ECOM168: Bayesian variable selection

Daniele Bianchi

School of Economics and Finance,
Queen Mary University of London

Outline

1. Introduction

2. Spike-and-slab priors

3. Variable selection as post model-fitting

Introduction

Introduction

Variable selection in linear regression models takes many different forms.

At a very general level, variable selection corresponds to hypothesis testing of the form

$$\mathcal{H}_0 : \ \beta_j = 0, \qquad \text{vs} \qquad \mathcal{H}_1 : \beta_j \neq 0,$$

From a frequentist perspective:

$\hookrightarrow$ testing the null hypothesis $\mathcal{H}_0 : \ \beta_j = 0$ has the ultimate aim to find out how likely it would be for a regressor (or a set of regressors) to occur under the null hypothesis.

Introduction

Hierarchical shrinkage priors put significantly probability mass in both the null $\mathcal{H}_0$ and the alternative $\mathcal{H}_1$ hypothesis.

$\hookrightarrow$ e.g., $\beta \sim N\left(0, \tau^2 I_p\right)$ does put a certain probability mass at $\beta \neq 0$.

$\hookrightarrow$ parameters are not zero a priori.

Put it differently, posterior densities from hierarchical shrinkage priors are not identically zero whenever a model parameter is equal to its null value.

Introduction

Two possible (partial) solutions:

- ↪ Mixture priors (e.g., spike-and-slab priors).
- ↪ Post-processing methods (e.g., thresholding).
- ↪ Non-local priors (see Johnson and Rossell, 2010).

In the following we are going to focus on mixture priors and post-processing methods.

Spike-and-slab priors

Spike-and-slab priors

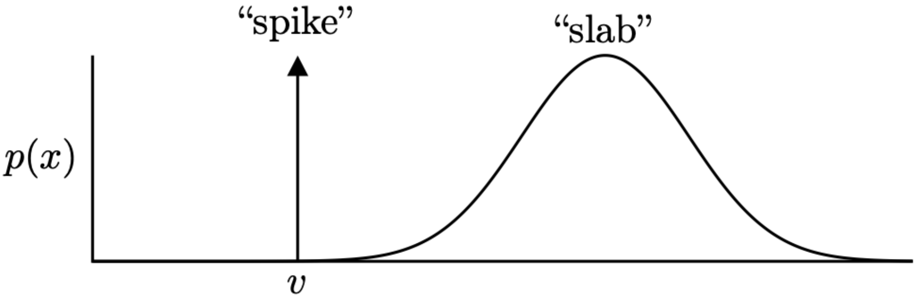
For a given random parameter x, a spike-and-slab prior attains a *mixture* of:

$\hookrightarrow$ some fixed value ν (called *spike*).

$\hookrightarrow$ some other prior $\pi\left(x\right)$ (called *slab*).

Contrary to hierarchical shrinkage priors, in the case ν implies $\beta = 0$, the spike-and-slab prior is sparsity inducing.

Spike-and-slab priors



An example of density of $p\left(x\right)$ for a spike-and-slab prior.

Spike-and-slab priors

In a standard regression context, the spike-and-slab prior pioneered by Mitchell and Beauchamp (1988) takes the form,

$$\beta_j|\gamma_j \sim (1-\gamma_j)\,\delta_0\left(\beta_j\right) + \gamma_j N\left(0,\tau^2\right), \qquad \gamma_j \sim \text{Ber}\left(\pi_0\right), \qquad j = 1,\ldots,p \tag{1}$$

where $\delta_0\left(\beta_j\right)$ is a *Dirac* function placing point mass at zero, i.e., the null hypothesis. The variable γ_j takes value 1 whether column j of X is included, and zero otherwise.

The mechanism is simple:

- ↪ $\gamma_j = 1$, the prior for β_j is $\beta_j \sim N\left(0,\tau^2\right)$, i.e., posterior estimate is unrestricted.
- ↪ $\gamma_j = 0$, the prior dominates the likelihood so that the posterior is concentrated at zero.

Spike-and-slab priors

The concept of variable selection is embedded in the posterior estimates of $\gamma_j, \; j = 1, \ldots, p$.

The posterior mean of the sequence of $\gamma_j = 0$ and $\gamma_j = 1$ denotes the *posterior inclusion probability* of each predictor as

$$\widehat{p}\,(\gamma_j|y) = \frac{1}{T}\sum_{t=1}^{T}\gamma_j^{(t)},$$

For e.g., if out of 10,000 draws we obtain $\gamma_j = 1$ for 2,000 times, then $\widehat{p}\,(\gamma_j|y) = 0.2 = 20\%$.

Barbieri and Berger (2004) suggest that the median probability model, i.e., model in which $\widehat{p}\,(\gamma_j|y) > 0.5$, is optimal for prediction.

$\hookrightarrow$ Truth is, there is some degree of subjectivity in the threshold for $\widehat{p}\,(\gamma_j|y)$.

Spike-and-slab priors

In the formulation of the spike-and-slab in Eq.(1) there are two key parameters:

- ↪ The prior on the variance τ^2.
- ↪ The prior on the selection parameter γ_j.

We are going to discuss in turn different approaches to set these parameters.

One can set a prior on τ^2, such as $\tau^2 \sim \text{Exp}\left(\lambda^2/2\right)$

- ↪ Caveat: careful not to set λ too large, since that would overshink the slab and make it indistinguishable from the spike.

Stochastic search variables selection

A computationally convenient approach is the stochastic search variable selection (SSVS) as originally proposed by George and McCulloch (1993, 1997),

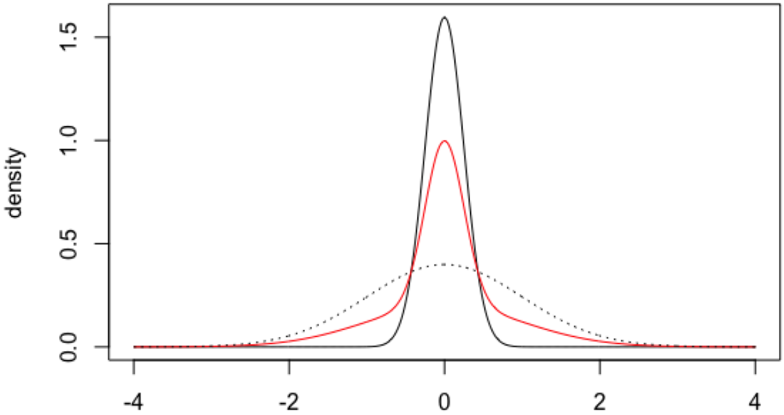
$$\beta_j|\gamma_j \sim (1-\gamma_j)\underbrace{N\left(0,\tau_0^2\right)}_{\text{spike}}+\gamma_j\underbrace{N\left(0,\tau_1^2\right)}_{\text{slab}}, \tag{2}$$

where both τ_0^2 and τ_1^2 are fixed, and $\tau_0^2 << \tau_1^2$.

This is a mixture of two continuous distributions, whereby for $\tau_0^2 \to 0$ the spike becomes a Dirac at zero.

N.B., for $\tau_0^2 \neq 0$, the spike is unable to shrink exactly $\beta_j = 0$, i.e., $\mathcal{H}_0 : \beta_j \approx 0$.

Stochastic search variables selection



An example of spike-and-slab prior as a mixture of Normals (see George and McCulloch, 1993)

Stochastic search variables selection

Clearly, the elicitation of τ_0^2, τ_1^2 is critical for variable selection given the prior in Eq.(2).

Narisetty and He (2014) show that fixing τ_0^2 and τ_1^2 to deterministic values may result in inconsistencies in variables selection.

Several alternatives have been proposed:

- $\hookrightarrow$ Ishwaran and Rao (2005): set $\tau_1^2 = c\tau_0^2$ and $\tau_0^2 \sim IG$, with $c >> 1$.
- $\hookrightarrow$ Frühwirth-Schnatter and Wagner (2010): a mixed strategy with $\tau_0^2 \sim \text{Exp}\left(\lambda^2\right)$ and $\tau_1^2 \sim IG$.
- $\hookrightarrow$ Ročková and George (2018): a mixed prior with $\tau_j^2 \sim \text{Exp}\left(\lambda_j^2/2\right)$ for $j = 0, 1$.

Stochastic search variables selection

Another important feature of Eq.(2) is the prior on γ_j. A relatively standard assumption is $\gamma_j \sim \text{Ber}\left(\pi_0\right)$.

However, this is a rather tight prior formulation.

$\hookrightarrow$ for e.g., $\pi_0 = 0.5$ implies that a priori we expect 50% of the predictors to be included.

$\hookrightarrow$ $\pi_0 = 0.1$ could be more sensible for large-dimensional models (aka more sparse a priori).

N.B., one could construct a hierarchical specification with $\pi_0 \sim \text{Beta}\left(1, \alpha_0\right)$.

$\hookrightarrow$ for instance, the choice of $\alpha_0 = 1$ makes this prior uniform.

Stochastic search variables selection

Computation with the spike-and-slab priors is often straightforward as is the case for other hierarchical priors.

Conditional on γ_j, the prior for β_j is either:

$\hookrightarrow$ a point mass at zero or a Normal (see, e.g., Mitchell and Beauchamp, 1988)

$\hookrightarrow$ one of the two Normal distributions (see, e.g., George and McCulloch, 1993).

$\hookrightarrow$ a mixture of non-Normal distributions (see, e.g., Frühwirth-Schnatter and Wagner, 2010).

Regarding the posterior computation of γ_js, this is often implemented element-by-element;

$\hookrightarrow$ $\gamma_j | \boldsymbol{\gamma}_{-j}$ (the set elements in $\boldsymbol{\gamma}$ with γ_j removed).

Stochastic search variables selection

Consider the SSVS prior,

$$\beta_j|\gamma_j, \sigma^2 \sim (1-\gamma_j)\, N\left(0, \sigma^2\tau_0^2\right) + \gamma_j N\left(0, \sigma^2\tau_1^2\right), \qquad \gamma_j \sim \text{Beta}\,(c, d)\,, \qquad \sigma^2 \sim IG\,(a, b)$$

and let D a diagonal matrix with elements $(1-\gamma_j)\,\tau_0^2 + \gamma_j\tau_1^2$. The conditional posteriors are:

$$\begin{aligned}
\beta|\ldots &\sim N\left(A^{-1}X'y, \sigma^2A^{-1}\right), \qquad \text{where} \qquad A^{-1} = \left(X'X + D\right)^{-1}, \\
\sigma^2|\ldots &\sim IG\left(a + \frac{n+p}{2}, b + \frac{s + \beta' D^{-1}\beta}{2}\right), \\
\gamma_j|\ldots &\sim \text{Ber}\left(\frac{N\left(\beta_j|0, \sigma^2\tau_1^2\right)\pi_0}{N\left(\beta_j|0, \sigma^2\tau_1^2\right)\pi_0 + N\left(\beta_j|0, \sigma^2\tau_0^2\right)(1-\pi_0)}\right), \qquad j = 1, \ldots, p \\
\pi_0|\ldots &\sim \text{Beta}\left(c + \sum_{j=1}^{p}\gamma_j, d + \sum_{j=1}^{p}(1-\gamma_j)\right), \qquad j = 1, \ldots, p
\end{aligned}$$

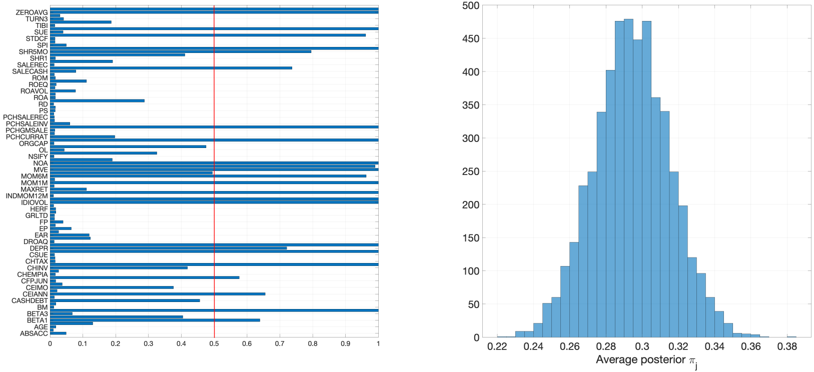
Stochastic search variables selection

The choice of τ_0^2, τ_1^2 is crucial for variables selection.

Narisetty et al. (2018) propose to fix the value of the prior variance parameters as $\tau_0^2 = \frac{\widehat{\sigma}^2}{10n}$ and $\tau_1^2 = \widehat{\sigma}^2 \max\left(\frac{p^2}{100n}, \log(n)\right)$ where $\widehat{\sigma}$ is the sample variance of y.

They also recommend to set a prior inclusion probability π_0 is chosen so that $Pr\left(\sum_{j=1}^{p} \gamma_j > K\right) = 0.1$ for $K = \max\left(10, \log(n)\right)$.

Example: Anomalies and the expected market returns



(a) Posterior estimates γ_js

(b) Cross-sectional distribution of π_js

An example of the posterior estimates of $\widehat{\gamma}_j$s and the cross-sectional average of the distribution of π_js.

Spike-and-slab lasso

Instead of fixing τ_j^2 for $j = 0, 1$, one could consider to elicit an hyper-prior for both.

Ročková and George (2018) propose to set two separate Laplace densities on the variance components,

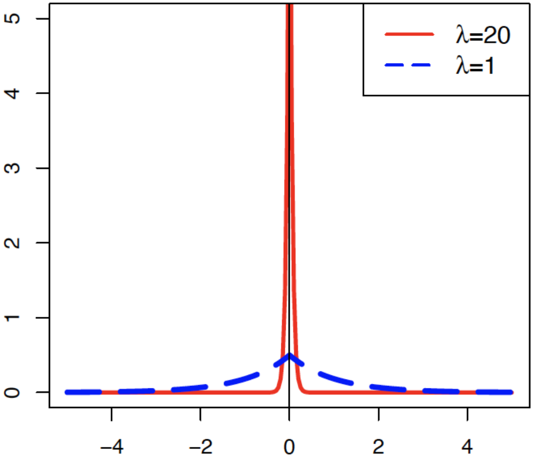
$$\tau_{j0}^2 | \lambda_0^2 \sim \text{Exp}\left(\frac{\lambda_0^2}{2}\right), \qquad \tau_{j1}^2 | \lambda_1^2 \sim \text{Exp}\left(\frac{\lambda_1^2}{2}\right), \qquad j = 1, \ldots, p$$

with $\lambda_0 >> \lambda_1$ so that $N\left(0, \sigma^2 \tau_{0j}^2\right)$ is the "spike" and $N\left(0, \sigma^2 \tau_{1j}^2\right)$ is the "slab".

The prior variances can be updated within a standard Gibbs sampler as

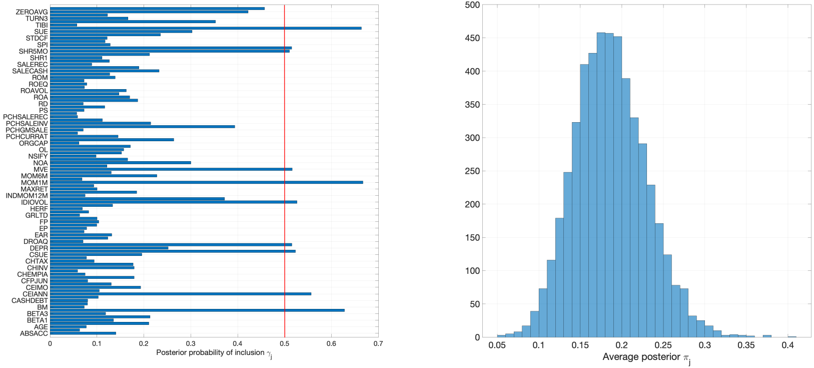
$$1/\tau_{ij}^2 | \ldots \sim IG\left(\sqrt{\lambda_i^2 \sigma^2 / \beta_j^2}, \lambda_i^2\right), \qquad \text{for} \qquad i = 1, 2,$$

Spike-and-slab lasso



Plot of the central region for the Laplace density with two different choices of scale parameters.

Example: Anomalies and the expected market returns



(c) Posterior estimates γ_js

(d) Cross-sectional distribution of π_js

An example of the posterior estimates of $\widehat{\gamma}_j$s and the cross-sectional average of the distribution of π_js from the spike-and-slab lasso.

Alternative formulations

The spike-and-slab prior can be written as

$$\beta_j|\gamma_j \sim N\left(0, \tau^2\gamma_j\right),$$

that is, the spike-and-slab belongs to the class of hierarchical priors.

In addition, if we introduce $\tau^2 \sim g$, then the spike-and-slab prior belongs to the class of "global-local" priors with:

- $\hookrightarrow$ τ^2 the global shrinkage.
- $\hookrightarrow$ γ_j the local shrinkage parameters, e.g., $\gamma_j \sim \text{Ber}\left(\pi_0\right)$.

Alternative formulations

Kuo and Mallick (1998) consider an alternative formulation for variables selection within the context of linear regressions,

$$y|\beta, \gamma, \sigma^2 \sim N\left(X\theta, \sigma^2 I_n\right), \qquad \text{where} \qquad \theta = (\beta_1\gamma_1, \ldots, \beta_p\gamma_p),$$

with $\gamma_j = 1$ if the jth variable is included in the model, and $\gamma_j = 0$ otherwise.

This formulation is equivalent to Mitchell and Beauchamp (1988), but it implies that the indicator γ_j enters via the likelihood and not through the prior for β_j.

- ↪ when $\gamma_j = 0$, then $\beta_j \sim p(\beta_j)$.
- ↪ not an issue, since we care about $\gamma_j\beta_j$, not β_j per se (when $\gamma_j = 0$, x_j is simply removed).

Several choices of priors: e.g., $\gamma_j \sim$ Ber(p_j), $\sigma^2 \sim IG(a, b)$, and $\beta \sim N(0, D)$.

Alternative formulations

The posterior densities take a familiar form and can be written as

$$\beta|\ldots \sim N\left(A^{-1}X^{*\prime}y/\sigma^2, A^{-1}\right), \qquad \text{with} \qquad A^{-1} = \left(X^{*\prime}X^{*}/\sigma^2 + D^{-1}\right)^{-1},$$

$$\sigma^2|\ldots \sim IG\left(a + \frac{n}{2}, b + \frac{1}{2}\left(y - X^{*\prime}\beta\right)'\left(y - X^{*\prime}\beta\right)\right)$$

$$\gamma_j|\ldots \sim \text{Ber}\left(\frac{c_j}{c_j + d_j}\right),$$

with

$$c_j = p_j \exp\left[-\frac{1}{2\sigma^2}\left(y - X\theta_j^*\right)'\left(y - X\theta_j^*\right)\right] \quad d_j = (1 - p_j)\exp\left[-\frac{1}{2\sigma^2}\left(y - X\theta_j^{**}\right)'\left(y - X\theta_j^{**}\right)\right],$$

where θ_j^* is θ with the jth component as β_j and θ_j^{**} is θ with the jth component as 0.

$\hookrightarrow$ The conditional posterior of γ_j depends on $\boldsymbol{\gamma}_{-j}$.

Improving sampling efficiency

In the context of the SSVS method, sampling from the posterior density requires a Cholesky decomposition and an inversion of a possibly large dimensional matrix D.

A consistent and scalable sampler for the SSVS – called *skinny Gibbs* – has been proposed by Narisetty et al. (2018). The starting point is a matrix formulation of the SSVS prior

$$\boldsymbol{\beta}|\boldsymbol{\gamma} \sim N\left(0, D_{\gamma}\right),$$

where $D_{\gamma} = \text{diag}\left(\left((1-\gamma_1\right)^2 \tau_0^2 + \gamma_1^2 \tau_1^2, \ldots, \left(1-\gamma_p\right)^2 \tau_0^2 + \gamma_p^2 \tau_1^2\right)$.

The conditional posterior takes the form

$$\boldsymbol{\beta}| \ldots \sim N\left(A^{-1} X^{\top} y / \sigma^2, A^{-1}\right),$$

where $A^{-1} = \left(X^{\top} X / \sigma^2 + D_{\gamma}^{-1}\right)^{-1}$ (see also, Korobilis et al., 2022).

Improving sampling efficiency

For a large p, inverting the matrix D_γ (and therefore A) could be prohibitive.

The skinny Gibbs algorithm of Narisetty et al. (2018) addresses this issue by splitting the posterior density of β in two components,

$$\begin{aligned} \beta_1 | \ldots &\sim N\left(A_1^{-1} X_1' y / \sigma^2, A_1'\right), \\ \beta_2 | \ldots &\sim N\left(0, A_2'\right), \end{aligned}$$

where β_1 is the p_1-dimensional vector of betas for which $\gamma_i = 1$, and β_2 the $p - p_1$-dimensional vector of betas for which $\gamma_j = 0$.

Improving sampling efficiency

The two posterior variances take the form

$$A_1^{-1} = \left(X^\top X/\sigma^2 + \frac{1}{\tau_1^2} I_{p_1} \right)^{-1}, \qquad \text{and} \qquad A_2^{-1} = \left(n + \frac{1}{\tau_0^2} \right)^{-1} I_{p-p_1},$$

In a very sparse setting, we could expect A_1^{-1} to be of moderate size.

Caveat: β_1 and β_2 can be highly correlated, depending on how "dense" are the covariates X.

Variable selection as post model-fitting

Variable selection as post model-fitting

Continuous shrinkage priors, such as global-local priors, induce “approximate” sparsity in the regression coefficients.

↪ allow a subset (“one group”) of coefficients to heavily shrunk towards zero.

↪ alternative to the “two-group” discrete mixture priors (see, e.g., Mitchell and Beauchamp, 1988; George and McCulloch, 1993).

A subclass of these priors, such as the Dirichlet-Laplace and the horseshoe possess a series of attractive theoretical properties, such as minimax optimality and frequentist validity.

Another key advantage is the (relatively) low computational complexity;

↪ block updating from conditionally conjugate Gaussian distributions.

Variable selection as post model-fitting

However, *shrinking* rather than *selecting* is a defining feature of these hierarchical priors.

An immediate consequence is that the posterior draws of the regression parameters are non-sparse with probability one:

$\hookrightarrow$ non-zero probability on $\mathcal{H}_0 : \beta_j = 0$.

That is, despite their tractability, hierarchical shrinkage priors do not automatically lead to variable selection.

A potential solution is *thresholding*; for e.g., Carvalho et al. (2010) defined a local shrinkage factor $(0, 1)$ which is analogous to posterior inclusion probability.

$\hookrightarrow$ Caveat: choice of the threshold an issue in practice.

Variable selection as post model-fitting

Ray and Bhattacharya (2018) propose a simple, yet effective method to select variables based on post model-fitting.

The Signal Adaptive Variable Selector (SAVS) approach *post-processes* a point estimate such as the posterior mean, obtained from a given hierarchical shrinkage prior.

Two key benefits:

$\hookrightarrow$ The procedure is automatic, i.e., no tuning parameters.

$\hookrightarrow$ Can be applied after any hierarchical shrinkage priors.

Signal Adaptive Variable Selector (SAVS)

For a given real number a, let $sign(a) \in (-1, 1)$ denote its sign with $sign(a) = 1$ for $a \geq 0$ and -1 otherwise.

Also, let $a_+ = \max\{a, 0\}$ denote the positive part of a, and $\|x\|$ be the Euclidean norm for $x \in \mathbb{R}^d$.

With these ingredients, let,

$$\widehat{\beta}_j^* = sign\left(\widehat{\beta}_j\right) \|X_j\|^{-2} \left(|\widehat{\beta}_j| \cdot \|X_j\|^2 - \mu_j\right)_+, \qquad j = 1, \ldots, p \tag{3}$$

where X_j is the jth column of X and $\mu_j = 1/|\widehat{\beta}_j|^2$ for $j = 1, \ldots, p$.

We henceforth refer to $\widehat{\beta}_j^*$ as the Signal Adaptive Variable Selector (SAVS) for the estimate $\widehat{\beta}_j$.

Signal Adaptive Variable Selector (SAVS)

```
Algorithm 1: SAVS algorithm
input  : Posterior mean $\widehat{\beta}$ and design matrix $X$
for $j = 1$ to $p$ do
    $\mu_j = 1/|\widehat{\beta}_j|^2$
    if $\widehat{\beta}_j| \cdot \|X_j\|^2 \leq \mu_j$ then
        $\widehat{\beta}_j^* = 0$
    else
        $\widehat{\beta}_j^* = sign\left(\widehat{\beta}_j\right) \|X_j\|^{-2} \left(|\widehat{\beta}_j| \cdot \|X_j\|^2 - \mu_j\right)$
    end
end
output: A sparse estimate $\widehat{\beta}^*$
```

Signal Adaptive Variable Selector (SAVS)

The SAVS algorithm takes a non-sparse point estimate $\widehat{\beta}_j$ and the design matrix X as input. Then returns a sparse estimate $\widehat{\beta}_j^*$ which can be readily used for variable selection.

In their original paper Ray and Bhattacharya (2018), $\widehat{\beta}_j$ is obtained from an horseshoe prior,

$\hookrightarrow$ Any other shrinkage prior can be used.

N.B., the quality of the SAVS is directly linked to the accuracy of the posterior estimates.

Signal Adaptive Variable Selector (SAVS)

To provide some motivation for the SAVS, notice that Eq.(3) can be obtained by solving an optimization problem closely related to the adaptive lasso,

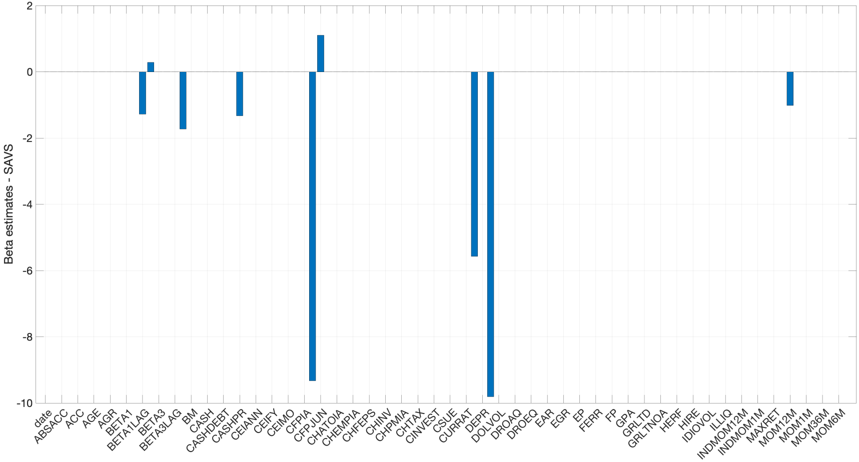
$$\widehat{\beta}^* = \arg\min_{\beta} \left\{ \frac{1}{2} \left\| X\widehat{\beta} - X\beta \right\|_2^2 + \sum_{j=1}^{p} \mu_j |\beta_j| \right\} \tag{4}$$

Equation (4) tries to find a sparse coefficient vector β that is close to $\widehat{\beta}$, while introducing a penalty in case of non-zero elements in β.

N.B., One shortcoming of SAVS is that there is no quantification of the uncertainty on $\widehat{\beta}^*$.

↪ Possible solution; replace $\widehat{\beta}_j$ in Eq.(3) with a draw from the posterior (see, e.g., Huber et al., 2021).

Example: Anomalies and the expected market returns



An example of the sparsified estimates from the SAVS. Posterior estimates are obtained from the horseshoe prior.

References

Barbieri, M. M. and Berger, J. O. (2004). Optimal predictive model selection. *The Annals of Statistics*.

Carvalho, C. M., Polson, N. G., and Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, 97(2):465–480.

Frühwirth-Schnatter, S. and Wagner, H. (2010). Stochastic model specification search for gaussian and partial non-gaussian state space models. *Journal of Econometrics*, 154(1):85–100.

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

George, E. I. and McCulloch, R. E. (1997). Approaches for bayesian variable selection. *Statistica Sinica*, pages 339–373.

Huber, F., Koop, G., and Onorante, L. (2021). Inducing sparsity and shrinkage in time-varying parameter models. *Journal of Business & Economic Statistics*, 39(3):669–683.

Ishwaran, H. and Rao, J. S. (2005). Spike and slab variable selection: frequentist and bayesian strategies. *The Annals of Statistics*.

Johnson, V. E. and Rossell, D. (2010). On the use of non-local prior densities in bayesian hypothesis tests. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(2):143–170.

Korobilis, D., Shimizu, K., et al. (2022). Bayesian approaches to shrinkage and sparse estimation. *Foundations and Trends® in Econometrics*, 11(4):230–354.

Kuo, L. and Mallick, B. (1998). Variable selection for regression models. *Sankhyā: The Indian Journal of Statistics, Series B*, pages 65–81.

Mitchell, T. J. and Beauchamp, J. J. (1988). Bayesian variable selection in linear regression. *Journal of the american statistical association*, 83(404):1023–1032.

Narisetty, N. N. and He, X. (2014). Bayesian variable selection with shrinking and diffusing priors. *The Annals of Statistics*.

Narisetty, N. N., Shen, J., and He, X. (2018). Skinny gibbs: A consistent and scalable gibbs sampler for model selection. *Journal of the American Statistical Association*.

Ray, P. and Bhattacharya, A. (2018). Signal adaptive variable selector for the horseshoe prior. *arXiv preprint arXiv:1810.09004*.

Ročková, V. and George, E. I. (2018). The spike-and-slab LASSO. *J. Amer. Statist. Assoc.*, 113(521):431–444.