$

where $0 \leq v_0 < v_1$ are hyperparameters such that:

$$D_{\sigma,\gamma} = \sigma^2 \begin{pmatrix} (1-\gamma_1)v_0 + \gamma_1 v_1 & & 0 \\ & \ddots & \\ 0 & & (1-\gamma_p)v_0 + \gamma_p v_1 \end{pmatrix} \tag{7}$$

Looking at the structure of this covariance matrix we see that depending on the value of γ_i, β_i will have a variance parameter given either by v_0 or v_1. More precisely:

$$\mathbb{V}(\beta_i) = v_0 \quad \text{if} \quad \gamma_i = 0 \tag{8}$$

$$\mathbb{V}(\beta_i) = v_1 \quad \text{if} \quad \gamma_i = 1 \tag{9}$$

In a nutshell, v_0 is the variance parameter of the spike distribution and is set to a small value, while v_1 is the variance parameter of the slab distribution and is set to a large value. In a the traditional spike-and-slab formulation, v_0 would be set to zero so as to have a Dirac mass at zero. However, it entails too much computational problems and keeping a slightly positive value for v_0 constraints very small coefficients to be in the spike part of the mixture without exactly being zero. Nevertheless, the γ_i will play the role of selectors in the sense that the probability for β_i to be in the spike part of the mixture is a signal that β_i should be exactly zero. The advantage of having such a specification is that we rae in a presence of a conjugate prior for β which allows a closed form and consequently does not require to use a Gibbs sampler.

Prior on α. Ročková and George (2014) suggests to remove the constant term by working on the centered outcome y and design matrix X.

Prior on σ^2. The prior distribution on the scaling parameter is supposed to be an inverse-gamma:

$$\pi(\sigma^2|\gamma) = \text{IG}(v/2, v\lambda/2) \tag{10}$$

with $v = 1$ and $\lambda = 1$.

Prior on γ. The parameter γ takes 2^p possible values, just as when penalizing with an ℓ_0-norm as we have seen in the introduction; its distribution is specified with respect to an hyperparameter θ with a specification of the form

$$\pi(\gamma) = \mathbb{E}\big[\pi(\gamma|\theta)\big] = \int \pi(\gamma|\theta)\pi(\theta)d\theta \tag{11}$$

If there is no a priori information on which regressor should be or shouldn't be included in the model (i.e. we don't have structural information on γ), a non informative choice is the i.i.d. Bernoulli prior

$$\pi(\gamma|\theta) = \theta^{\sum_{i=1}^n \gamma_i}(1-\theta)^{p-\sum_{i=1}^n \gamma_i} \tag{12}$$

This specification means that each variable has a probability θ to be included in the model.

Prior on θ**.** To complete the hierarchical structure parameter θ is assigned a beta distribution

$$\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1} \tag{13}$$

where a and b are positive reals arbitrarily determined.

II.2 Closed form EM algorithm

The estimation of the posterior distribution $\pi(\gamma|y)$ is not straightforward. Using the previous spike-and-slab formulation, Ročková and George (2014) propose a version of the EM algorithm where closed form solutions exist for the parameters β, σ, θ in the Maximisation step of the algorithm. The resulting method is less computationally intensive than the standard MC-MC stochastic search approach.

The EM algorithm maximises the posterior distribution of the parameters $\pi(\beta, \sigma, \theta)$ indirectly through an iterative process based on the objective function:

$$Q(\beta, \sigma, \theta|\beta^{(k)}, \sigma^{(k)}, \theta^{(k)}) = \mathbb{E}\left[\log \pi(\beta, \sigma, \theta, \gamma|y)\big|\beta^{(k)}, \sigma^{(k)}, \theta^{(k)}, \gamma^{(k)}\right] \tag{14}$$

where $\pi(\beta, \sigma, \theta, \gamma|y)$ is the *complete* posterior distribution, which is unobserved since γ is latent. The algorithm iters the following two steps:

(1) **(E step)** The first step corresponds to the evaluation of the conditional expectation

$$Q(\beta, \sigma, \theta|\beta^{(k)}, \sigma^{(k)}, \theta^{(k)}) \tag{15}$$

(2) **(M step)** The second step corresponds to a maximisation:

$$\left(\beta^{(k+1)}, \sigma^{(k+1)}, \theta^{(k+1)}\right) = \underset{\beta,\sigma,\theta}{\arg\max}\, Q(\beta, \sigma, \theta|\beta^{(k)}, \sigma^{(k)}, \theta^{(k)}) \tag{16}$$

The EM algorithm generates a sequence of parameter estimates which converges monotonically towards a local maximum of the objective posterior $\pi(\beta, \sigma, \theta)$. In the context of the spike-and-slab hierarchical structure Ročková and George (2014) show that the objective function Q has interesting properties that might significantly ease the computational burden of the algorithm. Their results are based on the following decomposition of the conditional expectation (14):

$$Q(\beta, \sigma, \theta|\beta^{(k)}, \sigma^{(k)}, \theta^{(k)}) = C + Q_1\left(\beta, \sigma\big|\beta^{(k)}, \sigma^{(k)}, \theta^{(k)}\right) + Q_2\left(\theta\big|\beta^{(k)}, \sigma^{(k)}, \theta^{(k)}\right)$$

A first simplification comes from the fact that the hierarchical posterior distribution of $\gamma|\beta^{(k)}, \sigma^{(k)}, \theta^{(k)}, y$ depends on y only through the current estimates $\beta^{(k)}, \sigma^{(k)}, \theta^{(k)}$. The hierarchical assumptions produce a simple closed formula to evaluate Q_1. Consequently, the E-step simply consists in the evaluation of this closed expression with no additional algorithm involved. A second simplification comes from the separability of Q: in the M-step, it is possible to maximise Q_1 and Q_2 separately. Analytical solutions for the values of the parameters β, σ, θ that maximise the objective function Q are given in the following paragraphs. Once again, their existence provides substantial computationnal savings.

We will not go into each details of the EM-algorithm, but we think it is important to comment on some of the features in the Maximization step that make the EMVS particularly interesting and see how it selects variables.

Iteration for β At each step, the next value of β is computed as a solution of a generalized Ridge problem:

$$\beta^{(k+1)} = (X^T X + D^\star)^{-1} X^T y \tag{17}$$

We can see that $D^\star$ is the penalty term of the Ridge regression. It is a $p \times p$ diagonal matrix with entries $d_i^\star$ as the individual penalty of the coefficient β_i. We focus on the form of this penalty for a moment.

$$d_i^\star = \frac{1 - p_i^\star}{v_0} + \frac{p_i^\star}{v_1} \tag{18}$$

$p_i^\star$ is the probability of being included in the model at each step. $\frac{1}{v_0}$ is large so when $p_i^\star$ is close to zero, the penalty applied to coefficient β_i is large, and it will be shrunk towards zero. This penalty

combines the data extracted in $p_i^\star$ and prior beliefs regarding the width of each component of the mixture in v_0, v_1. This closed form would not have been possible in the usual spike-and-slab as $v_0 = 0$ in this case. Moreover, it means that $\beta^{(k+1)}$ is defined even if $p >> n$ as the matrix inversion problem has been regularized.

Iteration for σ This iteration does not give a particular insight on the way the EMVS performs variable selection.

Iteration for θ At each, the parameter of the Bernoulli prior that a given variable is included in the model is given by:

$$\theta^{(k+1)} = \frac{a - 1 \sum_{i=1}^{p} p_i^\star}{a + b + p - 2} \tag{19}$$

This is again a combination between prior beliefs and the data.

II.3 The variable selection procedure and spike-and-slab regularization plot

The EMVS helps selecting a sparsity pattern in a specific way, still keeping the idea of using the MAP. Two features are particularly appealing. The first one is that the variable selection procedures operates by thresholding. The second one is the computational simplicity that allows to investigate several configuration of prior distribution.

Thresholding rule The posterior submodel $\hat{\gamma}$ is selected using the following rule:

$$\hat{\gamma} = \arg\max_{\gamma} \mathbb{P}(\gamma | \hat{\beta}, \hat{\theta}, \hat{\sigma}) \tag{20}$$

Since the posterior is i.i.d., the problem is separable and we obtain the following rule:

$$\hat{\gamma}_i = 1 \Leftrightarrow \mathbb{P}(\gamma_i = 1 | \hat{\beta}, \hat{\theta}, \hat{\sigma}) \geq 0.5 \tag{21}$$

Since $\mathbb{P}(\gamma_i = 1 | \hat{\beta}, \hat{\theta}, \hat{\sigma})$ is an increasing function of $|\hat{\beta}|$, equation (21) is equivalent to perform thresholding in the following way:

$$\hat{\gamma}_i = 1 \Leftrightarrow |\hat{\beta}| \geq \hat{\sigma} \sqrt{2 v_0 \log(\omega_i c) c^2 / (c^2 - 1)} \tag{22}$$

where $c^2 = v_1 / v_0$ and $\omega_i = (1 - \mathbb{P}(\gamma_i = 1 | \hat{\sigma})) / \mathbb{P}(\gamma_i = 1 | \hat{\sigma})$. We can note that the larger the estimated noise level $\hat{\sigma}$, the more coefficients will be set to zero.

Regularization plot Since the EMVS is computationally very fast, a nice property is that it is possible to compute the MAP of the coefficients for several values of v_0, which plays the role of a regularization parameter as we have seen in equation (18). Consequently, a "spike-and-slab" regularization plot that works in the same way as the Lasso regularization plot in Hastie *et al.* (2009, p. 65) helps the applied statistician to visualize the results and select the relevant features of the design matrix. As v_0 increases, the negligible coefficients are more and more absorbed in the spike part of the mixture. Contrary to shrinkage estimators such as the Lasso or the Ridge, as v_0 increases, the EMVS does not shrink the large coefficients to zero too much when negligible coefficients are getting closer to zero. This feature is appealing in the sense of the *unbiasedness* property that we have mentioned in the introduction. Section III displays examples of such plots.

Post-Estimation model selection Once the statistician has run the EMVS for many values of the regularization parameter $(v_0^1, \ldots, v_0^K)$, he his facing a set of possible solutions $(\hat{\gamma}^1, \ldots, \hat{\gamma}^K)$ and has to choose one among them. The selection criteria proposed by the authors is based on the maximization of the marginal probability of γ under the prior $v_0 = 0$, denoted $\pi_0(\gamma | y)$. This criteria is particularly interesting since it gives the probability, once we have observed the outcome y, that the accurate model is

indeed $\hat{\gamma}$. This probability integrates and summarizes all the information introduced by the hierarchical spike-and-slab structure. Furthermore it allows to evaluate each model by selecting only the variables for which $\gamma_i = 1$. In the spike-and-slab context $\pi_0(\gamma|y)$ is known up to a normalizing constant C, i.e. a closed form solution allows to compute a function

$$g(\gamma) = C\pi_0(\gamma|y) \tag{23}$$

That will then be evaluated at $(\hat{\gamma}^1, \ldots, \hat{\gamma}^K)$. This function is the one referred to in the graphic outputs for instance in figure 1.

II.4 Extensions

We have presented the workhorse EMVS model. Ročková (2013, Chapter 3) presents several extensions and problems that we briefly mention.

Multimodality As it is common with EM algorithms, getting stuck in a local optimum is a risk. However, as it is computationally very fast, it is possible to overcome that drawback with the EMVS. The author proposes using deterministic annealing as a way to flatten the objective function and facilitate the jump between local modes.

Heavy tail slab distribution The current slab distribution is Normal which is a thin tail distribution. However, this choice has the drawback that large coefficients may be shrunk towards zero too much. To overcome this phenomenon, it is proposed to replace the Normal slab distribution by a fat tail distribution that does not bias large coefficients.

Structured Prior Information Forms for $\pi(\gamma|\theta)$ This part of the paper considers other priors for the inclusion dummy. Instead of assuming that the inclusion of a variable in the model is independent from the inclusion of the others, the author considers different priors where groups have variables are likely to be relevant together.

Stochastic Dual Coordinate Ascent for EMVS Finally, this part is concerned with reducing the computational burden of the Ridge solution for β in the Maximization step, especially when $p > n$.

III Applications and comparison with other selectors

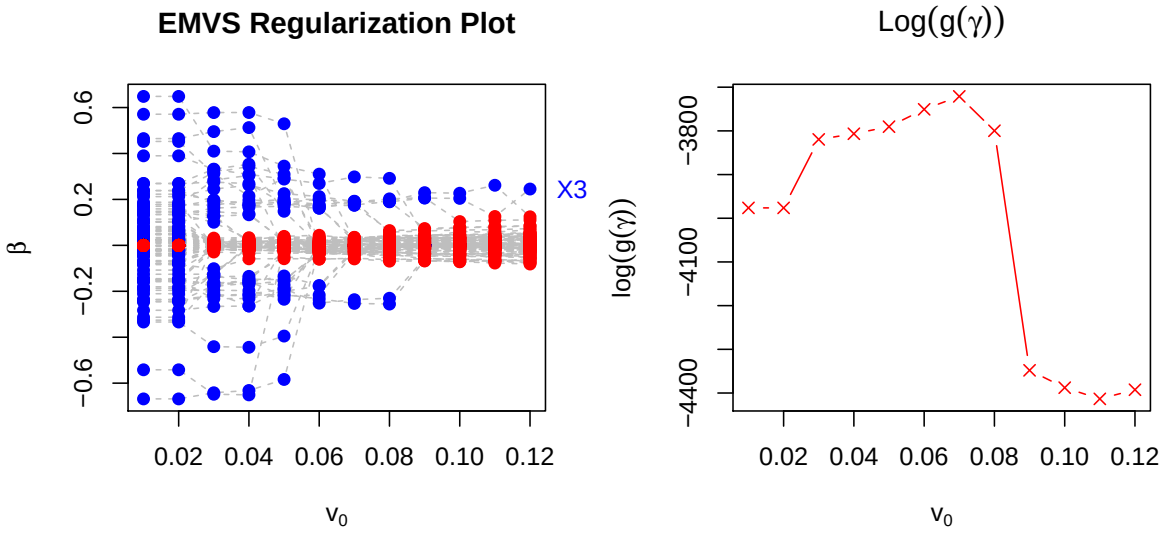
In this part, we use the C++ / R code kindly given by Veronika Ročková.

III.1 An application to violent crime data in the US

We use the EMVS to analyze the Violent Crime dataset[3]. This dataset gathers information regarding socio-demographic, crime and police force features of American cities. Once cleaned of missing cases, this dataset contains 1994 instances of 100 attributes. We try to see which factors influence the violent crime rate, defined as the number of crime per 1000 inhabitants, of US cities using the EMVS.

We run the EMVS for several values of v_0 (from 0.01 to 0.12). We use a Beta-binomial prior on the probability of inclusion, setting $a = b = 1$ and we set $v_1 = 1000$. This computation for twelve values of v_0 took no more than 3 seconds (guesstimate) on our personal computers. We plot the results in Figure 1:

Figure 1: Results from the EMVS on the crime dataset



Note: The left-hand panel shows the regularization plot as a function of v_0. The blue dots are variables that are in the slab part of the distribution. The red dots are the variables that are in the spike part of the distribution so considered as irrelevent. The right-hand panel shows the log of the g function which is a measure of the quality of the model.

The best model found attains a $\log(g)$ value of -3721.08 as can be infered from the plot. It is obtained for $v_0 = 0.07$ and selects 9 variables. Amongst them: the percentage of population that is african american, the percentage of males who are divorced, the percentage of kids with two parents, the percentage of kids born to never married, the percentage of persons in dense housing (more than 1 person per room), the number of vacant houses, the number of rental housing that are in the lower quartile, the median gross rent and the number of homeless people counted in the streets. We can clearly see that underlying these variables are the economic and social status of the living population. Indeed, the black population is amongst the poorest and the least educated in the US. Kids in the least stable families are also more prone to crime. Finally, housing factors plays a big role in the prediction of the crime rate: they are of course supported by economic factors, but we cannot pronounce ourselves on the causality link between the two in the sense that people may also shy away from these location precisely *because* of the crime rate.

A nice feature of the EMVS is the possibility of computing the probability of inclusion for each variable. For the best model found, the EMVS discriminates very well in the sense that no probability is

[3]Source: `https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime`

"in the middle" (*i.e.* close to 0.5): they are either very small or very large. The median of the inclusion probability is .009, the third quartile is 0.001 and the maximum is 1. The bottom line is that regarding variable selection, the EMVS gives a clear-cut answer.

III.2 Comparison with the Lasso

We now run a Lasso on the same dataset to see how it differs from the EMVS. The Lasso penalty level is arbitrarily set; we take it so that the procedure selects the same number of variables as EMVS. The two procedure share 6 covariates over 9. Our results are presented in Table 1.

Table 1: EMVS vs Lasso results

	EMVS	OLS (EMVS selected)	Lasso	Post-Lasso
Nb. vacant households	0.1742	0.1785	0.1333	0.1414
Pctg of divorced males	0.1759	0.1751	0.0772	0.1409
Median gross rent	0.2982	0.2810		
Number of homeless persons in the streets	0.1880	0.1519	0.0772	0.1569
Pctg of kids born to never married	0.1882	0.1616	0.1948	0.1952
Pctg Kids with two parents	-0.2365	-0.2940	-0.3010	-0.3025
Pctg persons in dense housing	0.1912	0.2358	0.0476	0.1024
Pctg people living in urban areas			0.0010	0.0382
Pctg vacant housing that is boarded up			0.0043	0.0434
Pctg Black	0.1922	0.1719		
Pctg White			-0.1771	-0.1730
Rental housing - lower quartile rent	-0.2528	-0.2080		

Two of the three regressors selected only by the Lasso (the percentage of people living in urban areas and the percentage of vacant housing that is boarded up) are associated to small coefficients in absolute terms. Intuitively, these coefficients have been absorbed by the spike distribution of the EMVS. Interestingly, two of the three variables selected only by EMVS are linked to the distribution of rents, for which estimated coefficients are large (in absolute terms). As discussed before it is rather natural to see these factors playing a significant part in the explanation of crime rates, even if causality is not straightforward. To compare the shrinkage bias induced by penalties, we ran OLS on the selected models. The bias induced by EMVS is lower than the one induced by the Lasso. A striking example is the coefficient associated to the number of homeless persons in the streets, which is strongly biased in the first step of the Lasso. This comes from the fact that the Lasso penalizes small and large coefficients in a similar way while EMVS is more adaptive. As a matter of fact EMVS seems better suited regarding the unbiasedness property advocated by Fan and Li (2001).

III.3 Main conclusions regarding the application

As it is common with many selectors, the choice of the regularization parameters matters a lot as we have seen in the crime data application. In the EMVS, the regularization parameter is embodied in equation (18). The choice of v_0 and v_1 appears to be crucial, just as in the choice of λ in the Lasso. However, the strenght of the Bayesian framework here is that the probability that the true model is equal to an estimated model can be computed by evaluating $\pi(\hat{\gamma}|y)$. This quantity gives a natural way of selecting a model which is not possible in the frequentist version of the Lasso. To us, it is definitely a strenght of the EMVS compared to its competitors

A potential extension of the model could be to build a not-so-informative prior on these two parameters so as to make the choice of the regularization parameters easier. For example, the Bayesian hierarchical specification in Giannone *et al.* (2012) put a prior distribution on few regularization parameters. These prior distribution are based on previous results from the macroeconomic forecasting literature, but they

are flat enough so as to let the "data speak" and find a good specification.

We have also noticed that contrary to other selectors, the EMVS does not shrink too much the coefficients which is definitely a strength.

IV Conclusion

The EMVS is a nice variable selection tools that displays four main strengths. Firstly, the Bayesian framework allows for flexible priors to incorporate the prior knowledge that we have regarding the structure of the coefficients (whether they are likely to be independently or jointly significant). Moreover, it is computationally very fast which means that we can explore a lot of sub-models in a very short span of time. This is clearly an advantage against traditional hierarchical Bayesian methods that require long and slow MCMC computations. It comes at the cost that the full posterior distribution of the parameter of interest β is no longer computed, but it is something that we can overlook if we are only interested in variable selection. Then, from the application we have done on US crime data, we have seen that the EMVS gives clear-cut answers in terms of variable selection, without biasing too much the estimate of the parameter of interest β. Finally, it gives an answer in terms of model selection thanks to the Bayesian framework which allows to compute $\pi(\hat{\gamma}|y)$, the posterior probability of a given selected model. This is a clear advantage over frequentist methods such as the Lasso or the Ridge that suffer a lot from the unanswered question of the penalty level choice.

A potential extension would be to extend the model to analyse time series. In this context, a prior often used in auto-regressive models is that old lags of a series are a priori less relevant that recent lags to predict the series itself (see for example Giannone *et al.* (2012)). In an application (the prediction of Euro-area inflation) which is not reported here, we have found that the EMVS was not very effective in selecting a convincing subset of lags/variables. Indeed, it ended up selecting a model where only one variable was relevant and showed unstable results.

V Bibliography

Belloni, A. and Chernozhukov, V. (2010): "Post-l1-penalized estimators in high-dimensional linear regression models". CeMMAP working papers CWP13/10, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Belloni, A., Chernozhukov, V., Fernández-Val, I., and Hansen, C. (2013): "Program evaluation with high-dimensional data". CeMMAP working papers CWP57/13, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Belloni, A., Chernozhukov, V., and Hansen, C. (2012): "Inference on treatment effects after selection amongst high-dimensional controls". CeMMAP working papers CWP10/12, Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009): "Simultaneous analysis of Lasso and Dantzig selector". *The Annals of Statistics*, 37(4):1705–1732.

De Mol, C., Giannone, D., and Reichlin, L. (2008): "Forecasting using a large number of predictors: Is Bayesian shrinkage a valid alternative to principal components?" *Journal of Econometrics*, 146(2):318–328.

Fan, J. and Li, R. (2001): "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties". 96:1348–1360.

Farrell, M. H. (2013): "Robust Inference on Average Treatment Effects with Possibly More Covariates than Observations". *ArXiv e-prints*.

George, E. I. and McCulloch, R. E. (1993): "Variable Selection via Gibbs Sampling". *Journal of the American Statistical Association*, 88(423):881–889.

——— (1997): "Approaches for Bayesian variable selection". *Statistica sinica*, 7(2):339–373.

Giannone, D., Lenza, M., Momferatu, D., and Onorante, L. (2010): "Short-term inflation projections: a Bayesian vector autoregressive approach". Working Papers ECARES 11, ULB Universite Libre de Bruxelles.

Giannone, D., Lenza, M., and Primiceri, G. E. (2012): "Prior Selection for Vector Autoregressions". Working Paper 18467, National Bureau of Economic Research.

Givord, P. (2010): "Econometric Methods for Public Policies Evaluation". Technical report.

Hastie, T., Tibshirani, R., and Friedman, J. (2009): *The Elements of Statistical Learning: Data mining, Inference and Prediction*. Springer, 2 edition.

Imbens, G. M. and Wooldridge, J. M. (2008): "Recent Developments in the Econometrics of Program Evaluation". NBER Working Papers 14251, National Bureau of Economic Research, Inc.

Litterman, R. B. (1986): "Forecasting with Bayesian Vector Autoregressions-Five Years of Experience". *Journal of Business & Economic Statistics*, 4(1):25–38.

Lounici, K. (2008): "Sup-norm convergence rate and sign concentration property of Lasso and Dantzig estimators". *Electronic Journal of Statistics*, 2:90–102.

Sala-i Martin, X. (1997): "I Just Ran Two Million Regressions". *American Economic Review, American Economic Association*, 87(2):178–83.

Meinshausen, N. and Yu, B. (2009): "Lasso-type recovery of sparse representations for high-dimensional data". *The Annals of Statistics*, 37(1):246–270.

Ročková, V. (2013): "Bayesian Variable Selection in High-dimensional Applications". Ph.D. thesis, Erasmus Universiteit Rotterdam.

Ročková, V. and George, E. I. (2014): "EMVS: The EM Approach to Bayesian Variable Selection". *Journal of the American Statistical Association*, 109(506):828–846.

Schwarz, G. (1978): "Estimating the Dimension of a Model". *The Annals of Statistics*, 6(2):461–464.

Tibshirani, R. (1996): "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society (Series B)*, 58:267–288.

Zhang, C.-H. and Huang, J. (2008): "The sparsity and bias of the Lasso selection in high-dimensional linear regression". *The Annals of Statistics*, 36(4):1567–1594.

Zhao, P. and Yu, B. (2006): "On Model Selection Consistency of Lasso". *J. Mach. Learn. Res.*, 7:2541–2563.