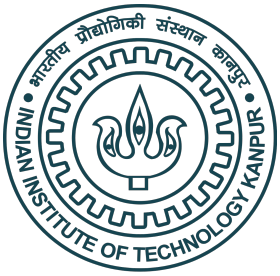
Efficient High-dimensional Robust Variable Selection via Rank-based LASSO Methods*

Submitted by:

Arkajyoti Bhattacharjee [†‡]
Rachita Mondal [§†]
Ritwik Vashishtha [¶†]
Shubha Sankar Banerjee [‖†]

Supervised by:

Dr. Subhra Sankar Dhar [†]



Submitted on:

8^{th} April, 2022

*This report has been prepared towards the partial fulfillment of the requirements of the course *MTH673A: Robust Statistical Methods.*

[†]Department of Mathematics & Statistics, Indian Institute of Kanpur, India.
[‡]201277, M.Sc. Statistics (Final year).
[§]201374, M.Sc. Statistics (Final year).
[¶]201389, M.Sc. Statistics (Final year).
[‖]201416, M.Sc. Statistics (Final year).

Abstract

Penalized variable selection is a popular approach for describing the relationship between the response, Y and explanatory variables, X. LASSO-based methods have received special attention throughout the literature of regression analysis. But stringent conditions are imposed on the $X - y$ relation and on the error distribution. In this report, we present Rank-LASSO as a robust, superior method over the general LASSO, which can be used even when number of predictors is much larger than the sample size. The major properties of the Rank-LASSO has been presented in a non-asymptotic fashion, which makes it useful for the aforementioned case of $p >> n$. The report also shows the superiority of the thresholded modified version of Rank-LASSO in more general scenarios. Apart from theoretical results, we present numerical experiments for demonstrating that performance of the Rank-LASSO is substantially better than regular LAD-LASSO in terms of robust model selection problems. The report is primarily based on Rejchel and Bogdan (2020).

Contents

1 Introduction

One of the most ubiquitous problems while dealing with high-dimensional datasets is that of variable selection. Many penalization-based model selection processes exist (see Hastie et al. (2009)). Under the linear regression model,

$$Y_i = \beta' X_i + \varepsilon_i, \quad i = 1, \ldots, n,$$

where $Y_i \in \mathbb{R}$ is a response variable, $X_i \in \mathbb{R}^p$ is a vector of covariates, $\beta \in \mathbb{R}^p$ is the vector of model parameters, and ε_i is a random error, the penalized variable selection approaches usually recommend estimating the vector of model parameters β by

$$\hat{\beta} = \arg\max_{\beta \in \mathbb{R}^p} \sum_{i=1}^{n} (Y_i - \beta' X_i)^2 + Pen(\beta),$$

where $\sum_{i=1}^{n} (Y_i - \beta' X_i)^2$ is the ℓ_2-loss function measuring the model fit and $Pen(\beta)$ is the penalty on the model complexity. One popular penalization-based estimation of β is called the LASSO (Tibshirani (1996)), which uses the ℓ_1-norm penalty. There is an extensive literature on LASSO-based variable selection, estimation and prediction, which are mainly in the context of (generalized) linear models and many properties of the LASSO hold under specific assumptions on the $X - Y$ relationship and/or the distribution of the random errors (for example, see Zhao and Yu (2006), Zou (2006)). However, real complex datasets may fail to satisfy such assumptions. Robust methods are, thus, required in such scenarios.

In this report, we consider the single-index model

$$Y_i = g(\beta' X_i, \varepsilon_i), \quad i = 1, \ldots, n, \tag{1}$$

where $g(.)$ is an unknown monotonic link function. Thus, we suppose that the covariates influence the response through the link function $g(.)$ of the scalar product $\beta' X_i$. No assumptions are made on the form of the link function g or the distribution of the error ε_i. In particular, the existence of expectation $\mathbb{E}(\varepsilon_i)$ is not assumed.

The goal of variable selection is the identification of the set of relevant covariates

$$T = \{j : 1 \leq j \leq p, \beta_j \neq 0\}. \tag{2}$$

In this report, we discuss simple robust variable selection methods that are computationally fast and can work efficiently in high-dimensional complex datasets. In the methods we discuss, the observed responses Y_i are replaced by their centered ranks. Ranks R_i are defined as

$$R_i = \sum_{j=1}^{n} \mathbb{I}(Y_j \leq Y_i), \quad i = 1, \ldots, n,$$

where $\mathbb{I}(.)$ is the indicator function. The relevant covariates are identified by solving the following LASSO problem:

$$\textbf{RankLASSO:} \quad \hat{\theta} = \arg\max_{\theta \in \mathbb{R}^p} Q(\theta) + \lambda |\theta|_1, \tag{3}$$

where

$$Q(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left(\frac{R_i}{n} - \frac{1}{2} - \theta' X_i \right)^2. \tag{4}$$

The above requires no separate algorithm for computation and can be done using the inbuilt R packages for the LASSO. In this project, we have used the "glmnet" package (Friedman et al. (2021)) in R.

Using ranks in place of response variables is quite common in non-parametric statistics and often lead to robust procedures. Zak et al. (2007), Bogdan et al. (2008) show the high efficiency of rank-based model selection in the sparse high-dimensional regression models, where the number of true non-zero regression coefficients is much smaller than the sample size n. The RankLASSO is based on a convex optimization algorithm, which can be solved easily even when $p >> n$ and the explanatory variables are highly correlated.

One of the disadvantages of the rank approach is the loss of information about the shape of the link function. Thus, RankLASSO cannot be directly used to build a predictive model for the response variable. Modifications of the RankLASSO are discussed that successfully enable the identification of significant covariates.

In this report, we discuss in Section (2) and (3) the model section consistency of RankLASSO. In Section (4), we discuss two extensions of RankLASSO- Thresholded and Weighted RankLASSO. Finally in Section (5) we perform some simulation experiments to validate the theoretical results.

1.1 Notations

In this report, we will be using the following notations:

- $X = (X_1, \ldots, X_n)$,
- $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$,
- $Z_i = (X_i, Y_i), i = 1, \ldots, n$,
- $T' = \{1, \ldots, p\} \backslash T$ is a complement of T
- X_T is a sub-matrix of X, with columns whose indices belong to the support T of β, see 2,
- θ_T is a restriction of a vector θ $in \mathbb{R}^p$ to the indices from T,
- p_0 ia the number of elements in T
- the ℓ_q-norm of a vector is defined as $|\theta|_q = (\sum_{j=1}^{p} |\theta_j|^q)^{1/q}$ for $q \in [1, \infty)$.

2 Identifiability of the support of β

We consider the single index model (1). Throughout the report, we assume that the design matrix X and the random error term ε satisfy the following assumptions.

Assumption 1. *We assume that $(X_1, \varepsilon_1), \ldots, (X_n, \varepsilon_n)$ are i.i.d. random vectors such that the distribution of X_1 is absolutely continuous and X_1 is independent of the noise variable ε_1. Additionally, we assume that $\mathbb{E}(X_1) = 0$, $H = \mathbb{E}(X_1 X_1')$ is positive definite and $H_{jj} = 1$ for $j = 1, \ldots, p$.*

The single index model (1) doesn't allow for the estimation of an intercept term and can identify β upto a multiplicative constant (see Theorem 1) since any change or shift in β can be absorbed in to g.

Assumption 2. *We assume that for each $\theta \in \mathbb{R}^p$, the conditional expectation $\mathbb{E}(\theta' X_1 | \beta' X_1)$ exists and $\mathbb{E}(\theta' X_1 | \beta' X_1) = d_\theta \beta' X_1$ for a real number $d_\theta \in \mathbb{R}$.*

Assumption 3. *We assume that the design matrix and the error term satisfy Assumptions 1 and 2, the cumulative distribution function F of the response variable Y_1 is increasing and g in 1 is increasing with respect to the first argument.*

RankLASSO does not estimate β, but the vector

$$\theta^0 = \arg\min_{\theta \in \mathbb{R}^p} \mathbb{E} Q(\theta), \tag{5}$$

where $Q(\theta)$ is defined in (4). Since H is positive definite, the minimizer θ^0 is unique and is given by the formula

$$\theta^0 = \frac{1}{n^2} H^{-1} \left(\mathbb{E} \sum_{i=1}^{n} R_i X_i \right) \tag{6}$$

See that

$$\sum_{i=1}^{n} R_i X_i = \sum_{i=1}^{n} \sum_{j=1}^{n} \mathbb{I}(Y_j \leq Y_i) X_i = \sum_{i \neq j} \mathbb{I}(Y_j \leq Y_i) X_i + \sum_{i=1}^{n} X_i$$

and that $\mathbb{E}(X_i) = 0$. So, we can rewrite (6) as

$$\theta_0 = \frac{n-1}{n} H^{-1} \mu \tag{7}$$

where $\mu = \mathbb{E}[\mathbb{I}(Y_2 \leq Y 1) X_1]$ is the expected value of the U-statistic

$$A = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}(Y_j \leq Y_i) X_i. \tag{8}$$

The following theorem shows the relation between θ^0 and β.

Theorem 1. *Consider the model (1). If Assumptions (1) and (2) are satisfied, then*

$$\theta_0 = \gamma_\beta \beta$$

with

$$\gamma_\beta = \frac{\frac{n-1}{n} \beta' \mu}{\beta' H \beta} = \frac{\frac{n-1}{n} Cov(F(Y_1), \beta' X_1)}{\beta' H \beta}, \tag{9}$$

where F is a cumulative distribution function of a response variable Y_1.

Additionally, if F is increasing and g is increasing with respect to the first argument, then $\gamma_\beta > 0$, so the signs of β coincide with the signs of θ^0 and

$$T = \{ j : \beta_j \neq 0 \} = \{ j : \theta_j^0 \neq 0 \}. \tag{10}$$

Note that for $Q(\theta)$ defined in (4), we have

$$Q(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(R_i/n - \theta' X_i)^2 + \theta'\bar{X}/2 - \frac{n+1}{4n} + 1/8.$$

Thereforem due to the fact the covariates X_i are centered, for all the proofs in this report, we consider $Q(\theta)$ without subtracting 0.5, that is

$$Q(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(R_i/n - \theta' X_i)^2.$$

This is to merely simplify notation.

We will use the following lemma to prove Theorem 1.

Lemma 1. *Let U be a random variable that is not degenerate i.e. $P(U = u) < 1$ for each $u \in \mathbb{R}$. Moreover, let $f, h : \mathbb{R} \to \mathbb{R}$ be an increasing function. Then $Cov(f(U), h(U)) > 0$.*

Proof. Clearly, we have

$$\mathbb{E}(Q(\theta)) = \frac{1}{2n^3}\sum_{i=1}^{n}\mathbb{E}R_i^2 - \frac{1}{n^2}\sum_{i=1}^{n}\mathbb{E}(R_i\theta' X_i) + \frac{1}{2n}\sum_{i=1}^{n}\mathbb{E}(\theta' X_i)^2.$$

Vectors $(X_1, Y_1), \ldots, (X_n, Y_n)$ are i.i.d. and X_i are centered, so for all $i \neq 1$

$$\begin{aligned}
\mathbb{E}(R_i\theta' X_i) &= \mathbb{E}(\mathbb{I}(Y_1 \leq Y_i)\theta' X_i) + \sum_{j\neq\{1,i\}}\mathbb{E}(\mathbb{I}(Y_j \leq Y_i)\theta' X_i) \\
&= \mathbb{E}(\mathbb{I}(Y_1 \leq Y_i)\theta' X_1) + \sum_{j\neq\{1,i\}}\mathbb{E}(\mathbb{I}(Y_j \leq Y_i)\theta' X_1) \\
&= \mathbb{E}(R_1\theta' X_1).
\end{aligned}$$

Moreover, ranks $R_1, \ldots, R_n$ are identically distributed; so $\sum_{i=1}^{n}\mathbb{E}(R_i^2) = n\mathbb{E}(R_1^2)$. Therefore, we obtain that $\mathbb{E}(Q(\theta)) = \frac{1}{2}\mathbb{E}(\frac{R_i}{n} - \theta' X_1)^2$. Using Jensen's inequality and Assumption (2), we have

$$\begin{aligned}
\mathbb{E}(Q(\theta)) &= \frac{1}{2}\mathbb{E}\mathbb{E}\left[\left(\frac{R_1}{n} - \theta' X_1\right)^2 | \beta' X_i, \varepsilon_i, i = 1, \ldots, n)\right] \\
&\geq \frac{1}{2}\mathbb{E}\left[\mathbb{E}\left(\frac{R_1}{n} - \theta' X_1\right)|\beta' X_i, \varepsilon_i, i = 1, \ldots, n)\right]^2 \\
&= \frac{1}{2}\mathbb{E}\left[\frac{R_1}{n} - \mathbb{E}(\theta' X_1 | \beta' X_1)\right]^2 \\
&= \frac{1}{2}\mathbb{E}\left[\frac{R_1}{n} - \mathbb{E}(\theta' X_1 | \beta' X_1)\right]^2 \\
&\geq \min_{d\in\mathbb{R}}\mathbb{E}(Q(d\beta)).
\end{aligned}$$

Obviously, we have $\min_{d\in\mathbb{R}}\mathbb{E}(Q(d\beta)) = \mathbb{E}(Q(\gamma_\beta\beta))$, where γ_β is defined in 9. Since θ^0 is the unique minimizer of $\mathbb{E}Q(\theta)$, we obtain the first part of the theorem. Again, denote $Z = \beta' X_1$ and $\varepsilon = \varepsilon_1$. It is clear that $\gamma_\beta > 0$ is equivalent to $Cov(Z, F(g(Z, \varepsilon))) > 0$. This covariance can be expressed as

$$\mathbb{E}(ZF(g(Z, \varepsilon))) = \mathbb{E}h(\varepsilon), \tag{11}$$

where $h(a) = \mathbb{E}[ZF(g(Z,\varepsilon))|\varepsilon = a] = \mathbb{E}(ZF(g(Z,a)))$ for arbitrary a. This fact simply follows from $\mathbb{E}(Z) = 0$ and independence between Z and ε. If F is increasing and g is increasing with respect to the first variable, then $h(a) > 0$ for arbitrary a by Lemma 1. Clearly, it implies that (11) is positive. □

3 Estimation and Separability of RankLASSO when $p >> n$

In this section the properties of Rank-Lasso will be discussed when the number of predictors is greater than the sample size. The following assumption is needed to obtain the results.

Assumption 4. *Let $(X_1)_T$ be the vector of significant predictors and suppose that it is subgaussian with coefficients $\tau_0 > 0$ i.e for each $u \in \mathbb{R}^{p_0}$ we have $\mathbb{E}exp(u^T(X_1)_T) \leq exp(\tau_0^2 u^T u/2)$. Also we have, the insignificant predictors are univariate subgaussian, i.e for each $a \in \mathbb{R}$ and $j \notin T$, $\mathbb{E}(aX_{1j}) \leq exp(\tau_j^2 a^2/2)$, for $\tau_j > 0$. Denote, $\tau = max(\tau_0, \tau_j, j \notin T)$.*

This assumption of subgaussianity is required to get exponential inequalities for the proofs which will be discussed in this section.This condition is used to work with random predictors in high dimensional model. Wang and Zhu (2015) has proven the model selection consistency of high dimensional Rank-Lasso based on restricted irrepresentable condition.Also, as it has been obtained using polynomial upper bound on the dependency of p on n, it fails to give clear idea of the tuning parameter selection. Instead of asymptotic results (Wang and Zhu (2015)) this report will state the non-asymptotic results which do not require irrepresentable condition. It allows p to increase exponentially as function of n. It also specifies the method od selecting the tuning parameter.
In case of $n > p$, usually the minimum eigen value of the matrix $X^T X/n$ is used to present the correlation between predictors. But, for large dimensions this value is zero. Then it is required to replace the minimum eigenvalue by by some other relevant measure.
Let T be the set of indices corresponding to the support of true vector β. Suppose that θ_T and $\theta_{T'}$ be the restrictions of the vector $\theta \in \mathbb{R}^p$ to indices of the indices from T and T', respectively. Now for $\zeta > 1$, a cone can be considered,

$$C(\zeta) = \{\theta \in \mathbb{R}^p : \ |\theta_{T'}|_1 \leq \zeta \, |\theta_T|_1\}$$

For $p > n$ three different characteristics have been introduced for measuring the potential for consistent estimation of model parameters.

- Restricted Eigen Value (Bickel et al. (2009)):

$$RE(\zeta) = \inf_{0 \neq \theta \in C(\zeta)} \frac{\theta^T X^T X \theta}{n \, |\theta_T|_2^2}$$

- Compatibility Factor (Van de Geer (2008)):

$$K(\zeta) = \inf_{0 \neq \theta \in C(\zeta)} \frac{p_0 \theta^T X^T X \theta}{n \, |\theta_T|_1^2}$$

- Cone Invertibility Factor(CIF, Ye and Zhang (2010)):

$$\bar{F}_q(\zeta) = \inf_{0 \neq \theta \in C(\zeta)} \frac{p_0^{1/q} \, |X^T X \theta|_\infty}{n \, |\theta_T|_q}$$

These conditions are much weaker than the irrepresentability condition. In this report the CIF will be used as it allows formulation of convergence results for any l_q norm, for $q \geq 1$. The population version of CIF is given by,

$$F_q(\zeta) = \inf_{0 \neq \theta \in C(\zeta)} \frac{p_0^{1/q} \mid H\theta \mid_\infty}{n \mid \theta_T \mid_q}$$

, where $H = E(X^T X)$.
The following theorem is presented for estimation accuracy of Rank-Lasso.

Theorem 2. *Let $a \in (0,1)$, $q \geq 1$ and $\zeta \geq 1$ be arbitrary. Suppose that Assumptions 3 and 4 are satisfied. Also,*

$$n \geq \frac{K_1 p_0^2 \tau^4 (1+\zeta)^2 log(p/a)}{F_q^2(\zeta)} \tag{12}$$

and

$$\lambda \geq K_2 \frac{\zeta+1}{\zeta-1} \tau^2 \sqrt{\frac{log(p/a)}{kn}} \tag{13}$$

where K_1, K_2 are universal constants and k is the smallest eigen value of the correlation matrix between true predictor $H_T = (H_{i,j})_{j,k \in T}$. Then there exists a universal constant K_3 such that,

$$\mid \hat{\theta} - \theta^0 \mid_q \leq \frac{4\zeta p_0^{1/q} \lambda}{(\zeta+1) F_q(\zeta)} \tag{14}$$

with probability at leasr $1 - K_3 a$.
Moreover, if $X_1 \sim Normal(0, H)$, then k, τ can be dropped from inequalities 12 and 13.

Proof. Proving the theorem will require the following lemmas.

Lemma 2. *Suppose that $Z_1, Z_2, ..., Z_n$ are i.i.d random variables and there exists $L > 0$ so that $C^2 = \mathbb{E} exp(\mid Z_1 \mid / L)$ is finite. Then for arbitrary $u > 0$,*

$$P\left(\frac{1}{n} \sum_{i=1}^{n} (Z_i - \mathbb{E}(Z_i)) > 2L \left(C\sqrt{\frac{2u}{n}} + \frac{u}{n} \right) \right) \leq exp(-u)$$

Lemma 3. *Consider a U-statistic, $U = \frac{1}{n(n-1)} \sum_{i \neq j} h(Z_i, Z_j)$, where h is a kernel based on i.i.d random variables $Z_1, ..., Z_n$. Suppose that, there exists $L > 0$ so that $C^2 = \mathbb{E} exp(\mid h(Z_1, Z_2) \mid / L)$ is finite. Then for arbitrary $u > 0$,*

$$P\left(U - \mathbb{E}(U) > 2L \left(C\sqrt{\frac{6u}{n}} + \frac{3u}{n} \right) \right) \leq exp(-u)$$

Lemma 4. *If assumptions 3 and 4 are satisfied, then for any $j = 1, ..., p$ and $u > 0$,*

$$P\left(\frac{1}{n} \sum_{i=1}^{n} X_{ij} X_i' \theta^0 - \frac{n-1}{n} \mu_j > 5 \frac{\tau^2}{\sqrt{k}} \left(2\sqrt{\frac{2u}{n}} + \frac{u}{n} \right) \right) \leq exp(-u)$$

Moreover, if X_1 follows $Normal(0, H)$, then τ and k can be dropped.

Lemma 5. *Suppose that assumption 4 and condition 12 are satisfied. Then for arbitrary* $a \in (0,1)$, $q \geq 1, \zeta > 1$ *we have* $\bar{F}_q(\zeta) \geq F_q(\zeta)/2$.

Let us denote $\Omega = \{ \mid \nabla Q(\theta^0) \mid_\infty \leq \frac{\zeta-1}{\zeta+1}\lambda \}$. Then for $A = \frac{1}{n(n-1)} \sum_{i \neq j} \mathbb{I}(Y_j \leq Y_i) X_i$ and for every $j = 1, 2, ..., p$ we get,

$$\nabla_j Q(\theta^0) = \left[\frac{1}{n} \sum_{i=1}^{n} X_{ij} X_i' \theta^0 - \frac{n-1}{n} \mu_j \right] + \frac{n-1}{n} (\mu_j - A^j) - \frac{1}{n^2} \sum_{i=1}^{n} X_{ij} \tag{15}$$

Now we can find the probabilistic bound of each term of equation 15.
Consider the middle term of equation 15. Take $h(z_1, z_2) = \frac{1}{2}[\mathbb{I}(y_2 \leq y_1) x_{1j} + \mathbb{I}(y_1 \leq y_2) x_{2j}]$. Now we can apply Lemma 3. Since X_{1j} and X_{2j} are i.i.d, for arbitrary $L > 0$ we have,

$$\mathbb{E} exp(\mid h(Z_1, Z_2) \mid /L) \leq [\mathbb{E} exp(\mid X_{1j} \mid /(2L))]^2 \tag{16}$$

Since X_{1j} is subgaussian 16 can be bounded by $4exp\left(\frac{\tau^2}{4L^2}\right)$. Putting $L = \tau$ and $u = log(p/a)$ in Lemma 3 we get the universal constant K_1,

$$P\left(A^j - \mu_j > K_1 \tau \sqrt{\frac{log(p/a)}{n}} \right) \leq \frac{a}{p}$$

The bound for the first and third term can be obtained similarly using Lemma 4 and 2 respectively. Combining these results we get, $P(\Omega) \geq 1 - K_2 a$ provided λ satisfies the condition 13.
Let us denote $\tilde{\theta} = \hat{\theta} - \theta^0$, where $\hat{\theta}$ is the minimizer of convex function 4, also we have , for $\hat{\theta}_j \leq 0, \nabla_j Q(\hat{\theta}) = -\lambda sign(\hat{\theta}_j)$ and for $\hat{\theta}_j = 0$, $\mid \nabla_j Q(\hat{\theta}) \mid \leq \lambda$, for any $j = 1, ..., p$.
Note that, $\mid \tilde{\theta} \mid_1 = \mid \tilde{\theta}_T \mid_1 + \mid \tilde{\theta}_{T'} \mid_1$. Now we can get,

$$\begin{aligned}
\tilde{\theta}' X' X \tilde{\theta}/n &= \tilde{\theta}' [\nabla Q(\hat{\theta}) - \nabla Q(\theta^0)] \\
&= \sum_{j \in T} \tilde{\theta}_j \nabla_j Q(\hat{\theta}) + \sum_{j \in T'} \hat{\theta}_j \nabla_j Q(\hat{\theta}) - \tilde{\theta}' \nabla Q(\theta_0) \\
&\leq \lambda \sum_{j \in T} \mid \tilde{\theta}_j \mid - \lambda \sum_{j \in T} \mid \hat{\theta}_j \mid + \mid \tilde{\theta} \mid_1 \mid \nabla Q(\theta^0) \mid_\infty \\
&= (\lambda + \mid \lambda Q(\theta^0) \mid_\infty) \mid \tilde{\theta}_T \mid_1 + (\mid \nabla Q(\theta^0) \mid_\infty - \lambda) \mid \tilde{\theta}_T' \mid_1
\end{aligned}$$

As we are considering the event Ω we get,

$$\mid \tilde{\theta}_T' \mid_1 \leq \frac{\lambda + \mid \nabla Q(\theta^0) \mid_\infty}{\lambda - \mid \nabla Q(\theta^0) \mid_\infty} \mid \tilde{\theta}_T \mid_1 \leq \zeta \mid \tilde{\theta}_T \mid_1$$

Thus we prove $\tilde{\theta} \in C(\zeta)$. Now, from the definition of $\bar{F}_q(\zeta)$ we get,

$$\begin{aligned}
\mid \hat{\theta} - \theta^0 \mid_q &\leq \frac{p_0^{1/q} \mid X' X (\hat{\theta} - \theta^0)/n \mid_\infty}{\bar{F}_q(\zeta)} \\
&\leq p_0^{1/q} \frac{\mid \nabla Q(\hat{\theta}) \mid_\infty + \mid \nabla Q(\theta^0) \mid_\infty}{\bar{F}_q(\zeta)} \\
&\leq \frac{4\zeta p_0^{1/q} \lambda}{(\zeta+1) F_q(\zeta)}
\end{aligned}$$

□

Theorem 2 gives the bound to the estimation error in Rank-Lasso.This result can also be used to get the consistency conditions for the estimates. By replacing a by a sequence a_n, that does not decreases too fast and replacing λ by corresponding sequence λ_n satisfying condition 13 the consistency conditions can be presented. The consistency holds even when number of predictors is significantly larger than sample size. For example, when $p = exp(n^{\alpha_1}), p_0 = n^{\alpha_2}, a = exp(-n^{\alpha_1})$, where $\alpha_1 + 2\alpha_2 < 1$, and λ takes the value exactly equal to the right hand side of the inequality 13 then the consistency in l_∞ norm holds provided $F_\infty(\zeta)$ and k are either bounded below or slowly converge to 0 and τ is bounded above or slowly diverges to ∞.
As a consequence of Theorem 2 the following corollaries can be proven.

Corollary 2.1. *If the conditions of Theorem 2 are satisfied for $q = \infty$, then for $\theta^0_{min} \geq \frac{8\zeta\lambda}{(\zeta+1)F_\infty(\zeta)}$ we have,*

$$P(\forall_{j\in T, k\notin T} \ |\hat{\theta}_j| > \ |\hat{\theta}_k|) \geq 1 - K_3 a$$

where $\theta^0_{min} = \min_{j\in T} \ |\theta^0_j|$

In the above corollary 2.1 it is stated that θ^0_{min} can not be too small. From Theorem 1 we have $\theta^0 = \gamma_\beta \beta$. Hence according to Corollary 2.1 ,

$$\min_{j\in T} \ |\beta_j| \geq \frac{8\zeta\lambda}{\gamma_\beta(\zeta+1)F_\infty(\zeta)}$$

Here the denominator contains $\gamma_\beta = \frac{\frac{n-1}{n}\beta'\mu}{\beta' H\beta}$. This factor is usually smaller than 1, hence large sample size is required for Rank-Lasso to work well.
Now, the following corollary is stated which is a simplified version of Theorem 2.

Corollary 2.2. *Let $a \in (0,1)$ be arbitrary and Assumptions 3 and 4 are satisfied. Suppose that, there exists $\zeta_0 > 1, C_1 > 0$ and $C_2 < \infty$ such that $k \geq C_1, F_\infty(\zeta_0) \geq C_1$ and $\tau \leq C_2$. Then for,*

$$n \geq K_1 p_0^2 log(p/a)$$

$$\lambda \geq K_2 \sqrt{\frac{log(p/a)}{n}}$$

we have ,

$$P(\ |\hat{\theta} - \theta^0|_\infty \leq 4\lambda/C_1) \geq 1 - K_3 a \tag{17}$$

where K_1, K_2 depend only on ζ_0, C_1, C_2 and K_3 is a universal constant as mentioned in Theorem 2.

4 Modifications of RankLASSO

RankLASSO model considered in previous section has a major drawback, that only if the restrictive irrepresentable condition is satified, only then can it recover the true model. If the condition is not satisfied, then RankLASSO can achieve a high power only if it includes a large number of irrelevant predictors. To overcome this limitation, we state the following two theorems which is based on application of weighted and thresholded versions of RankLASSO. We use the initial RankLASSO estimator $\hat{\theta}$ of θ^0, which is a consistent estimator under the assumptions of the previous theorem. We state the following theorems under simplified assumptions made under Corollary 2.2.

4.1 Thresholded RankLASSO

We first consider the thresholded RankLASSO, denoted by $\hat{\theta}^{th}$ which is defined as follows:

$$\hat{\theta}_j^{th} = \hat{\theta}_j \mathbb{I}(|\hat{\theta}_j| > \delta),\ j = 1,2,\ldots,p \tag{18}$$

where $\hat{\theta}$ is the RankLASSO estimator in (3) and $\delta > 0$ is a threshold.

Theorem 3. *Assuming Corollary 2.2 holds and that the sample size and the tuning parameter λ for RankLASSO are selected according to Corollary 2.2. Further, suppose that $\theta^0_{\min} = \min_{j\in T} |\theta^0_j|$ is such that it is possible to select the threshold δ so that*

$$\theta^0_{\min}/2 \geq \delta > K_4\lambda$$

where $K_4 = 4/C1$ is the constant from (17), then the following holds,

$$P(\hat{T}^{th} = T) \geq 1 - K_3 a$$

where K_3 is the universal constant from Theorem 2 and $\hat{T}^{th} = \{1 \leq j \leq p : \hat{\theta}_j^{th} \neq 0\}$ is the estimated set of relevant predictors by thresholded RankLASSO.

Proof. We proceed with the proof as a consequence of the uniform bound (17) from Corollary 2.2. For any $j \in T$, we obtain,

$$|\hat{\theta}_j| = |\hat{\theta}_j - \theta^0_j| \leq K_4\lambda < \delta,$$

so $j \notin \hat{T}^{th}$. Further if $j \in T$, then,

$$|\hat{\theta}_j| \geq |\theta^0_j| - |\hat{\theta}_j - \theta^0_j 0| \geq 2\delta - K_4\lambda > \delta$$

□

The importance of this theorem is highlighted by the fact that the thresholded RankLASSO can potentially identify the support of β under milder regularity conditions. This implies that the sequence of nested models based on the ranking provides by the RankLASSO estimates under these conditions contains the true model.

4.2 Weighted RankLASSO

We now consider the weighted RankLASSO that minimizes the following:

$$Q(\theta) + \lambda_a \sum_{j=1}^{p} w_j |\theta_j| \tag{19}$$

where $\lambda_a > 0$ and the weights are chosen according to the following scheme: for arbitrary number $K > 0$ and the RankLASSO estimator $\hat{\theta}$, from section, we have $w_j = |\hat{\theta}|_j^{-1}$ for $|\hat{\theta}_j| \leq \lambda_a$, and $w_j \leq K$, otherwise. We discuss the properties of weighted RankLASSO estimator in the following theorem:

Theorem 4. *We assume that Corollary 2.2 holds and the sample size and the tuning parameter λ for RankLASSO are selected accordingly as done in Corollary 2.2. Let $\lambda_a = K_4\lambda$, where $K_4 = 4/C1$ is from (17). Additionally, we suppose that the signal strength and sparsity satisfy $\theta^0_{min}/2 > \lambda_a$ and $p_0\lambda \leq K_5$, where K_5 is sufficiently small constant. Then, with a probability of atleast $1 - K_6 a$ there exists a global minimizer $\hat{\theta}^a$ of (19), such that $\hat{\theta}^a_{T'} = 0$ and*

$$|\hat{\theta}^a_T - \theta^0_T|_1 \leq K_7 p_0 \lambda \tag{20}$$

where K_6 and K_7 are constants dependling only on $K_1, \ldots, K_5$ and constant K, that is used in the definition of weights.

Proof. We begin by defining the following function:

$$\Gamma^a(\theta) = Q(\theta) + \lambda_a \sum_{j=1}^{p} w_j |\theta_j| \tag{21}$$

We fix $a \in (0,1)$ and set $\mathscr{E}_0 = 3$ (which is considered for simplicity). We consider the event $\Omega = \{|\nabla Q(\theta^0)|_\infty \leq \lambda/2\}$. We know from Theorem 2 that $P(\Omega) \geq 1 - K_3 a$ which also satisfies the inequality (17). The proof involves two steps.
We first show that with high probability there exists a minimizer of the function,

$$g(\theta_T) = \Gamma^a(\theta_T, 0)$$

that is close to θ^0_T in the ℓ_1 norm. Let that mnimizer be $\hat{\theta}^a_T$. In the second part, we obtain the vector $(\hat{\theta}^a_T, 0)$, that is, $\hat{\theta}^a_T$ augmented by $p - p_0$ zeros, is the minimizer of the function (21).

We consider the vectors $v \in \mathbb{R}^{p_0}$ having fixed common ℓ_1-norm and a sphere

$$\{\theta_T = \theta^0_T + p_0\lambda_v\} \tag{22}$$

Suppose that $|v|_1$ is sufficiently large. We take arbitrary θ_T from the sphere (22) and calculate that

$$Q(\theta_T, 0) - Q(\theta^0) = \frac{1}{2} p_0^2 \lambda^2 v' \frac{1}{n} X_T' X_T v + p_0 \lambda v' [\nabla Q(\theta^0)]_T$$

Let $\hat{\kappa}$ be the minimal eigenvalue of the matrix $\frac{1}{n} X_T' X_T$, then we have $v' X_T' X_T v \geq \hat{\kappa} |v|_1^2 / p_0$. Moreover, for the event Ω, we obtain

$$|v' [\nabla Q(\theta^0)]_T| \leq |v|_1 [\nabla Q(\theta^0)]_T|_\infty \leq \lambda |v|_1 / 2$$

Proceeding in similar lines of previous lemma, we can show that $\hat{\kappa} \geq \kappa/2$ with probability close to one. We therefore obtain,

$$Q(\theta_T, 0) - Q(\theta^0) \geq \kappa p_0 \lambda^2 |v|_1^2 / 4 - p_0 \lambda^2 |v|_1 / 2 \tag{23}$$

Now we focus on the penalty term and obtain the following:

$$\left| \lambda_a \sum_{j=1}^{p_0} w_j [|\theta^0_j + p_0 \lambda v_j| - |\theta^0_j|] \right| \leq \lambda_a p_0 \lambda \sum_{j=1}^{p_0} w_j |v_j| \tag{24}$$

Moreover, for $j \in T$, we have from Corollary 2.2 that

$$|\hat{\theta}_j| \geq |\theta^0_j| - |\hat{\theta}_j - \theta^0_j| \geq \theta^0_{min} - K_4 \lambda > \lambda_a$$

so $w_j \leq K$. Thus the right side of (24) is bounded by $K\lambda\lambda_a p_0|v|_1$. Combining with (23), we get

$$g(\theta_T) - g(\theta_T^0) \geq p_0\lambda^2|v|_1(\kappa|v|_1/4 - 1/2 - K_4K) \tag{25}$$

Clearly, the right hand side of (25) is bounded, because the norm $|v|_1$ can be taken sufficiently large, K, K_4 are constants and κ is lower bounded by a constant. Thus, the convex function $g(\theta_T)$ takes on a sphere (22) values larger than the in the center θ_T^0. So, there exists a minimizer inside the sphere.

We next show that the random vector $(\hat{\theta}_T^a, 0)$ minimizes (21)with high probability. Thus we have to prove that the event

$$\{|\nabla_j Q(\hat{\theta}_T^a, 0)| \leq w_j\lambda_a \text{ for every } j \notin T\} \tag{26}$$

has probability close to one. by Corollary 2.2we have for $j \notin T$,

$$|\hat{\theta}_j| = |\hat{\theta}_j - \theta_j^0| \leq K_4\lambda$$

Therefore, $w_j \geq \lambda_a^{-1}$. Further, we can calculate that

$$\nabla Q(\theta_T, 0) = \frac{1}{n}X_T'X_T\theta_T - \left[\frac{n-1}{n}A + \frac{1}{n^2}\sum_{i=1}^{n}X_i\right]$$

So we obtain the inequality

$$\left|\left[\nabla Q(\hat{\theta}_T^a, 0\right]_{T'}\right|_\infty \leq \left|\frac{1}{n}X_{T'}'X_T(\hat{\theta}_T^a - \theta_T^0)\right|_\infty + |[\nabla Q(\theta^0)]_{T'}|_\infty \tag{27}$$

Consider the event $\Sigma = \{|\nabla Q(\theta^0)|_\infty \leq \lambda/2\}$ that has probability close to 1 by proof of Theorem 14. The second term on the right-hand side of (27) can be bounded by $\lambda/2$. The former one can be decomposed as

$$\begin{aligned}\left|\frac{1}{n}X_{T'}'X + T(\hat{\theta}_T^a - \theta_T^0)\right|_\infty &\leq \left|\left(\frac{1}{n}X_{T'}'X_T - H_2'\right)(\hat{\theta}_T^a - \theta_T^0)\right|_\infty + |H_2'(\hat{\theta}_T^a - \theta_T^a)|_\infty \\ &\leq \left|\frac{1}{n}X_{T'}'X_T - H_2'\right|_\infty |\hat{\theta}_T^a - \theta_T^0|_1 + |H_2'|_\infty|\hat{\theta}_T^a - \theta_T^0|_1 \end{aligned} \tag{28}$$

The expression $|H_2|_\infty$ is bounded by 1, so from the first part of the proof, we can bound, with high probability, the second term in (28) by $K_6p_0\lambda$. The ℓ_∞-normin the former expression can be bounded with probability close to one, by $K_7\sqrt{\frac{\log(p/a)}{n}}$ as in proof of a previous lemma. Therefore, we have just proved that with probability close to one,

$$\left|\left[\nabla Q(\hat{\theta}_T^a, 0)\right]_{T'}\right|_\infty \leq K_8p_0\lambda$$

Combining it with the fact that $w_j \geq \lambda_a^{-1}$ we obtain that the event (26) has high probability close to one, since from assumptions of the theorem, $p_\lambda \leq K_5$ for K_5 small enough. □

5 Data Analysis

5.1 Simulations

In this section we present results of our simulation study illustrating the properties of RankLasso and its variants in variable selection.

We consider the setup, where the number of explanatory variables p increases with n according to the formula $p = 0.01n^2$. We consider the following pairs $(n,p) : (100,100), (200,400), (300,900), (400,1600)$. For each of these combinations we assume that the number of true variables in the model is given by p_0, where $p_0 = \#\{j : \beta_j \neq 0\} \in \{3,10,20\}$.

In three of the simulation scenarios the rows of the design matrix are generated as random vectors from the multivariate normal distribution with the covariance matrix Σ defined as follows

- for the independent case $\Sigma = I$,
- for the correlated case $\Sigma_{ii} = 1$ and $\Sigma_{ij} = 0.3$ for $i \neq j$.

In one of the scenarios the design matrix is created by simulating the genotypes of p independent Single Nucleotide Polymorphisms (SNPs). In this case the explanatory variables can take only three values: 0 for the homozygote for the minor allele (genotype $\{a,a\}$), 1 for the heterozygote (genotype $\{a,A\}$) and 2 for the homozygote for the major allele (genotype $\{A,A\}$). The frequencies of the minor allele for each SNP are independently drawn from the uniform distribution on the interval $(0.1,0.5)$. Then, given the frequency π_j for j-th SNP, the explanatory variable X_{ij} has the distribution: $P\left(X_{ij} = 0\right) = \pi_j^2$, $P\left(X_{ij} = 1\right) = 2\pi_j\left(1-\pi_j\right)$ and $P\left(X_{ij} = 2\right) = \left(1-\pi_j\right)^2$.

The full description of the simulation scenarios is provided below:

- Scenario 1

$$Y = X\beta + \varepsilon,$$

where X matrix is generated according to the independent case, $\beta_1 = \ldots = \beta_{p_0} = 3$ and the elements of $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n)$ are independently drawn from the standard Cauchy distribution,

- Scenario 2 - the regression model, values of regression coefficients and ε are as in Scenario 1, design matrix contains standardized versions of genotypes of p independent SNPs,

- Scenario 3 - the regression model, values of regression coefficients and ε are as in Scenario 1 and the design matrix X is generated according to the corvelated case,

- Scenario 4 - the design matrix X is generated according to the correlated case and the relationship between Y_i and $\beta' X_i$ is non-linear:

$$Y_i = \exp\left(1 + 0.05\beta^{'} X_i\right) + \varepsilon_i$$

and $\varepsilon_1, \ldots, \varepsilon_n$ are independent random variables from the standard Cauchy distribution.

In our simulation study we compare five different statistical methods:

- rL: RankLasso defined in (3) with $\lambda := \lambda_{rL}$ and

$$\lambda_{rL} = 0.3\sqrt{\frac{\log p}{n}}$$

- arL: adaptive RankLasso (19), with $\lambda_\alpha = 2\lambda_{rL}$ and weights

$$w_j = \begin{cases} \frac{0.1\lambda_{rL}}{\left|\hat{\theta}_j\right|} & \text{when} \quad \left|\hat{\theta}_j\right| > 0.1\lambda_{rL}, \\ \left|\hat{\theta}_j\right|^{-1} & \text{otherwise,} \end{cases}$$

where $\hat{\theta}$ is the RankLasso estimator computed above. If $\hat{\theta}_j = 0$, then $\left|\hat{\theta}_j\right|^{-1} = \infty$ and j^{th} explanatory variable is removed from the list of predictors before running weighted RankLasso,
- thrL: thresholded RankLasso (18), where the tuning parameter for RankLasso is selected by cross-validation and the threshold is selected in such a way that the number of selected predictors coincides with the number of predictors selected by adaptive RankLasso,
- cv: regular Lasso with the tuning parameter selected by cross-validation. The values of the tuning parameters for RankLasso and LADLasso were selected empirically so that both methods perform comparatively well for $p_0 = 3$ and $n = 200, p = 400$.

We compare the quality of the above methods by performing 200 replicates of the experiment, where in each replicate we generate the new realization of the design matrix X and the vector of random noise ε. We calculate the NMP: the average value of Numbers of Misclassified Predictors, i.e false positives plus false negatives.

Figure 1 illustrates the average number of falsely classified predictors for different methods and under different simulation scenarios. In the case when predictors are independent, RankLasso satisfies assumptions of Wang and Zhu (2015) and its NMP decreases with $p = 0.01n^2$. We can also observe that for independent predictors, the adaptive and thresholded versions perform similarly to the standard version of RankLasso. As expected, regular cross-validated Lasso performs very badly, when the error terms come from the Cauchy distribution. Also, it is interesting to observe that the first two rows in Figure 1 do not differ significantly, which shows that the performance of RankLasso for the realistic independent SNP data is very similar to its performance when the elements of the design matrix are drawn from the Gaussian distribution.

The behaviour of RankLasso changes significantly in the case when predictors are correlated. Namely, NMP of RankLasso increases with p. On the other hand, NMP of both adaptive and thresholded versions of RankLasso decrease with p, so these two methods are able to find the true model consistently. As shown in Figure 1, in the case of correlated predictors thresholded RankLasso is systematically better than adaptive RankLasso, even though both methods always select the same number of predictors.

6 Supplementary Material

The interested reader is directed to https://github.com/ArkaB-DS/rankLASSO which contains all the figures present here in the directory `images` and the corresponding codes to generate them in the `R` directory.

7 Acknowledgements

We take this opportunity to heartily thank our supervisor Prof. Subhra Sankar Dhar for his valuable feedback and constant guidance on this project.

References

Bickel, P. J., Ritov, Y., and Tsybakov, A. B. (2009). Simultaneous analysis of lasso and dantzig selector. *The Annals of statistics*, 37(4):1705–1732.

Bogdan, M., Ghosh, J. K., and Zak-Szatkowska, M. (2008). Selecting explanatory variables with the modified version of the bayesian information criterion. *Quality and Reliability Engineering International*, 24(6):627–641.

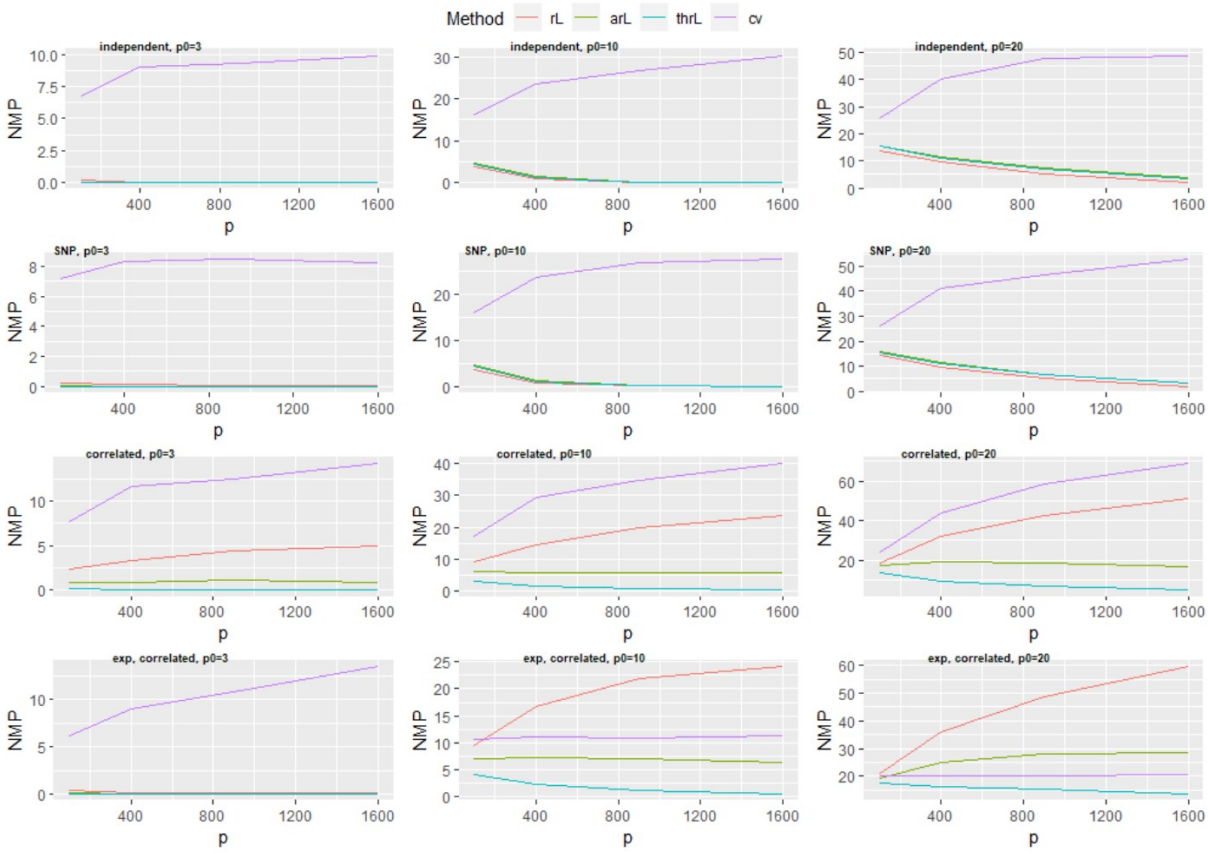


Figure 1: Plots of NMP (average number of misclassified predictors) as the function of p.

Friedman, J., Hastie, T., Tibshirani, R., Narasimhan, B., Tay, K., Simon, N., and Qian, J. (2021). Package 'glmnet'. *Journal of Statistical Software. 2010a*, 33(1).

Hastie, T., Tibshirani, R., Friedman, J. H., and Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction*, volume 2. Springer.

Rejchel, W. and Bogdan, M. (2020). Rank-based lasso-efficient methods for high-dimensional robust model selection. *Journal of Machine Learning Research*, 21(244):1–47.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Van de Geer, S. A. (2008). High-dimensional generalized linear models and the lasso. *The Annals of Statistics*, 36(2):614–645.

Wang, T. and Zhu, L. (2015). A distribution-based lasso for a general single-index model. *Science China Mathematics*, 58(1):109–130.

Ye, F. and Zhang, C.-H. (2010). Rate minimaxity of the lasso and dantzig selector for the lq loss in lr balls. *The Journal of Machine Learning Research*, 11:3519–3540.

Zak, M., Baierl, A., Bogdan, M., and Futschik, A. (2007). Locating multiple interacting quantitative trait loci using rank-based model selection. *Genetics*, 176(3):1845–1854.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association*, 101(476):1418–1429.