Stability Feature Selection using Cluster Representative LASSO

Niharika Gauraha
Systems Science and Informatics Unit
Indian Statistical Institute, 8th Mile, Mysore Road RVCE Post Bangalore, Bangalore, India

Keywords: Lasso, Stability Selection, Cluster Representative Lasso, Cluster Group Lasso.

Abstract: Variable selection in high dimensional regression problems with strongly correlated variables or with near linear dependence among few variables remains one of the most important issues. We propose to cluster the variables first and then do stability feature selection using Lasso for cluster representatives. The first step involves generation of groups based on some criterion and the second step mainly performs group selection with controlling the number of false positives. Thus, our primary emphasis is on controlling type-I error for group variable selection in high-dimensional regression setting. We illustrate the method using simulated and pseudo-real data, and we show that the proposed method finds an optimal and consistent solution.

1 INTRODUCTION

We consider the usual linear regression model

$$\mathbf{Y} = \mathbf{X}\beta + \varepsilon, \qquad (1)$$

where $Y_{n\times 1}$ is a univariate response vector, $X_{n\times p}$ is the design matrix, $\beta_{p\times 1}$ is the true underlying coefficient vector and $\varepsilon_{n\times 1}$ is an error vector. when the number of variables (p) is much larger than the number of observations (n), $p >> n$, the ordinary least squares estimator is not unique and mostly overfits the data. The parameter vector β can only be estimated based on given very few observations, if β is sparse. The Lasso (Tibshirani, 1996) and other regularized regression methods are mostly used for sparse estimation and variable selection. However, variable selection in situations involving high empirical correlation or near linear dependence among few variables remains one of the most important issues. This problem is encountered in many applications such as in microarray analysis, a group of genes sharing the same biological pathway tend to have highly correlated expression levels (Segal et al., 2003) and it is often desirable to identify all(rather than a few) of them if they are related to the underlying biological process.

Various algorithms have been proposed which are based on the concept of clustering variables first and then pursuing variable selection. In this paper, we propose the Stability feature selection using CRL (SCRL), an approach for first identifying clusters among the variables using some criterion(discussed in section 2.2) and then subsequently performing stability feature selection on cluster representatives while controlling the number of false positives. The stability feature selection consists of repeatedly applying the baseline feature selection method to random data sub-samples of half-size, and finally selecting the features that have larger selection frequency than a predefined threshold value. Thus, The proposed algorithm, SCLR, is an application of stability feature selection where the base selection procedure is the Lasso and the Lasso is applied on the reduced design matrix of cluster representatives. Since, the Lasso is repeatedly applied on the reduced design matrix, the SCRL method is computationally fast as well.

Basically, The proposed method, SCRL is a two-stage procedure: at the first stage we cluster the variables and at the second stage we do group selection by stability feature selection using Lasso for cluster-representatives. When the group sizes are all one, it reduces to stability selection. In order to illustrate the performance of SCRL we carry out a simple simulation. We consider a fixed design matrix $X_{n\times p}$ generated as

$$\begin{aligned}
&x_i = Z_1 + \varepsilon_i^x, \quad Z_1 \sim N(0,1), \quad i = 1,...,5\\
&x_i = Z_2 + \varepsilon_i^x, \quad Z_2 \sim N(0,1), \quad i = 6,...,10\\
&x_i = Z_3 + \varepsilon_i^x, \quad Z_3 \sim N(0,1), \quad i = 11,...,15\\
&x_i \; i.i.d. \sim N(0,1), \quad i = 16,...,20\\
&\varepsilon_i^x \; i.i.d. \sim N(0,0.01), i = 1,...,15
\end{aligned}$$

In this example, the predictors are divided into three equally important groups and within each group there

Gauraha, N.
Stability Feature Selection using Cluster Representative LASSO.
DOI: 10.5220/0005827003810386
In *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2016)*, pages 381-386
ISBN: 978-989-758-173-1

are five members. The three groups have pairwise empirical correlations $\rho \approx 0.9$ and the remaining five are noise features. The true active set is $S_0 = \{1,2,...,15\}$. we generate the response according to $y = X\beta + \varepsilon$, where the elements of ε are i.i.d. draws from a $N(0,\sigma^2)$ distribution. We simulated data with sample size n = 100 and with predictors p = 20 and $\sigma = 3$ and 15.

As variable selection or screening method we use the following five methods and compare the true positives rate(TPR) and the number of false positives(FP): The Lasso (Tibshirani, 1996), stability selection using Lasso (Meinshausen and Bühlmann, 2010), CLR (Bühlmann et al., 2012), CGL (Bühlmann et al., 2012) and the proposed method SCRL. For the Lasso, CRL and CGL, we run 50 simulations and we choose the model with the smallest prediction error among 50 runs. The results are reported in table 1.

Table 1: Comparision of TPR and FP of different methods.

σ	Method	TPR	FP
3	Lasso	0.6	4
	Stability Selection	0	0
	CGL	1	4
	CRL	1	4
	SCRL	1	0
15	Lasso	0.4	2
	Stability Selection	0	0
	CGL	1	1
	CRL	1	1
	SCRL	1	0

An ideal variable selection method would select only 15 true predictors and no noise features. The Lasso tends to select single variable from the group of correlated or linearly dependent variables. In the case of Stability Selection none is selected. CGL and CRL select all true predictors but also select some noise features. The SCRL selects all true variables and no noise features. Thus, SCRL gives an optimal solution.

The rest of this paper is organized as follows. In Section 2, we provide background, review of relevant work and we discuss our contribution. In section 3, we describe the proposed algorithm which mostly selects more adequate models in terms of model interpretation with reduced type I error(false positives). In section 4, we provide simulation studies. Section 5 contains the computational details and we shall provide conclusion in section 6.

2 BACKGROUND AND NOTATIONS

In this section, we state notations and define required concepts. We also provide review of relevant work and our contribution.

2.1 Notations and Assumptions

We mostly follow the notations in (Bühlmann and van de Geer, 2011). We consider the usual linear regression setup with univariate response variable $Y \in R$ and p-dimensional variables $X \in R^p$:

$$Y_i = \sum_{j=1}^{p} X_i^{(j)} \beta_j + \varepsilon_i \quad i = 1,...,n \;\; j = 1,...,p \qquad (2)$$

where $\varepsilon_i \sim N(0,\sigma^2)$
or, in matrix notation (as in Equation 1)

$$\mathbf{y} = \mathbf{X}\beta + \varepsilon$$

where $\beta \in R^p$ are unknown coefficients to be estimated, and the components of the noise vector $\varepsilon \in R^n$ are i.i.d. $N(0,\sigma^2)$

$L1$-norm is defined as:

$$\|\beta\|_1 = \sum_{j=1}^{p} |\beta_j| \qquad (3)$$

$L2$-norm squared is defined as:

$$\|\beta\|_2^2 = \sum_{j=1}^{p} \beta_j^2 \qquad (4)$$

The infinite norm is defined as:

$$\|\beta\|_\infty = max_{1 \le i \le N} |\beta_j| \qquad (5)$$

The true active set is denoted as S_0 and defined as

$$S_0 = \{j; \beta_j \neq 0\} \qquad (6)$$

The estimated parameter vector is denoted as $\hat{\beta}$. The estimated active set is denoted as $\hat{S}$ and defined as

$$\hat{S} = \{j; \hat{\beta}_j \neq 0\} \qquad (7)$$

We consider true positive rate as a measure of performance, which is defined as:

$$TPR = \frac{|\hat{S} \bigcap S_0|}{|S_0|} \qquad (8)$$

The number of clusters are denoted by q. The partition, $G = \{G_1,...,G_q\}$ with $\cup_{r=1}^{q} G_r = \{1,...,p\}$ and $G_r \cap G_l = \emptyset$, represents group structure among variables. The clusters $G_1,...,G_q$ are generated from the

design matrix X, using methods as described in section 2.2. The representative for each cluster is then defined as (Bühlmann et al., 2012)

$$\bar{X}^{(r)} = \frac{1}{|G_r|} \sum_{j \in G_r} X^{(j)}, \quad r = 1, ..., q,$$

where $X^{(j)}$ denotes the jth $n \times 1$ column-vector of X. The design matrix of cluster representatives is denoted as $\bar{X}$.

2.2 Clustering of Variables

To cluster variables we use two methods: correlation based and canonical correlation based bottom-up agglomerative hierarchical clustering methods. The first method forms groups of variables based on correlations between them. The second method uses canonical correlation for clustering variables. The construction of groups based on canonical correlations addresses the problem of linear dependence among variables, whereas the standard correlation based hierarchical clustering addresses only correlation problems. For further details on clustering of variables and determining the number of clusters, we refer to (Bühlmann et al., 2012).

2.3 The Lasso and the Group Lasso

The Least Absolute Shrinkage and Selection Operator (Lasso), introduced by Tibshirani (Tibshirani, 1996), is a penalized least squares method that imposes an L1-penalty on the regression coefficients. The lasso does both shrinkage and automatic variable selection simultaneously due to the nature of the L1-penalty.

The lasso estimator is defined as

$$\hat{\beta}_{Lasso} = \underset{\beta}{\operatorname{argmin}}(\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1) \qquad (9)$$

But the Lasso-estimator has some limitations, i.e., if some variables are highly correlated with each other, the lasso tends to select a single variable out of a group of correlated variables. In certain situations, when the distinct groups or clusters among the variables are known a priory and it is desirable to select or drop the whole group instead of single variables. The group Lasso (Yuan and Lin, 2007) is used, that imposes an L2-penalty on the coefficients within each of q known groups to achieve such group sparsity.

The Group Lasso estimator(with known q groups) is defined as

$$\hat{\beta}_{GL} = \underset{\beta}{\operatorname{argmin}}(\|\mathbf{y} - \sum_{j=1}^{j=K} \mathbf{X}_j\beta_j\|_2^2 + \lambda \sum_{j=1}^{j=q} m_j\|\beta_j\|_2) \qquad (10)$$

where the $m_j = \sqrt{|G_j|}$ serves as balancing term for varying group sizes. The group Lasso behaves like the lasso at the group level, depending on the value of the regularization parameter λ, the whole group of variables may drop out of the model. For singleton groups (the group sizes are all one), it reduces exactly to the lasso.

2.4 Cluster Group Lasso

The cluster group Lasso, first identifies groups among the variables using hierarchical clustering methods described in section 2.2, and then applies the group lasso(Equation 10) to the resulting clusters. For more details on CGL, we refer to (Bühlmann et al., 2012).

2.5 Cluster Representative Lasso

Similar to the CGL, the cluster representative Lasso, first identifies groups among the variables using hierarchical clustering and then applies the lasso for cluster representatives (Bühlmann et al., 2012).

The optimization problem for CRL is defined using response y and the design matrix of cluster representatives $\bar{X}$ as:

$$\hat{\beta}_{CRL} = \underset{\beta}{\operatorname{argmin}}(\|\mathbf{y} - \bar{\mathbf{X}}\beta\|_2^2 + \lambda_{CRL}\|\beta\|_1) \qquad (11)$$

The selected clusters are then denoted as:

$$\hat{S}_{clust,CRL} = \{r; \hat{\beta}_{CRL,r} \neq 0, r = 1, ..., q\}$$

and the selected variables are obtained as the union of the selected clusters as:

$$\hat{S}_{CRL} = \cup_{r \in \hat{S}_{clust,CRL}} G_r$$

2.6 Stability Feature Selection

The stability feature selection, introduced by N. Meinhausen and P. Buhlmann (Meinshausen and Bühlmann, 2010), is a general technique for performing feature selection while controlling the type-I error. It is combination of sub-sampling and high-dimensional feature selection algorithms, i.e., the Lasso. It provides a framework for the baseline feature selection method, to identify a set of stable variables that are selected with high probability. Mainly, it consists of repeatedly applying the baseline feature selection method to random data sub-samples of half-size, and finally selecting the variables which have selection frequency larger than a fixed threshold value (usually in the range (0.6, 0.9)). For further details see (Meinshausen and Bühlmann, 2010).

2.7 Review of Relevant Work and Our Contribution

This section provides a review of relevant work in order to show that how our proposal differs or extend the previous studies.

It is known that the Lasso can not handle the situations where predictors are highly correlated or group of predictors are linearly dependent. In order to deal with such situations, various algorithms have been proposed which are based on the concept of clustering variables first and then pursuing variable selection, or clustering variables and model selection simultaneously.

The methods that perform clustering and model fitting simultaneously are The Elastic Net (Zou and Hastie, 2005), Fused LASSO (Tibshirani et al., 2005), OSCAR(octagonal shrinkage and clustering algorithm for regression) (H. and B., 2008) and Mnet (Huang et al., 2010). ENet uses a combination of the L_1 and L_2 penalties, OSCAR uses a combination of L_1 norm and and L_∞ norm and Mnet uses a combination of the MCP(minimum concave penalty) and L_2 penalties. As these methods are based on combination of penalties, they do not use any specific information on the correlation pattern among the predictors, Hence they can not handle linear dependency problem. Moreover Fused Lasso is applicable only when the variables have a natural ordering and not suitable to perform automated variable clustering to unordered features.

We list few methods that perform clustering and model fitting at different stages: Principal component regression (M., 1957) , Tree Harvesting (Hastie et al., 2001), Cluster Group Lasso (Bühlmann et al., 2012), Cluster representative Lasso(CRL) (Bühlmann et al., 2012) and the sparse Laplacian shrinkage (SLS) (J. et al., 2011)). All these methods have been proven to be consistent variable selection techniques but they fail to control the false positive rate.

Since the Lasso tends to select only one variable from the group of strongly correlated variables(even if many or all of these variables are important), the stability feature selection using Lasso does not choose any variable from the group of highly correlated variables because correlated variables split the vote. To overcome this problem we propose to cluster the variables first and then do stability feature selection using Lasso for cluster-representatives. Basically, our work can be viewed as an extension of CRL (Bühlmann et al., 2012) and an application of stability feature selection (Meinshausen and Bühlmann, 2010). We compare our algorithm with the CRL, in terms of variable selection in section 4. Our simulation studies shows that our method outperforms the CRL.

3 STABILITY FEATURE SELECTION USING CRL

We consider high dimensional setting, where group of variables are strongly correlated or there exist near linearly dependency among few variables. It is known that the Lasso tends to select only one variable from the group of highly correlated or linearly dependent variables even though many or all of them belong to the active set S_0. Various techniques based on clustering in combination with sparse estimation have been proposed in past for variable selection, or in more mathematical terms, to infer the true active set S_0, but mostly they fail to control the selection of false positives. In this article, Our aim is to identify the true active set and to control false positives simultaneously. We use the concept of clustering the correlated or linearly dependent variables and then selecting or dropping the whole group instead of single variables similar to the CRG method proposed in (Bühlmann et al., 2012). The stability feature selection has been proven for identifying the most stable features and for providing control on the family-wise error rate, we recommend (Meinshausen and Bühlmann, 2010) for theoretical proofs. In order to reduce the selection of false positive groups, we propose to combine the CRL with sub-sampling, we call it SCRL, stability feature selection using CRL.

The proposed SCLR method can be seen as an application of stability feature selection where the base selection procedure is the Lasso and the Lasso is applied on the reduced design matrix of cluster representatives. The advantage of using reduced design matrix of cluster representatives are as follows:

(a) The plain stability feature selection, where the baseline feature is the Lasso and the Lasso is applied on the whole design matrix X(there is no pre-processing step of clustering the variables). It is a special case of SCRL when the group sizes are all one. The plain stability feature selection is not suitable for variable selection when the variables are highly correlated because the underlying selection method Lasso tends to select single variables per cluster, it selects none from the correlated groups as the vote gets split within the cluster variables. Therefore, using the reduced design matrix ensures the most stable group selection.
(b)The Lasso is repeatedly applied on the reduced design matrix, therefore the SCRL method is computationally fast as well.

Algorithm 1: SCRL Algorithm.

Input: dataset (y, X), ClusteringMethod
Output: $\hat{S}$:= set of selected variables
Steps:
stage 1:
1. Perform clustering, Denote clusters as $G_1, ..., G_q$
2. Compute the matrix of cluster representatives, denote it as $\bar{X}$
stage 2:
3. perform stability feature selection using response Y and the reduced design matrix $\bar{X}$ Denote the selected set of groups as $\hat{S}_G = \{r;$ cluster $G_r$is selected, $r = 1, ..., q\}$.
4. The union of the selected groups is the selected set of variables $\hat{S} = \cup r \in \hat{S}_{cluster}$
return $\hat{S}$

Similar to the CRL method, The proposed method SCRL(Algorithm 1) is a two-stage procedure: the first stage is exactly the same as the CRL, where we cluster the variables based on the criterion discussed previously and compute the design matrix of cluster representatives. At the second stage we perform group selection by stability feature selection using Lasso for cluster-representatives. When the group sizes are all one, it reduces to the plain stability selection.

4 NUMERICAL RESULTS

In this section, we consider two simulation settings and a semi-real data example. We compare the performances of CRL and SCRL. In each example, data is simulated from the linear model (Equation 1) with fixed design **X**. These examples are similar to the examples used in the paper (Bühlmann et al., 2012).

We consider the true positive rate(and also the number of false positives) as a measure of performance, which is defined in Equation 8.

4.1 Example 1: Block Diagonal Model

We generate covariates from $N_p(0, \Sigma_1)$, where Σ_1 consists of 10 block matrices $\mathcal{T}$, where $\mathcal{T}_{10\times 10}$ is a block diagonal matrix, defined as

$$\mathcal{T}_{j,k} = \begin{cases} 1, & j = k \\ .9, & else \end{cases}$$

The true active set and true parameters β are defined as: $S_0 = \{1, 2, ..., 20\}$ and for any $j \in S_0$ we sample β_j from the set $\{.1, .2, .3, ..., 2\}$ without replacement. This setup has all the active variables in the first two blocks of highly correlated variables.

Simulation results are reported in table 2. We notice that the TPR is the same for both the methods, but SCRL has lower false positives than CRL.

Table 2: Performance measures for example 1.

σ	Method	TPR	FP
3	SCRL	1	0
	CRL	1	40
15	SCRL	1	0
	CRL	1	60

4.2 Example 2: Single Block Design

We generate covariates from $N_p(0, \Sigma_2)$, where Σ_2 consisted of a single group of strongly correlated variables of size 30, it is defined as

$$\Sigma_{2;j,k} = \begin{cases} 1, & j = k \\ 0.9 & i, j \in \{1, ..., 30\} \text{ and } i \neq j, \\ 0 & otherwise \end{cases}$$

The remaining 70 variables are uncorrelated. The true active set and true parameters beta are defined as: $S_0 = \{1, 2, ..., 15\} \cup \{31, 32, ..., 35\}$ and for any $j \in S_0$ we sample β_j from the set $\{.1, .2, .3, ..., 2\}$ without replacement. The first block of size 30 contains 25, the most of the active predictors.

Simulation results are reported in table 3. The TPR is the same for both the methods, but SCRL has lower false positives than CRL.

Table 3: Performance measures for example 2.

σ	Method	TPR	FP
3	SCRL	0.9	5
	CRL	0.9	24
15	SCRL	0.9	5
	CRL	0.9	33

4.3 Example 3: Pseudo-real Data

We consider a real dataset, riboflavin($n = 71, p = 4088$) data for the design matrix X with synthetic parameters beta and simulated Gaussian errors $N_n(0, \sigma^2 I)$. See (Bühlmann et al., 2014) for details on riboflavin dataset. We fix the size of the active set to $s_0 = 10$. For the true active set, we randomly select a variable k, and the nine variables which have highest correlation to the variable k, and for each $j \in S_0$ we set $\beta_j = 1$.

The performance measures are reported in table 4. The TPR is the same for both the methods, but SCRL has lower false positives than CRL.

Table 4: Performance measures for semi-real dataset.

σ	Method	TPR	FP
3	SCRL	1	0
	CRL	1	5
15	SCRL	0.7	3
	CRL	0.7	7

4.4 Empirical Results

We clearly see that in both of the simulation settings and in the pseudo-real example, the SCRL method outperform the CRL, since the number of false positives selected by CRL is much larger than the SCRL.

5 COMPUTATIONAL DETAILS

Statistical analysis was performed in R 3.2.2. We used, the packages "glmnet" for penalized regression methods(the Lasso) , the package "gglasso" to perform group Lasso, the package "ClustOfVar" for clustering of variables and the package 'hdi" for stability selection. All mentioned packages are available from the Comprehensive R Archive Network (CRAN) at http://cran.r-project.org/.

6 CONCLUSIONS

In this article, we proposed a two stage procedure, SCRL, for variable selection with controlled false positives in high-dimensional regression model with strongly correlated variables. At the first stage, SCRL identifies the clusters or group structures using some criterion and clusters representatives are computed for each cluster. At the second stage these cluster representatives are then used in order to more accurately perform stability feature selection while controlling the type-I error. Our algorithm is an application of stability feature selection with the baseline feature selection method used as the Lasso. Since the Lasso tends to select only one variable from the group of strongly correlated or linearly dependent variables, the stability feature selection using Lasso selects none of the variables from the correlated/linearly dependent group because the vote gets spit among the correlated variables. To address this issue we use the reduced design matrix of cluster representatives for stability feature selection. The stability feature selection has been proven for identifying the most stable features and for providing control on the family-wise error rate. Therefore, the SCRL reports most stable groups with controlled false positives. In addition, it also offers computational advantage, as the Lasso method only has to be applied on reduced design matrix.

REFERENCES

Bühlmann, P., Kalisch, M., and Meier, L. (2014). High-dimensional statistics with a view towards applications in biology. *Annual Review of Statistics and its Applications*, 1:255–278.

Bühlmann, P., Rütimann, P., van de Geer, S., and Zhang, C.-H. (2012). Correlated variables in regression: clustering and sparse estimation. *Journal of Statistical Planning and Inference*, 143:1835–1871.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Springer Verlag.

H., B. and B., R. (2008). Simultaneous regression shrinkage, variable selection and clustering of predictors with oscar. *Biometrics*, pages 115–123.

Hastie, T., Tibshirani, R., Botstein, D., and Brown, P. (2001). Supervised harvesting of expression trees. *Genome Biology*, 2:1–12.

Huang, J., Breheny, P., Ma, S., and hui Zhang, C. (2010). The mnet method for variable selection. *Department of Statistics and Actuarial Science, University of Iowa*.

J., H., S, M., H., L., and CH., Z. (2011). The sparse laplacian shrinkage estimator for high-dimensional regression. *statistical signal processing, in SSP09. IEEE/SP 15th Workshop on Statistical Signal Processing*, pages 2021–2046.

M., K. (1957). A course in multivariate analysis. *Griffin: London*.

Meinshausen, N. and Bühlmann, P. (2010). Stability selection (with discussion). *J. R. Statist. Soc*, 72:417–473.

Segal, M., Dahlquist, K., and Conklin, B. (2003). Regression approaches for microarray data analysis. *Journal of Computational Biology*, 10:961–980.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Statist. Soc*, 58:267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society Series B*, pages 91–108.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in regression with grouped variables. *J. R. Statist. Soc*, 68(1):49–67.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *J. R. Statist. Soc*, 67:301–320.