Proceedings of the KI 2017 Workshop on Formal and Cognitive Reasoning

A New Combined Approach for Inference in High-Dimensional Regression Models with Correlated Variables

Niharika Gauraha and Swapan Parui

Indian Statistical Institute

Abstract. We consider the problem of model selection and estimation in sparse high dimensional linear regression models with strongly correlated variables. First, we study the theoretical properties of the dual Lasso solution, and we show that joint consideration of the Lasso primal and its dual solutions are useful for selecting correlated active variables. Second, we argue that correlation among active predictors is not problematic, and we derive a new weaker condition on the design matrix, called Pseudo Irrepresentable Condition (PIC). Third, we present a new variable selection procedure, Dual Lasso Selector, and we show that PIC is a necessary and sufficient condition for consistent variable selection for the proposed method. Finally, by combining the dual Lasso selector further with the Ridge estimation even better prediction performance is achieved. We call the combination, DLSelect+Ridge. We illustrate the DLSelect+Ridge method and compare it with popular existing methods in terms of variable selection and prediction accuracy by considering a real dataset.

Keywords: Correlated Variable Selection, High-dimensional Regression, Lasso, Dual Lasso, Ridge Regression

1 Introduction and Motivation

We start with the standard linear regression model as

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

with response vector $\mathbf{Y}_{n\times 1}$, design matrix $\mathbf{X}_{n\times p}$, true underlying coefficient vector $\beta_{p\times 1}$ and error vector $\epsilon_{n\times 1} \sim N_n(0, I)$. In particular, we consider the case of sparse high dimensional linear model ($p \gg n$) with strong correlation among a few variables. The Lasso is a widely used regularized regression method to find sparse solutions, the Lasso estimator is defined as

$$\hat{\beta}_{Lasso} = \arg\min_{\beta\in\mathbb{R}^p} \left\{ \frac{1}{2}\|\mathbf{Y} - \mathbf{X}\beta\|_2^2 + \lambda\|\beta\|_1 \right\}, \tag{2}$$

where $\lambda \geq 0$ is the regularization parameter that controls the amount of regularization. It is known that the Lasso tends to select a single variable from a

group of strongly correlated variables even if many or all of these variables are important. In the presence of correlated predictors, the concept of clustering or grouping correlated predictors and then pursuing group-wise model fitting was proposed (see [4] and [5]). When the dimension is very high or in case of overlapping clusters, finding an appropriate group structure remains as difficult as the original problem. An alternative approach is simultaneous clustering and model fitting that involves combination of two different penalties. For example, Elastic-Net [15] is a combination of two regularization techniques, the ℓ_2 regularization provides grouping effects and ℓ_1 regularization produces sparse models. Therefore, Elastic-Net selects or drops highly correlated variables together that depends on the amount of ℓ_1 and ℓ_2 regularization.

The influence of correlations on Lasso prediction has been studied in [7] and [6], and it is shown that Lasso prediction works well in presence of any degree of correlations with an appropriate amount of regularization. However, studies show that correlations are problematic for parameter estimation and variable selection. It has been proven that the design matrix must satisfy the following two conditions for the Lasso to perform exact variable selection: irrepresentability condition (IC) [14] and beta-min condition [2]. Having highly correlated variables implies that the design matrix violates IC, and the Lasso solution is not stable. When active covariates are highly correlated, the Lasso solution is not unique and Lasso randomly selects one or a few variables from a correlated group. However, even in case of highly correlated variables the corresponding dual Lasso solution is always unique. The dual of the Lasso problem as given in equation (2), as shown in [13] is given by

$$\begin{aligned} &\sup_{\theta} \ \frac{1}{2}\|\mathbf{Y}\|_2^2 - \|\theta - \mathbf{Y}\|_2^2 \\ &\text{subject to } |X_j^T\theta| \leq \lambda \text{ for } j = 1, ..., p, \end{aligned} \tag{3}$$

where θ is the dual vector. The intuitions drawn from the articles [11] and [13] further motivate us to consider the Lasso optimal and its dual optimal solution together, that yields in selecting correlated active predictors.

Exploiting the fact about uniqueness of the dual Lasso solution, we propose a new variable selection procedure; the Dual Lasso Selector (DLS). For a given Lasso estimator $\hat{\beta}_{Lasso}(\lambda)$, we can compute the corresponding dual Lasso solution by the following relationship between the Lasso solution and its dual (see [13] for the derivation):

$$\hat{\theta}(\lambda) = \mathbf{Y} - \mathbf{X}\hat{\beta}_{Lasso}(\lambda). \tag{4}$$

Basically, the DLS active set (to be defined later), corresponds to the predictors that satisfy dual Lasso feasible boundary conditions (we discuss it in details in a later section). We argue that correlation among active predictors is not problematic, and we define a new weaker condition on the design matrix that allows for correlation among active predictors, called Pseudo Irrepresentable Condition (PIC). We show that the PIC is a necessary and sufficient condition

for the proposed dual Lasso selector to select the true active set (under the assumption of beta-min condition) with a high probability. Moreover, we use the ℓ_2 penalty (the Ridge regression, [8]) which is known to perform best in case of correlated variables, to estimate the coefficients of the predictors selected by the dual Lasso selector. We call the combination of the two, DLSelect+Ridge. The DLSelect+Ridge resembles the Ridge post Lasso, but it is conceptually different and behaves differently from the Lasso followed by the Ridge, especially in the presence of highly correlated variables. Moreover, DLSelect+Ridge sounds like Elastic-net, since both are combinations of ℓ_1 and ℓ_2 penalties but Elastic-net is a combination of the Ridge Regression followed by the Lasso. In addition, Elastic-Net needs to cross-validate on a two-dimensional surface $O(k^2)$ to select its optimal regularization parameters, whereas DLSelect+Ridge needs to cross validate twice on one-dimensional surface $O(k)$, where k is the length of the search space for a regularization parameter.

We have organized the rest of the paper in the following manner. We start with background in Section 2. In Section 3, we present Dual Lasso Selector, we define PIC and discuss variable selection consistency under this assumption on the design matrix and we illustrate the proposed method on a real set. We shall provide some concluding remarks in Section 4.

2 Notations and Assumptions

In this section, we state notations and assumptions, used throughout the paper. We consider usual sparse high-dimensional linear regression model as given in equation (1) with $p \gg n$. For the design matrix $\mathbf{X} \in \mathbb{R}^{n \times p}$, we represent rows by $x_i^T \in \mathbb{R}^p,\ i = 1, ..., n$, and columns by $X_j \in \mathbb{R}^n,\ j = 1, ..., p$. We assume that the design matrix $\mathbf{X}_{n \times p}$ is fixed, the data is centred and the predictors are standardized, so that $\sum_{i=1}^n \mathbf{Y}_i = 0$, $\sum_{i=1}^n (X_j)i = 0$ and $\frac{1}{n}\mathbf{X}_j^T\mathbf{X}_j = 1$ for all $j = 1, ..., p$. We denote by $S = \{j \in \{1, ..., p\} : \beta_j \neq 0\}$, the true active set and cardinality s of the set S. We assume that the true coefficient vector β is sparse, that is $s \ll p$. We denote $\mathbf{X}_S$ as the restriction of $\mathbf{X}$ to columns in S, and β_S is the vector β restricted to the support S, with zero outside the support S. Without loss of generality we can assume that the first s variables are the active variables, and we partition the covariance matrix, $C = \frac{1}{n}\mathbf{X}^T\mathbf{X}$, for the active and the redundant variables as follows.

$$C = \begin{bmatrix} C_{11}\ C_{12} \\ C_{21}\ C_{22} \end{bmatrix} \tag{5}$$

Similarly, the coefficient vector β can be partitioned as $(\beta_1\ \beta_2)^T$. The ℓ_1-norm and ℓ_2-norm (square) are defined as $\|\beta\|_1 = \sum_{j=1}^p |\beta_j|$ and $\|\beta\|_2^2 = \sum_{j=1}^p \beta_j^2$ respectively. The sub-gradient $\partial\|\beta\|_1$ and sign function $sign(\beta)$ are defined as follows.

$$\partial\|\beta\|_1 = \begin{cases} 1 & \text{if } \beta_i > 0 \\ [-1, 1] & \text{if } \beta_i = 0 \\ -1 & \text{if } \beta_i < 0 \end{cases}, \quad sign(\beta) = \begin{cases} 1 & \text{if } \beta_i > 0 \\ 0 & \text{if } \beta_i = 0 \\ -1 & \text{if } \beta_i < 0 \end{cases} \tag{6}$$

3 Dual Lasso Selector

In this section, we present the dual Lasso selector, a variable selection method for sparse high-dimensional regression models with correlated variables. First, we state the basic properties of the Lasso and its dual, which have already been derived and studied by various authors, see [12] and [13] for more details.

1. **Uniqueness of the Lasso-fit**: There may not be a unique solution for the Lasso problem because for the criterion as given in equation (2) is not strictly convex in β. But the least square loss is strictly convex in $\mathbf{X}\beta$, hence there is always a unique fitted value $\mathbf{X}\hat{\beta}$.
2. **Uniqueness of the dual vector**: The dual problem is strictly convex in θ, therefore the dual optimal $\hat{\theta}$ is unique. Another argument for the uniqueness of $\hat{\theta}$ is that it is a function of $\mathbf{X}\hat{\beta}$ as given in equation (4), which itself is unique. The fact that the DLS can achieve consistent variable selection for situations (with correlated active predictors) when the Lasso is unstable for estimation of the true active set is related to the uniqueness of the dual Lasso solution.
3. **Uniqueness of the Sub-gradient**: Sub-gradient of ℓ_1 norm of any Lasso solution $\hat{\beta}$ is unique because it is a function of $\mathbf{X}\hat{\beta}$. More specifically, suppose that $\hat{\beta}$ and $\tilde{\beta}$ are two Lasso solutions for a fixed λ value, then they must have the same signs $sign(\hat{\beta}) = sign(\tilde{\beta})$. It is not possible that $\hat{\beta}_j > 0$ and $\hat{\beta}_j < 0$ for some j.

Let $\hat{S}_{Lasso}$ denote the support set or active set of the Lasso estimator $\hat{\beta}$ which is given as $\hat{S}_{Lasso}(\lambda) = \{j \in \{1, ..., p\} : (\hat{\beta}_{Lasso})_j \neq 0\}$. Similarly, we define the active set of the dual Lasso vector that corresponds to the active constraints of the dual optimization problem, $\hat{S}_{dual}(\lambda) = \{j \in \{1, ..., p\} : \ |X_j^T\theta| = \lambda\}$. Now, we state the following lemmas that will be used later for our mathematical derivations.

Lemma 1. *The active set selected by the Lasso $\hat{S}_{Lasso}(\lambda)$ is always contained in the active set selected by the dual Lasso $\hat{S}_{dual}(\lambda)$, that is $\hat{S}_{Lasso}(\lambda) \subseteq \hat{S}_{dual}(\lambda)$.*

Proof. The proof is rather easy. From KKT conditions (see [13]), we have

$$|X_j^T\theta| < \lambda \implies \hat{\beta}_j = 0 \tag{7}$$

The proof lies in the implication in the above equation (7).

It is known that IC (assuming beta-min conditions holds throughout the paper) is a necessary and sufficient condition for the Lasso to select the true model (see [14]).

Lemma 2. *Under the assumption of IC on the design matrix, the active set selected by the Lasso $\hat{S}_{Lasso}(\lambda)$ is equal to the active set selected by the dual Lasso $\hat{S}_{dual}(\lambda)$, that is $\hat{S}_{Lasso}(\lambda) = \hat{S}_{dual}(\lambda)$.*

For proof of the above lemma lies in the uniqueness of the Lasso solution under IC assumption [14]. Suppose that we partition the covariance matrix as given in equation (5), then IC is said to be met for the set S with a constant $\eta > 0$, if the following holds:

$$\|C_{12}C_{11}^{-1} sign(\beta_1)\|_\infty \leq 1 - \eta. \tag{8}$$

The IC may fail to hold due to violation of any one (or both) of the following two conditions: 1. When C_{11} is almost not invertible, and thus there is strong correlation among variables of the true active set, 2. The active predictors are correlated with the noise features.

When there is strong correlation among variables of the active set, then C_{11} is (almost) not invertible and the IC does not hold, and the Lasso fails to do variable selection. In the following, we argue that the dual Lasso can still perform variable selection consistently even when C_{11} not invertible under the assumption of a milder condition on the design matrix, called Pseudo Irrepresentable Condition. The Pseudo Irrepresentable Condition is defined as follows.

Definition 1 (Pseudo Irrepresentable Condition (PIC)). *We partition the covariance matrix as given in equation (5). Then the PIC is said to be met for the set S with a constant $\eta > 0$, if the following holds:*

$$|X_j^T G\ sign(\beta_1)| \leq 1 - \eta,\ \textit{for all } j \in S^c, \tag{9}$$

where G is a generalized inverse of the form $\begin{bmatrix} C_A^{-1} & 0 \\ 0 & 0 \end{bmatrix}$, and equation (9) holds for each $C_A \in C_R$, where C_R is defined as $C_R := \{C_{rr} : rank(C_{rr}) = rank(C_{11}) = r, r \subseteq S\}$.

The following lemma gives a sufficient condition for the dual Lasso for support recovery. This lemma is similar in spirit to Lemma 2 defined in [10]. Here, we do not assume that C_{11} is invertible.

Lemma 3 (Primal-dual Condition for Variable Selection). *Suppose that we can find a primal-dual pair $(\hat{\beta}, \hat{\theta})$ that satisfies the following conditions:*

$$\boldsymbol{X}^T(\boldsymbol{Y} - \boldsymbol{X}\hat{\beta}) + \lambda\hat{v} = 0, \quad \textit{where } \hat{v} = sign(\hat{\beta}) \tag{10}$$

$$\hat{\theta} = \boldsymbol{Y} - \boldsymbol{X}\hat{\beta}, \tag{11}$$

$$\hat{\beta}_j = 0 \textit{ for all } j \in S^c, \tag{12}$$

$$|\hat{v}_j| < 1 \textit{ for all } j \in S^c. \tag{13}$$

Then $\hat{\theta}$ is the unique optimal solution to the dual Lasso and $\hat{S}_{dual}$ recovers the true active set.

Proof. We have shown that the dual Lasso optimal $\hat{\theta}$ is always unique, and it remains to show that $\hat{S}_{dual}$ recovers the true active set S. Under the assumption as given in equation (13), we can derive that $|X_j^T\hat{\theta}| < \lambda$ for all $j \in S^c$. Therefore $\hat{S}_{dual} = S$.

Theorem 1. *Under the assumption of PIC on the design matrix* $\boldsymbol{X}$*, the active set selected by the dual Lasso* $\hat{S}_{dual}$*, is the same as the true active set* S *with a high probability, that is,* $\hat{S}_{dual} = S$.

The proof of the above theorem is similar to the proof of Theorem 7.1 in [2], if inverse of the matrix C_{11} is replaced with its generalized inverse . We note that PIC may hold even when C_{11} is not invertible, which implies that PIC is weaker than IC. It is illustrated with the following examples.

Let $S = \{1, 2, 3, 4\}$ be the active set, and let the covariance matrix $C = \frac{\mathbf{X}^T\mathbf{X}}{n}$ of the design matrix $\mathbf{X}$ is given as $C = \begin{bmatrix} 1 & 0 & 0 & 0 & \rho \\ 0 & 1 & 0 & 0 & \rho \\ 0 & 0 & 1 & 0 & \rho \\ 0 & 0 & 0 & 1 & \rho \\ \rho & \rho & \rho & \rho & 1 \end{bmatrix}$. Here, the active variables are uncorrelated and the noise variable is equally correlated with all active covariates. First of all, it is easy to check that only for $|\rho| \leq \frac{1}{2}$, C is positive semi definite, and for $|\rho| < \frac{1}{4}$, C satisfies the IC. Now, we augment this matrix with two additional columns, one copy of the first and one copy of the second active variables, and we rearrange the columns such that we get covariance matrix, $C_1 = \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & \rho \\ 1 & 1 & 0 & 0 & 0 & 0 & \rho \\ 0 & 0 & 1 & 1 & 0 & 0 & \rho \\ 0 & 0 & 1 & 1 & 0 & 0 & \rho \\ 0 & 0 & 0 & 0 & 1 & 0 & \rho \\ 0 & 0 & 0 & 0 & 0 & 1 & \rho \\ \rho & \rho & \rho & \rho & \rho & \rho & 1 \end{bmatrix}$. Suppose that the set of active variables is $S = \{1, 2, 3, 4, 5, 6\}$ and we assume that $|\rho| < \frac{1}{4}$. We partition C_1 as given in equation (5), and it is clear that the corresponding sub-matrix C_{11} is not invertible and IC does not hold, hence the Lasso may not perform variable selection. The rank of the matrix C_{11} is 4. Let us consider any (4×4) sub matrix of the matrix C_{11} such that its rank is four ($A \subset S, rank(C_A) = 4$, Here $C_R = \{\{1, 3, 5, 6\}, \{1, 4, 5, 6\}, \{2, 3, 5, 6\}, \{2, 4, 5, 6\}\}$). Further, we consider the generalized inverse of C_{11} as $C_{11}^{+} = \begin{bmatrix} C_A^{-1} & 0 \\ 0 & 0 \end{bmatrix}$, where $C_A \in C_R$ is invertible. With the above inverse C_{11}^{+} PIC holds for the design matrix $\mathbf{X}$, and the dual Lasso will select the true active set S with a high probability and will set zero to the coefficient of the noise features.

3.1 Dual Lasso Selection and Ridge Estimation

Now, we combine the dual Lasso selection with the Ridge estimation. Mainly, we consider the ℓ_2 penalty (Ridge penalty) which is known to perform best in case of correlated variables, to estimate the coefficients of the predictors selected by the dual Lasso. We develop an algorithm called DLSelect+Ridge, which is a two stage procedure, the dual selection followed by the Ridge Regression.

If model selection works perfectly (under strong assumptions, i.e. IC), then the post-model selection estimators are the oracle estimators with well behaved properties (see [1]). It has been already proven that the Lasso+OLS [1] estimator performs at least as good as Lasso in terms of the rate of convergence, and it has a smaller bias than the Lasso. Further Lasso+mLS (Lasso+ modified OLS) or Lasso+Ridge estimator have been also proven to be asymptotically unbiased under the IC, see [9]. Under the IC the Lasso solution is unique and

Algorithm 1: DLSelect+Ridge

Input: dataset $(\mathbf{Y}, \mathbf{X})$
Output: $\hat{S}$:= the set of selected variables, $\hat{\beta}$:= the estimated coefficient vector
Steps:
1. Perform Lasso on the data ($\mathbf{Y}$, $\mathbf{X}$). Denote the Lasso estimator as $\hat{\beta}_{Lasso}$.
2. Compute the dual optimal as $\hat{\theta} = \mathbf{Y} - \mathbf{X}\hat{\beta}_{Lasso}$, and denote the dual Lasso active set as $\hat{S}_{dual}$
3. Compute the reduced design matrix as $\mathbf{X}_{red} = \{X_j : j \in \hat{S}_{dual}\}$.
4. Perform Ridge regression based on the data ($\mathbf{Y}$, $\mathbf{X}_{red}$) and obtain the ridge estimator β_j for $j \in S_{dual}$. Set the remaining coefficients to zero.
return $(\hat{S}, \hat{\beta})$

the DLSelect+Ridge is the same as the Lasso+Ridge and the same argument holds for the DLSelect+Ridge. In the following section, we empirically compare the performance of DLSelect+Ridge with other methods.

3.2 Empirical Results with Riboflavin Dataset

The dataset of riboflavin consists of $n = 71$ observations of $p = 4088$ predictors (gene expressions) and univariate response, riboflavin production rate (log-transformed), see [3] for details on riboflavin dataset. Since the ground truth is not available, we consider Riboflavin data for the design matrix $\mathbf{X}$ with synthetic parameters β and simulated Gaussian errors $\epsilon \sim \mathbb{N}_n(0, \sigma^2 I)$. We fix the size of the active set to $s = 20$ and $\sigma = 1$, and for the true active set S select ten predictors which are highly correlated with the response and another ten variables which are most correlated with those selected variables. The true coefficient vector is: $\beta_j = \begin{cases} 1 & \text{if } j \in S \\ 0 & \text{if } j \notin S \end{cases}$. Then we compute the response using equation (1). We compute Mean Squared Error (MSE) and True Positive Rate (TPR) as performance measures, which are defined as follows: $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ and $TPR = |\hat{S} \bigcap S|/|S|$, where $\hat{y}_i's$ are estimated responses and $\hat{S}$ is the estimated active set. The performance measures (the median MSE with standard deviation for 100 runs, and the median TPR) are reported in Table 1. From Table 1, we conclude that DLSelect+Ridge performs better than others in terms of prediction performance, and DLSelect+Ridge is as good as Elastic-Net in terms of variable selection.

Table 1: Performance measures for Riboflavin data

Method	MSE (SE)	TPR
Lasso	135.37(62.50)	0.30
Ridge	293.96(83.42)	NA
Enet	126.31(65.67)	0.45
DLSelect+Ridge	88.31(54)	0.45

4 Concluding Remarks

The main achievements of this work are summarized as follows: we argued that the correlation among active predictors is not problematic, as long as PIC is satisfied by the design matrix. In particular, we showed that the dual Lasso performs consistent variable selection under the assumption of PIC. Exploiting this result we proposed DLSelect+Ridge method. We compared DLSelect+Ridge with the popular existing methods by considering a real dataset. The numerical studies show that the proposed method is very competitive in terms of variable selection and prediction accuracy.

References

1. Belloni, A., Chernozhukov, V.: Least squares after model selection in high-dimensional sparse models. Bernoulli 19, 521–547 (2013)
2. Bühlmann, P., van de Geer, S.: Statistics for High-Dimensional Data: Methods, Theory and Applications. Springer Verlag (2011)
3. Bühlmann, P., Kalisch, M., Meier, L.: High-dimensional statistics with a view towards applications in biology. Annual Review of Statistics and its Applications 1, 255–278. (2014)
4. 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)
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., Lederer, J.: The Lasso, correlated design, and improved oracle inequalities, Collections, vol. Volume 9, pp. 303–316. Institute of Mathematical Statistics, Beachwood, Ohio, USA (2013)
7. Hebiri, M., Lederer, J.: How correlations influence lasso prediction. IEEE Trans. Inf. Theor. 59(3), 1846–1854 (2013)
8. Hoerl, A.E., Kennard, T.W.: Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12, 55–67 (1970)
9. Liu, H., Yu, B.: Asymptotic properties of lasso+mls and lasso+ridge in sparse high-dimensional linear regression. Electron. J. Statist. 7, 3124–3169 (2013)
10. Omidiran, D., Wainwright, M.J.: High-dimensional variable selection with sparse random projections: Measurement sparsity and statistical efficiency. J. Mach. Learn. Res. 11, 2361–2386 (2010)
11. Osborne, M.R., Presnell, B., Turlach, B.A.: A new approach to variable selection in least squares problems. IMA Journal of Numerical Analysis 20(3), 389–403 (2000)
12. Tibshirani, R., Taylor, J.: The solution path of the generalized lasso. Ann. Statist. 39(3), 1335–1371 (06 2011)
13. Wang, J., Zhou, J., Wonka, P., Ye, J.: Lasso screening rules via dual polytope projection. Journal of Machine Learning Research 16, 1063–1101 (2015)
14. Zhao, P., Yu, B.: On model selection consistency of lasso. Journal of Machine Learning Research 7, 2541–2563 (2006)
15. Zou, H., Hastie, T.: Regularization and variable selection via the elastic net. J. R. Statist. Soc 67, 301–320 (2005)