Proceedings of the KI 2017 Workshop on Formal and Cognitive Reasoning

Pre-Selection in Cluster Lasso Methods for Correlated Variable Selection in High-Dimensional Linear Models

Niharika Gauraha and Swapan Parui

Indian Statistical Institute

Abstract. We consider variable selection problems in high dimensional sparse regression models with strongly correlated variables. To handle correlated variables, the concept of clustering or grouping variables and then pursuing model fitting is widely accepted. When the dimension is very high, finding an appropriate group structure is as difficult as the original problem. We propose to use Elastic-net as a pre-selection step for Cluster Lasso methods (i.e. Cluster Group Lasso and Cluster Representative Lasso). The Elastic-net selects correlated relevant variables, but it fails to reveal the correlation structure among the active variables. We use cluster Lasso methods to address shortcoming of the Elastic-net, and the Elastic-net is used to provide reduced feature set for the cluster Lasso methods. We theoretically explore, the group selection consistency of the proposed combination of algorithms under various conditions, i.e. Irrepresentable Condition (IC), Elastic-net Irrepresentable Condition (EIC) and Group Irrepresentable Condition (GIC). We support the theory using simulated and real dataset examples.

Keywords: Correlated Variable Selection, Group Lasso, Cluster Lasso Methods, High-dimensional Linear models

1 Introduction

We consider the usual linear regression model

$$\mathbf{Y} = \mathbf{X}\beta^0 + \epsilon, \tag{1}$$

with response vector $\mathbf{Y}_{n\times 1}$, design matrix $\mathbf{X}_{n\times p}$, true underlying coefficient vector $\beta^0_{p\times 1}$ and error vector $\epsilon_{n\times 1}$. When the number of predictors (p) is much larger than the number of observations (n), $p \gg n$, the Lasso [11] and its variants are mostly used for sparse estimation and variable selection. However, variable selection in situations involving high empirical correlation remains one of the most important issues. This problem is encountered in many applications such as microarray data analysis and genome analysis, see [10].

It has been proven that the design matrix must satisfy the following conditions for the Lasso to perform exact variable selection: irrepresentable condition (IC)

[16] and beta-min condition [1]. Having highly correlated variables implies that the design matrix violates the IC. To deal with correlated variables, mainly two approaches have been proposed in literature: simultaneous clustering and model fitting (see [7]), and clustering followed by the sparse estimation (see [2]) and [5]). The former approach imposes restrictive conditions on the design matrix. However, the time complexity for clustering of variables severely limits the dimension of data sets that can be processed by the later approach. Moreover, group selection in models with a larger number of groups is more difficult (see [14]). To overcome the limitations of the later approach, we propose to use Elastic-net [17] as a pre-selection procedure for the Cluster Lasso [2] methods. Basically, we try to reduce the noise (dimension) first using the Elastic-net which is known to select correlated variables, before applying the Cluster Lasso (CL) methods. This scheme allows for a significant decrease in the cost of clustering, especially in the high dimensional problems.

Basically, we propose to combine Elastic-net and Cluster Lasso methods to improve both speed of computation and accuracy of the results in the case of sparsity. This goal is achieved by using Elastic-net to first reduce the number of variables under consideration and then using CL methods on the reduced data to select correlated variables. We theoretically explore, how the proposed combination of algorithms will perform under various conditions, i.e. Irrepresentable Condition (IC), Elastic-net Irrepresentable Condition (EIC) and Group Irrepresentable Condition (GIC). This theoretical analysis is validated by experiments on simulated datasets by comparison with different methods: Lasso, Elastic-net and Cluster Group Lasso. Moreover, we shows that the algorithm is able to improve the results compared to the mentioned methods on a real-world dataset.

The rest of this paper is organized as follows. In section 2, we provide notations, assumptions and background to be used later. In section 3, we review mathematical theory of the Lasso, the Elastic-net and the group Lasso. In section 4, we describe the proposed algorithm which mostly selects more adequate models in terms of model interpretation and prediction performance. In section 5, we provide numerical results based on simulated and real dataset. Section 6 contains the computational details and we shall provide conclusion in section 7.

2 Basic Assumptions, Notations and Concepts

In this section, we state notations and assumptions and we define the required concepts.

2.1 Notations and Assumptions

We consider the usual linear regression set up as given in Equation (1). We assume that the components of the noise vector $\epsilon \in R^n$ are i.i.d. $N(0, \sigma^2)$. The columns of the design matrix $\mathbf{X}$ are denoted by X^j. We assume that the design matrix $\mathbf{X}$ is fixed, the data is centred and the predictors are standardized, so that $\sum_{i=1}^n Y_i = 0$, $\sum_{i=1}^n X_i^j = 0$ and $\frac{1}{n} X^{j'} X^j = 1$ for all $j = 1, ..., p$. The ℓ_1-norm is

denoted by $\|.\|_1$, the ℓ_2-norm is denoted by $\|.\|_2$ and the ℓ_∞ norm is denoted by $\|.\|_\infty$. The minimum eigenvalue of a matrix A is denoted as $\Lambda_{min}(A)$. The true active set S_0 denotes the support of the subset selection solution ($S_0 = supp(\beta_0)$) and defined as $S_0 = \{j; \beta_j^0 \neq 0\}$. For any given $S \subset \{1, 2, ..., p\}$, the β_S is a $p \times 1$ vector which has zeroes outside the set S, as given by

$$\beta_S = \{\beta_j I(j \in S)\},$$

where I is the indicator function. Then we have

$$\beta = \beta_S + \beta_{S^c}.$$

The (scaled) Gram matrix (covariance matrix) is defined as

$$\hat{\Sigma} = \frac{\mathbf{X}'\mathbf{X}}{n}.$$

The covariance matrix can be partitioned for the subset S as

$$\Sigma = \begin{bmatrix} \Sigma_{11} = \Sigma(S) & \Sigma_{12}(S) \\ \Sigma_{21}(S) & \Sigma_{22} = \Sigma(S^c) \end{bmatrix}. \tag{2}$$

Similarly, we partition the parameter vector for the subset S as

$$\beta_{p\times 1} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix}. \tag{3}$$

The sign function is defined as

$$sign(x) = \begin{cases} -1 & \text{if } x < 0 \\ 0 & \text{if } x = 0 \\ 1 & \text{if } x > 0 \end{cases} \tag{4}$$

2.2 Clustering of Variables

We use correlation based, bottom-up agglomerative hierarchical clustering methods to cluster predictors, which forms groups of variables based on correlations between them. To determine number of clusters we use the bootstrap approach, see *stability* feature in [4].

2.3 The Regularized Regression Methods

In this section, we briefly review the Lasso, the Group Lasso and the Elastic-net regularized regression methods.

The Least Absolute Shrinkage and Selection Operator (Lasso) was introduced by Tibshirani [11]. It is a penalized least squares method that imposes an ℓ_1-penalty on the regression coefficients, which does both shrinkage and automatic

variable selection simultaneously due to the nature of the ℓ_1-penalty. We denote $\hat{\beta}_{LASSO}$, as a Lasso estimated parameter vector, which is defined as:

$$\hat{\beta}_{LASSO} \in \underset{\beta \in \mathbb{R}^p}{argmin}\{\frac{1}{n}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1\}, \tag{5}$$

where λ is the regularization parameter. The Lasso estimated active set is denoted as $\hat{S}_{LASSO}$ and defined as $\hat{S}_{LASSO} = \{j; (\hat{\beta}_{LASSO})_j \neq 0\}$.

The major disadvantages of the Lasso are: the Lasso tends to select single variable from any group and, it can select at most n variables for $p \gg n$ case. To overcome the above limitations of the Lasso, the Elastic-net was proposed by [17]. The estimated parameter vector by naive Elastic-net is denote by β_{EN}, and defined as

$$\hat{\beta}_{EN} = \underset{\beta \in \mathbb{R}^p}{argmin}\{\frac{1}{n}\|\mathbf{y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1 + \alpha\|\beta\|_2\}, \tag{6}$$

where $\alpha \in [0, 1]$ and $\lambda \geq 0$ are the regularization parameters. Since the Elastic-net estimate is $(1 + \lambda^2)\hat{\beta}_{EN}$, it selects the same variable as the $\hat{\beta}_{EN}$, we consider the $\hat{\beta}_{EN}$ as the Elastic-net estimate. The Elastic-net estimated active set is denoted as $\hat{S}_{EN}$ and defined as $\hat{S}_{EN} = \{j; (\hat{\beta}_{EN})_j \neq 0\}$.

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, then the Group Lasso (see [15]) or its variants (i. e., Group Square-Root Lasso [3] and Adaptive group Lasso [14] are used, that imposes an L_2-penalty on the coefficients within each group to achieve group sparsity.

Here we define some more notations and state assumptions for the group Lasso. We may interchangeably use β^0 and β for the true regression coefficient vector, the later one is without the superscript. Let us assume that the parameter vector β is structured into groups, $G = \{G_1, ..., G_q\}$, where $q < n$, denotes the number of groups. The partition G basically builds a partition of the index set $\{1, ..., p\}$ with $\cup_{r=1}^{q} G_r = \{1, ..., p\}$ and $G_r \cap G_l = \emptyset, \quad \text{for } r \neq l$. The parameter vector β, then has the structure $\beta = \{\beta_{G_1}, ..., \beta_{G_q}\}$ where $\beta_{G_j} = \{\beta_r : r \in G_j\}$. The columns of each group are represented by $\mathbf{X}^{G_j}$, then the response vector $\mathbf{Y}$ can also be written as

$$\mathbf{Y} = \sum_{j=1}^{q} \mathbf{X}^{(G_j)}\beta_{G_j} + \epsilon,$$

The loss function of the group Lasso is the same as the loss function of the Lasso $\frac{1}{n}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2$. The group Lasso penalty is defined as

$$\|\beta\|_{2,1} = \textstyle\sum_{j=1}^{q} \|\mathbf{X}^{G_j}\beta_{G_j}\|_2\sqrt{\frac{m_j}{n}},$$

where $m_j = |G_j|$ is the group size. Since the penalty is invariant under parametrizations within-group. Therefore, without loss of generality, we can assume $\Sigma_{rr} = I$. Hence the group Lasso penalty can be written as

$$\|\beta\|_{2,1} = \sum_{j=1}^{q} \sqrt{m_j}\|\beta_{G_j}\|_2$$

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

$$\hat{\beta}_{grp} \in \underset{\beta}{argmin}\{\frac{1}{n}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_{2,1}\} \tag{7}$$

Let W denote the active group set, $W \subset \{1, ..., q\}$, with cardinality $w = |W|$, then we assume: (i) the size of the each group is less than the number of observations $m_{max} < n$, and (ii) the number of active groups, w, is less than the number of observations (sparsity assumption). We denote the clusters selected by the group Lasso as $\hat{S}_{clust}$, which is defined as

$$\hat{S}_{clust} = \{r : \text{ cluster } G_r \text{ is selected, } r = 1, ..., q\} \tag{8}$$

The union of the selected clusters gives the selected set of variables.

$$\hat{S}_{grp} = \cup_{r \in \hat{S}_{clust}} G_r \tag{9}$$

2.4 The Cluster Lasso Methods

When group structure is not known, the cluster Lasso methods are preferred, they perform clustering followed by the sparse estimation. The clusters $G_1, ..., G_q$ are generated from the design matrix $\mathbf{X}$ (i.e. using correlation based method etc.). When the group Lasso is applied to the resulting clusters, it is called *Cluster Group Lasso*, and when the Lasso is applied for the cluster representatives, it is called *Cluster Representative Lasso*, see [2].

3 Mathematical Theory of the Lasso and its Variants

In this section, we briefly review the results required for proving consistent variable selection (and group variable selection) in high dimensional linear models. For more details on mathematical theory for the Lasso and the group Lasso, we refer to: [6], [16], [14], [2] and [1], and for the Elastic-net we refer to [9].

Let us assume a fixed set $S \subset \{1, ..., p\}$ with cardinality $s = |S|$ and partitions of the covariance matrix and parameter vector as given by the Equations (2) and (3) respectively, for the following definitions.

Definition 1 (Irrepresentable Condition (IC)). *The strong irrepresentable condition is said to be met for the set S with a constant $\eta > 0$, if the following holds:*

$$\|\Sigma_{12}\Sigma_{11}^{-1}(S)sign(\beta_1)\|_\infty \leq 1 - \eta. \tag{10}$$

Sufficient conditions (eigenvalue and mutual incoherence) on the design matrix to hold IC are discussed in [16] and [8].

Definition 2 (Beta-min Condition). *The Beta Min Condition is met for the regression coefficient β, if* $\min |\beta_S| \geq \frac{4\lambda s}{\phi^2(S)}$,

where $\phi(S)$ is the compatibility condition, see [1].

Lemma 1. *Under the following assumptions the Lasso selects the set S with a high probability:*

- *Irrepresentable condition holds for S.*
- *Beta-min condition holds for the parameter vector β.*

Definition 3 (The Elastic-net Irrepresentable Condition (EIC)). *The EIC is met for the set S, with a constant $\eta > 0$, if the following holds.*

$$\|\Sigma_{21}(\Sigma_{11} + \frac{\alpha}{n}I)^{-1}(sign(\beta_1^0) + \frac{2\alpha}{\lambda}\beta_1^0)\|_\infty \leq 1 - \eta$$

Proposition 1. *For a given set S, if IC holds then it implies that for any $\lambda > 0$, there exists α, such that EIC holds, but EIC does not imply IC.*

When IC holds it is easy to show that the EIC also holds, since $\alpha = 0$ makes the EIC same as IC. The EIC may hold even when Σ_{11} is not invertible, it proves that the EIC does not imply the IC .

Now we discuss necessary conditions for the group variable selection. Our error analysis for the group Lasso is based on the pure active group and pure noise group assumptions, that is all variables are active variables within an active group and no variables are active in a noise group and we also assume that the clustering process identifies the group structure correctly.

It is convenient to assume the following. Let $W \subset \{1, ..., q\}$ be a group index set, say, $W = \{1, ..., w\}$ Consider the full index set corresponding to W as

$$S = \{(1,1), ..., (1,m_1), ..., (w,1), ..., (w,m_w)\} = \{1, ..., s\}$$

where $s = \sum_{j=1}^{w} m_j$. We partition the $\Sigma_{11}(S)$ covariance matrix group wise, and denote its inverse as R_S. (here we assume that each $\Sigma_{r,r}$ is non-singular, or one may use the pseudo inverse)

$$R_S = \begin{bmatrix} I & \Sigma_{11}^{-1/2}\Sigma_{12}\Sigma_{22}^{-1/2} & ... & \Sigma_{11}^{-1/2}\Sigma_{1w}\Sigma_{ww}^{-1/2} \\ \Sigma_{22}^{-1/2}\Sigma_{21}\Sigma_{11}^{-1/2} & I & ... & \Sigma_{22}^{-1/2}\Sigma_{2w}\Sigma_{ww}^{-1/2} \\ \vdots & \vdots & \ddots & \vdots \\ \Sigma_{ww}^{-1/2}\Sigma_{w1}\Sigma_{11}^{-1/2} & \Sigma_{ww}^{-1/2}\Sigma_{w2}\Sigma_{22}^{-1/2} & ... & I \end{bmatrix}$$

We note that diagonal elements are $I_{m_r \times m_r}$ identity matrix due to within group parameterization invariance property.

Definition 4 (Group Irrepresentability Condition (GIC)). *The GIC is met for the set W with a constant $\eta > 0$, if the following holds*

$$\|(\Sigma_{21}R_S sign(\beta))_{G_r}\| \leq 1 - \eta \quad \forall r \notin W, \tag{11}$$

where the inequality holds group wise.

We note that the GIC definition reduces to the IC for singleton groups.

Definition 5 (Group beta-min Condition). *The group beta-min Condition is met for β , if $\|\beta^{G_r}\|_\infty \geq \frac{D\lambda\sqrt{m_r}}{n} \quad \forall r \in W$, where $D > 0$ is a constant which depends on σ, ϕ_{grp} and other constants used in cone constraints and GIC.*

For its exact form, we refer to [3]. We note that, only one component of the $\beta^{G_r}, \forall r \in W$ has to be sufficiently large, because we aim to select groups as a whole, and not individual variables.

Theorem 1. *Under the following assumptions the group Lasso selects the set of active groups W with a high probability.*

- *The GIC holds for W.*
- *Group beta-min condition holds for $\beta^{G_r}, \forall r \in W$.*

Now we show that the IC implies the GIC, and the converse is not true.

Lemma 2. *If the IC holds for the set S, then the GIC holds for the set W, but the converse is not true.*

Here we give sketch of the proof. If the IC holds for the set S, then the GIC holds for the singleton groups, therefore the GIC holds for any group structure within S. It is easy to show that the Σ_{11} is not invertible when active variables are correlated, but the Σ_{11} may be invertible after within group transformation and GIC may hold. Therefore IC is more restrictive than GIC, hence converse of the theorem is not true.

4 Pre-Selection for the Cluster Lasso Methods

We consider high dimensional settings, where group of variables are highly correlated. It is known that the Lasso tends to select one or few variables from the group of highly correlated variables, even though many or all of them belong to the active set. The CL methods have proven to be effective in selecting group variables and reducing false negatives. The two major drawbacks of the CL methods are: selection of groups does not work well when there is a large number of groups, and the time complexity for clustering when p is large is unacceptable. We try to address these problems using Elastic-net as a pre-selection for the CL methods, as described in the Algorithm 1. The variable selection consistency for the Elastic-net and the CL methods have been already proven under the EIC and the GIC respectively, see [9] and [2]. In the following, we discuss various situations where Elastic-net+CL consistently selects the active set with a high probability, and situations where our scheme may fail to select the true variables.

Algorithm 1: The Elastic-net+CL Algorithm

Input: dataset **(Y,X)**
Output: $\hat{S}$:= set of selected variables
Steps:
1. Perform Elastic-net on data **(Y,X)**, and denote $\hat{S}_{EN}$ as variable set selected.
Let $\mathbf{X}_{red} = \mathbf{X}^{\hat{S}_{EN}}$ be the reduced design matrix.
2. Perform Clustering of variables on data $\mathbf{X}_{red}$, and denote the clusters as $G_1, ..., G_q$.
3. Perform group Lasso on $(\mathbf{Y}, \mathbf{X}_{red})$ with group information $G_1, ..., G_q$, and denote the selected set of groups as
$\hat{S}_{cluster} = \{r;\ \text{cluster } G_r \text{ is selected}, r = 1, ..., q\}$.
The union of the selected groups is then the selected set of variables
$\hat{S} = \cup_r r \in \hat{S}_{cluster}$
return $\hat{S}$

Case 1 The IC holds:
Suppose that for a given set S, the design matrix $\mathbf{X}$ satisfies the IC, which implies that the EIC and the GIC will also be satisfied for the set S, see Proposition 1 and Lemma 2. It follows that with a high probability (w.h.p.) the Elastic-net will select the true variables and similarly, for any group structure within S will yield the group selection consistency for the group Lasso, w.h.p. Therefore in this situation, the Elastic-net+CL shows variable selection consistency.

Case 2 The EIC and the GIC both hold:
Apart from the previous case, this situation may also arise when two or more variables within set S are highly correlated, then IC does not hold for the set S (since Σ_{11} is not invertible), but the EIC and the GIC will be satisfied. Using the same argument as for the preceding case, it is easy to show that the combination has variable selection consistency.

Case 3 The EIC holds but the GIC does not hold:
This situation may arise for the overlapping groups, when the active groups are highly correlated, then the design matrix does not satisfy the GIC, but the EIC may hold for the set S. In such cases, though the Elastic-net may select true active set w.h.p., but the subsequent group Lasso may fail to select the active set.

Case 4 The GIC holds but the EIC does not hold:
This situation can come when there are near linear dependences among set of active variables, the EIC may not hold and Elastic-net tends to select single or a few variables from the group of linearly dependent variables. If we correctly put the linearly dependent variables in appropriate clusters (one may use the canonical correlation based clustering, see [2]) then GIC may hold. Since, the pre-selection step itself fails to select the true active set, hence our scheme of pre-selection for CL methods may not work in this case.

We illustrate the above four cases with simulation studies in the next section. We do not consider the case when the EIC and the GIC both do not hold, because

neither the pre-selection nor the CL methods are guaranteed to select the true variables.

5 Numerical Results

In this section, we consider four simulation settings and a real data example in order to empirically compare performance of the proposed scheme with the other existing methods; the Lasso [11], the Elastic-net [17] and the CGLCor [2].

5.1 Simulation Studies

Four examples are considered in this simulation that correspond to the cases discussed in the preceding section. In each example, data is simulated from the linear model in Equation (1) with fixed design $\mathbf{X}$. For each example, our simulated dataset consisted of a training and an independent validation set and 50 such datasets were simulated for each example. The models were fitted on the training data for each 50 datasets and the model with the lowest test set Mean Squared Error (MSE) was selected as the final model. For model interpretation, we consider the pair (TN, FN), where TN denotes true negative, that is the number of zero coefficients that are correctly estimated by zero and FN denotes false negative, that is the number of non zero coefficients that are incorrectly estimated by zero. As our aim is to avoid false negatives, we mainly focus on FN while analysing the different estimators. The MSE and the (TN, FN) are defined as follows.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{12}$$

$$TN = |\hat{S}^c \bigcap S^c| \tag{13}$$

$$FN = |\hat{S}^c \bigcap S| \tag{14}$$

We generate covariates from $N_p(0, \Sigma_i)$ with p=1000 and n=100, where Σ_i for four different cases are defined later. $\epsilon \sim N_p(0, \sigma^2 I)$, we set the $\sigma = 3$. For the regression coefficient we set $\beta^0 = \{3, ..., 3, 0, ..., 0\}$, where the first twenty variables are set as active variables, $S = \{1, ..., 20\}$, and the remaining variables are noise features.

The Orthogonal Model: In this case, the variables are uncorrelated that implies Σ_1 is a 40×40 identity matrix, hence the IC, EIC and GIC hold for the set S.

The simulation results (the minimum MSE with standard deviation for 50 runs, and the minimum TN and FN) are presented in the Table 1, from which it is easy to interpret that all the estimators report no false negatives.

Table 1. The Orthogonal Model

Model	MSE	TN	FN
Lasso	34.234 (36.6)	919	0
Enet	35.847 (35.9)	916	0
CGLCor	36.332 (35.5)	929	0
Enet+CGLCor	28.112 (35.9)	933	0

Block Diagonal Model: We consider a block diagonal model, where we simulate $\mathbf{X} \sim N_p(0, \Sigma_2)$, where Σ_2 consists of 100 block matrices of size 10×10, defined as

$$block_{j,k} = \begin{cases} 1, & j = k \\ .9, & otherwise \end{cases}$$

The Table 2 shows the simulation results for the block diagonal model. As expected, the Lasso fails to select the correlated active components. However, the Elastic-net, CGLCor and Elastic-net+CGLCor report no false negatives.

Table 2. Block Diagonal Model

Model	MSE	TN	FN
Lasso	8.19 (1.9)	944	12
Enet	8.18 (1.7)	950	0
CGLCor	10.97 (2.5)	979	0
Enet+CGLCor	7.52 (1.8)	950	0

Linear Dependency Model: This situation arises, when a covariate is near linear dependent on two or more covariates. Here we simulate $\mathbf{X} \sim N_p(0, I)$, then we set the linear dependency as follows.

$$\begin{aligned} X_{10} &= X_1 + ... + X_9 + N(0, .01) \\ \vdots &= \quad \vdots \\ X_{1000} &= X_{991} + ... + X_{999} + N(0, .01) \end{aligned}$$

Here we use the canonical correlation based clustering technique for CL methods, see CGL method in [2]. The Table 3 shows the simulation results for the linear dependency model, from which it is clear that the Lasso, the Elastic-net and the Elastic-net+CGL could not avoid false negatives.

Table 3. Linear Dependency Model

Model	MSE	TN	FN
Lasso	6.65 (1.6)	973	15
Enet	6.86 (1.6)	972	15
CGL	9.09 (2.0)	979	0
Enet+CGL	7.41 (2.1)	968	18

Correlation within active groups: We consider similar set up as block diagonal model except we make the first two groups (active groups) correlated. In this case, the smallest eigenvalue of the Σ_{11} is very small and the GIC does not hold. But the EIC will be satisfied since the correlation is high within active groups only. Though, the Elastic-net is consistent but the Elastic-net+CGLCor may perform poorly.

Table 4. Results for Correlated Groups

Model	MSE	TN	FN
Lasso	8.19 (1.9)	944	12
Enet	8.16 (1.6)	950	0
CGLCor	11.23 (2.5)	979	5
Enet+CGLCor	7.67 (1.5)	965	4

The simulation results are presented in the Table 4, from which it is clear that only Elastic-net method correctly identifies the true active set.

5.2 Real Data Example

In this section, we consider the gene selection problem in leukaemia data (see [13]) to compare the Elastic-net+CL scheme with other methods. The leukaemia data consists of 7129 genes and 72 samples, among which 49 are type 1 leukaemia and 23 are type 2 leukaemia. We consider part of the leukaemia dataset, we first select the 50 most significant genes and then the 50 least significant genes, according to their t-statistic scores (see [12]), so that the true active set is $S = \{1, ..., 50\}$. We used tenfold Cross-Validation method for choosing the regularization parameters.

The simulation results (TN and FN along with cross-validation error) are presented in the Table 5, from which it is clear that the Elastic-net+CGLCor outperforms in terms of model interpretation and variable selection.

Table 5. Leukemia Gene Selection Results

Model	CVError	TN	FN
Lasso	0.021	42	43
Enet	0.028	12	3
CGLCor	0.034	37	5
Enet+CGLCor	0.027	44	3

6 Computational Details

Statistical analysis was performed in R 3.2.2. We used the package *glmnet* for penalized regression methods (the Lasso and the Elastic-net), the package *gglasso* to perform group Lasso and the package *ClustOfVar* for clustering of variables.

7 Conclusion

In this article, we proposed pre-selection using Elastic-net for CL methods for variable selection in high-dimensional linear model with strongly correlated variables. We proved that the variable set selected by the Elastic-net+CL method contains the true active set consistently under IC or EIC+GIC and beta-min conditions. We discussed that reducing dimensionality improves the speed and accuracy of the clustering process, and we explored theoretically how the proposed combination of algorithms performed under various conditions. Our simulation studies showed that combining Elastic-net and Cluster Lasso methods improved variable selection and predictive performance.

Acknowledgment

The authors would like to thank the anonymous reviewers of this paper for their suggestions and constructive comments.

References

1. Bühlmann, P., van de Geer, S.: Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Verlag (2011)
2. Bühlmann, P., Rütimann, P., van de Geer, S., Zhang, C.H.: Correlated variables in regression: clustering and sparse estimation. Journal of Statistical Planning and Inference 143, 1835–1871 (2012)
3. Bunea, F., Lederer, J., She, Y.: The group square-root lasso: Theoretical properties and fast algorithms. nformation Theory, IEEE Transactions on 60, 1313–1325 (2014)
4. Chavent, M., Kuentz, V., Liquet, B., Saracco, J.: Clustofvar: an r package for the clustering of variables. The R user conference, University of Warwick Coventry UK pp. 63–72 (2011)
5. Gauraha, N.: Stability feature selection using cluster representative lasso. In Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods pp. 381–386 (2016)
6. van de Geer, S., Bühlmann, P.: On the conditions used to prove oracle results for the lasso. Electronic Journal of Statistics 3, 1360–1392 (2009)
7. H., B., B., R.: Simultaneous regression shrinkage, variable selection and clustering of predictors with oscar. Biometrics pp. 115–123 (2008)
8. Hastie, T., Tibshirani, R., Wainwright, M.: Statistical Learning with Sparsity: The Lasso and Generalizations. CRC Press (2015)
9. Jia, J., Yu, B.: On model selection consistency of the elastic net when $p \gg n$. Technical Report (2008)
10. Segal, M., Dahlquist, K., Conklin, B.: Regression approaches for microarray data analysis. Journal of Computational Biology 10, 961–980 (2003)
11. Tibshirani, R.: Regression shrinkage and selection via the lasso. J. R. Statist. Soc 58, 267–288 (1996)
12. Tibshirani, R., Hastie, T., Narasimhan, B., , Chu, G.: Diagnosis of multiple cancer types by shrunken centroids of gene expression. PNAS pp. 6567–6572 (2002)
13. TR, G., DK, S., P, T., C, H., M, G., JP, M., H, C., ML, L., JR, D., MA, C., CD, B., ES, L.: Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science pp. 531–537 (1999)
14. Wei F, H.J.: Consistent group selection in high-dimensional linear regression. Bernoulli 16, 1369–1384 (2010)
15. Yuan, M., Lin, Y.: Model selection and estimation in regression with grouped variables. J. R. Statist. Soc 68(1), 49–67 (2007)
16. Zhao, P., Yu, B.: On model selection consistency of lasso. Journal of Machine Learning Research 7, 2541–2563 (2006)
17. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Statist. Soc 67, 301–320 (2005)