Bayesian Variable Selection and Estimation for Group Lasso

Bayesian Statistics Project Report

Georges SARR georges.sarr@ensae.fr
Hamdi BEL HADJ HASSINE hamdi.belhadjhassine@ensae.fr
Lasme Ephrem ESSOH lasmeephremdominique.essoh@ensae.fr

April 27, 2022

1 Introduction

Regression is one of the ubiquitous tools in data analysis with applications in various domains. For some specific applications, the number of covariates in the regression can be higher than the number of covariates that actually explain the target variable, in which case lasso regression can be used to perform variable selection and eliminate a number of covariates depending on the regularization penalty. In some applications, covariates can be partitioned into groups that are semantically related, and in such case we can apply group lasso to select the groups of covariates that best explain the target variable. In this paper, the authors present different methods to perform variable selection both at the group level and individual covariate level, and propose a new method, Bayesian Sparse Group Selection with Spike and Slab Prior (BSGS-SS), which significantly reduces the number of false positives for a better selection of explanatory variables while still retaining good prediction accuracy.

2 Methods

In this section, we present a quick overview of the theoretical framework of the paper we study. Then, we present our methodology as well as our results.

2.1 Brief theoretical framework

The paper deals with feature selection in a sparse regression model. We consider a general linear regression problem with G factors :

$$\boldsymbol{Y}_{n\times 1} = \sum_{g=1}^{G} \boldsymbol{X}_g \boldsymbol{\beta}_g + \boldsymbol{\varepsilon}_{n\times 1}, \;\; \boldsymbol{\varepsilon}_{n\times 1} \sim \boldsymbol{N}_n(\boldsymbol{0}, \sigma^2 \boldsymbol{I}_n) \tag{1}$$

where $\boldsymbol{\beta}_g$ is a coefficients vector of length m_g, and $\boldsymbol{X}_g$ is an $n \times m_g$ covariate matrix corresponding to the factor $\boldsymbol{\beta}_g$, $g = 1, 2, ..., G$.

In the literature, an appropriate method to solve selection problem (1) when $m_g = 1$ is the LASSO which consists in minimizing the sum of least squares in the regression problem under the constraint that the regression coefficients are all in the unit ball in the sense of the L1 norm : this is the standard LASSO. This method also admits a Bayesian formulation. To deal with the general case i.e. when $m_g \geq 1$, we present the appropriate framework given by section 2.1.1 which also admits a Bayesian alternative presented in section 2.1.2.

2.1.1 Group LASSO (GL)

Group LASSO is a generalization of the standard LASSO method. This approach consists in solving

where G is the number of covariate groups and the $\boldsymbol{\beta}_g$s represent the vectors of grouped covariates. This model shrinks the coefficients at the group level but according to the paper and our experiments it doesn't perform variable selection well because of its high false-positive rate.

2.1.2 Bayesian Group LASSO with Spike and Slab Prior (BGL-SS)

The authors proposed this model to improve the Group LASSO (GL). It is a bayesian conception of GL with spike and slab prior which introduces sparsity at the group level.

Given the model $\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}_g, \sigma^2 \sim \boldsymbol{N}_n(\boldsymbol{X}\boldsymbol{\beta}_g, \sigma^2 \boldsymbol{I}_n)$, we assume that :

$$\begin{aligned}
\boldsymbol{\beta}_g|\sigma^2, \tau_g^2 &\overset{ind}{\sim} (1-\pi_0)\mathcal{N}(0, \sigma^2\tau_g^2 I_{m_g}) + \pi_0\delta_0(\boldsymbol{\beta}_g) \\
\tau_g^2 &\overset{ind}{\sim} \Gamma(\frac{m_g+1}{2}, \frac{\lambda^2}{2}) \\
\sigma^2 &\overset{ind}{\sim} I\Gamma(\alpha, \gamma)
\end{aligned}$$

The $\beta_{g,j}$s are estimated using their posterior median approximated by Gibbs sampling.

2.1.3 Sparse Group LASSO : frequentist (SGL) and Bayesian (BSGL, BSGS-SS) approaches

For some situations it may be desirable to select variables at both the group level and the individual level when the true coefficients present bi-level sparsity.

Sparse Group LASSO (SGL): A frequentist approach tackling this problem is the Sparse Group Lasso (SGL) method solving:

$$\min_{\boldsymbol{\beta}} ||\mathbf{Y} - \sum_{g=1}^{G} \mathbf{X}_g\boldsymbol{\beta}_g||_2^2 + \lambda_1||\boldsymbol{\beta}||_1 + \lambda_2\sum_{g=1}^{G}||\boldsymbol{\beta}_g||_2 \tag{2}$$

Bayesian Sparse Group LASSO (BSGL): The Bayesian counterpart of SGL is the Bayesian Sparse Group Lasso (BSGL) model with prior given by:

$$\pi(\beta) \propto \exp\big(-\lambda_1||\boldsymbol{\beta}||_1 - \lambda_2\sum_{g=1}^{G}||\boldsymbol{\beta}_g||_2\big) \tag{3}$$

for which the posterior mode estimates the $\beta_{g,j}$s

Bayesian Sparse Group Selection with Spike and Slab Prior (BSGS-SS): Although BSGL has shrinkage effects at both the group level and also within a group, it does not produce sparse model since the posterior mean/median estimators are never exact 0. The authors propose the Bayesian Sparse Group Selection with Spike and Slab prior (BSGS-SS) to achieve sparsity at both levels for variable selection purpose with a lower false-positive rate and more accurate predictions. Its formulation is given by:

$$
\begin{aligned}
&\boldsymbol{\beta}_g = \boldsymbol{V}_g^{\frac{1}{2}} \boldsymbol{b}_g, \text{ where } \boldsymbol{V}_g^{\frac{1}{2}} = \operatorname{diag}\left\{\tau_{g1}, \ldots, \tau_{gm_g}\right\}, \tau_{gj} \geq 0 \\
&\boldsymbol{b}_g \overset{ind}{\sim} (1-\pi_0)\, \boldsymbol{N}_{m_g}\left(\boldsymbol{0}, \boldsymbol{I}_{m_g}\right) + \pi_0 \delta_0\left(\boldsymbol{b}_g\right), \quad g = 1, \ldots, G \\
&\tau_{gj} \overset{ind}{\sim} (1-\pi_1)\, \boldsymbol{N}^{+}\left(0, s^2\right) + \pi_1 \delta_0\left(\tau_{gj}\right)
\end{aligned}
$$

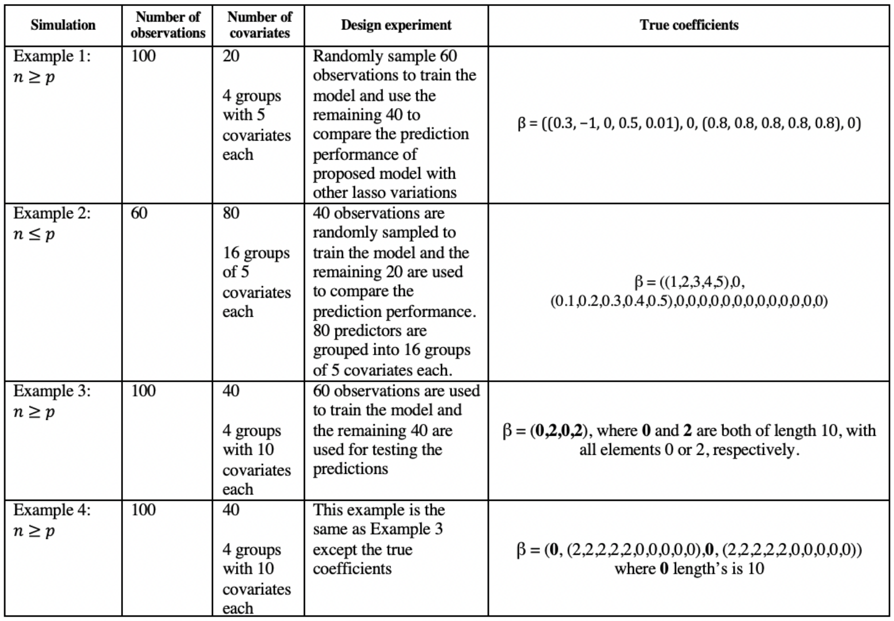
We obtain spike and slab posteriors from which we can sample using Gibbs sampler.

3 Experiments

3.1 Methodology

Our methodology focuses on the statistical simulation of the different models with the primary objective of reproducing the authors' results as far as possible with respect to Bayesian approaches and comparing them to frequentist results. Thus, we implement each of the models as well as the simulations starting from scratch, and to confirm that our implementation is correct we evaluated it using the same data examples used in the paper.

The table below describes the various simulation experiments we perfom. For each simulation, we estimate β coefficients using GL, SGL, BGL-SS and BSGS-SS and compare their performance.



Simulation	Number of observations	Number of covariates	Design experiment	True coefficients
Example 1: $n \geq p$	100	20 4 groups with 5 covariates each	Randomly sample 60 observations to train the model and use the remaining 40 to compare the prediction performance of proposed model with other lasso variations	β = ((0.3, −1, 0, 0.5, 0.01), 0, (0.8, 0.8, 0.8, 0.8, 0.8), 0)
Example 2: $n \leq p$	60	80 16 groups of 5 covariates each	40 observations are randomly sampled to train the model and the remaining 20 are used to compare the prediction performance. 80 predictors are grouped into 16 groups of 5 covariates each.	β = ((1,2,3,4,5),0, (0.1,0.2,0.3,0.4,0.5),0,0,0,0,0,0,0,0,0,0,0,0,0)
Example 3: $n \geq p$	100	40 4 groups with 10 covariates each	60 observations are used to train the model and the remaining 40 are used for testing the predictions	β = (**0**,**2**,**0**,**2**), where **0** and **2** are both of length 10, with all elements 0 or 2, respectively.
Example 4: $n \geq p$	100	40 4 groups with 10 covariates each	This example is the same as Example 3 except the true coefficients	β = (**0**, (2,2,2,2,2,0,0,0,0,0),**0**, (2,2,2,2,2,0,0,0,0,0)) where **0** length's is 10

Figure 1: Experiments Design

We evaluate the models using 3 metrics:

Selection True Positive Rate: TPR = $\frac{TP}{TP+FN}$

Selection False Positive Rate: FPR = $\frac{FP}{FP+TN}$

Prediction Mean Squared Error: MSE = $\frac{1}{N}\sum_{i=1}^{N}(Y_i - X_i\hat{\beta})^2)$

For selection evaluation, we denote by True Positives (TP) β coefficients that are correctly estimated to be non-zero, and by False Positives (FP) coefficients which are estimated to be non-zero while they are actually zero. Similarly TN and FN denote coefficients correctly or incorrectly estimated to be zero. We consider an estimated coefficient to be zero when its absolute value is below a given threshold (chosen to be 10^{-6}).

For Bayesian models, we obtain a relatively large number of samples from the posterior distribution of β. To choose which values to use for variable selection, we compared the sample mean, median, and the credible intervals. For the sample mean and median, we used them to calculate TPR and FPR. We also calculated credible intervals from the posterior distribution at different levels and counted the TP and FP based on whether the true coefficients are included in the credible intervals. The results of those experiments are provided in the next section.

4 Results

In table 1 we present the average TPR and FPR for each data example over 50 runs. We observe that BGL-SS and BSGS-SS greatly reduce the FPR and allow for parcimonious selection of variables while still maintaining a good TPR.

	GL	SGL	BGL-SS	BSGS-SS
Example 1				
TPR	0.90 (0.18)	0.90 (0.18)	0.78 (0.17)	0.48 (0.18)
FPR	0.51 (0.40)	0.42 (0.32)	0.28 (0.23)	0.07 (0.09)
Example 2				
TPR	1 (0)	0.99 (0.07)	0.70 (0.22)	0.76 (0.16)
FPR	0.42 (0.35)	0.15 (0.09)	0 (0)	0.01 (0.01)
Example 3				
TPR	1 (0)	1 (0)	1 (0)	1 (0)
FPR	0.55 (0.34)	0.59 (0.33)	0 (0)	0 (0)
Example 4				
TPR	1 (0)	1 (0)	1 (0)	1 (0)
FPR	0.56 (0.20)	0.5 (0.26)	0 (0)	0 (0.02)

Table 1: Average TPR and FPR (standard deviation in parentheses) of each model

In table 2 we present the median prediction MSE over 50 runs. As we can see, we could replicate the paper's results and achieve similar values. We observe here that althought they are more parcimonious, BGL-SS and BSGS-SS still provide competitive prediction performance. We also see that they perform better on data that presents more sparsity, as we can expect.

In table 3 we see that using the median β from the posterior samples greatly reduces the FPR while it only slightly deteriorates the TPR and the MSE, which can be a good tradeoff for false-positive

	GL	SGL	BGL-SS	BSGS-SS
Example 1	10.22 (2.49)	9.46 (3.23)	9.98 (2.21)	12.10 (2.56)
Example 2	6.55 (2.22)	6.04 (2.11)	6.78 (2.53)	6.14 (2.44)
Example 3	6.00 (1.31)	5.85 (1.26)	5.92 (1.43)	6.14 (1.61)
Example 4	4.87 (1.35)	5.20 (1.34)	4.90 (0.92)	4.95 (0.95)

Table 2: Median Mean Squared Error (standard deviation in parentheses) of each model over 50 runs

sensitive applications. Using the credible interval for variable selection accentuates this effect and allows for almost-zero false positives.

	TPR	FPR	MSE
BGL-SS(Mean)	1 (0)	053 (0.38)	6.82 (2.58)
BGL-SS(Median)	0.87 (0.19)	0.07 (0.17)	6.98 (2.70)
BGL-SS(CI 50%)	0.77 (0.25)	0.01 (0.05)	ø
BSGS-SS(Mean)	1.00 (0.00)	1.00 (0.00)	6.83 (2.57)
BSGS-SS(Median)	0.81 (0.25)	0.02 (0.05)	7.45 (3.38)
BSGS-SS(CI 50%)	0.70 (0.35)	0.01 (0.02)	ø

Table 3: TPR,FPR and MSE for each bayesian posterior criterion

5 Conclusion

Although implementing the studied models took a considerable amount of time and unfortunately prevented us from performing further interesting experiments with the models, thanks to this project we had the opportunity to study, implement and evaluate several algorithms based on the Bayesian theory that can be useful for real-world applications. Namely, the BSGS-SS model with median posterior sampling outperforms all the other studied algorithms in parsimonious variable selection and reducing the false-positive rate while still maintaining good prediction accuracy. This study also proves the usefulness of Spike-and-slab priors for bayesian regression, especially in sparse settings where models using Spike-and-slab priors such as BSGS-SS can perform variable selection both at the group level and intra-group.