Recursive bootstrap L_1-type regularized regression modeling based on parametric statistical test

Heewon Park*, Seiya Imoto and Satoru Miyano
Human Genome Center, Institute of Medical Science, The University of Tokyo, Tokyo, Japan
hwpark@ims.u-tokyo.ac.jp

Abstract

The L_1-type regularization approaches have been widely used for uncovering cancer driver gene based on various genome-scale information. Although the existing L_1-type regularization methods have been widely used to various fields of research, there are several drawbacks as a tool for feature selection in high dimensional data analysis: limitation of subset size, erroneous estimation result, multicollinearity problem and time consuming procedures. We propose a novel statistical strategy, called a Recursive Random Lasso (RRLasso) for high dimensional data analysis in line with a random lasso. In order to time effective analysis, we consider recursive bootstrap procedure based on random forest method. Furthermore, we introduce a parametric statistical test for variable selection based on bootstrap regression modeling results. We can see through Monte Carlo simulations that the proposed RRLasso not only provides time effective performances but also performs well for high dimensional data analysis.

Keywords: bootstrap method; L_1-type regularization; parametric statistical test; random forest method.

1. Introduction

A crucial issue of cancer research is to identify cancer driver genes based on various genomic data analysis (e.g., expression levels, copy number variations, methylation, etc.), since efficiently identified anti-cancer drug target plays a key role in cancer therapy. Although various L_1-type regularization approaches, e.g., lasso (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005), etc., have been widely used to identify cancer driver gene, there are several drawbacks as a tool for feature selection based on high dimensional data analysis (i.e., lasso and adaptive lasso: selecting features at most sample size n, adaptive L_1-type regularization method: multicollinearity problem, elastic net: unbiased estimation result for coefficients of highly correlated variables with different magnitudes).

To settle on the issues, Wang et al. (2011) proposed a random lasso based on bootstrap regression modeling with random forest method. Although the random lasso overcomes the drawbacks of existing L_1-type regularization methods, the method is computationally intensive, due to two step bootstrap procedures. Furthermore, Wang et al. (2011) performed final feature selection based on an arbitrarily decided threshold, even though the variable selection result heavily depends on a threshold.

We propose a novel statistical strategy for high dimensional regression modeling in line with the random lasso. We introduce recursive bootstrap approaches to measure significance of predictor variables and estimate regression coefficients. We also propose a novel threshold based on a parametric statistical test to effectively perform for feature selection. By using recursive bootstrap procedure, we perform time effective bootstrap regression modeling for high dimensional data analysis without loss of efficiency of modeling accuracy. Furthermore, the proposed parametric statistical test based on recursive bootstrap results leads to effective variable selection results without false positive of selected variables.

We demonstrate through Monte Carlo simulations with various scenarios the effectiveness of the proposed recursive random lasso and elastic net.

The rest of this paper is organized as follows. In Section 2, we introduce the existing L_1-type regularization approaches, and point out drawbacks of the existing methods. Section 3 presents the existing random lasso. We then propose a recursive random lasso and elastic net in Section 4. Monte Carlo simulations are conducted to examine the effectiveness of the proposed statistical strategy in Section 5. Some conclusions are given in Section 6.

2. L_1-type regularization methods

Suppose we have n independent observations $\{(y_i, \boldsymbol{x}_i); i = 1, ..., n\}$, where y_i are random response variables and $\boldsymbol{x}_i$ are p-dimensional vectors of the predictor variables. Consider the linear regression model,

$$y_i = \boldsymbol{x}_i^T \boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, ..., n, \tag{1}$$

where $\boldsymbol{\beta}$ is an unknown p-dimensional vector of regression coefficients and ε_i are the random errors which are assumed to be independently and identically distributed with mean 0 and variance σ^2. We assume that the y_i are centered and x_{ij} are standardized by their mean and standard deviation: $\sum_i^n y_i/n = 0$, $\sum_i^n x_{ij}/n = 0$ and $\sum_i^n x_{ij}^2/n = 1$, thus an intercept term is excluded from the regression model in (**??**). For the regression modeling, a great deal of studies is being carried out, especially high dimensional data analysis.

Tibshirani (1996) proposed the lasso, which minimizes residual sum of squares subject to the constraint $\lambda \sum_{j=1}^p |\beta_j|$, and its solution is given by

$$\hat{\boldsymbol{\beta}}^{\text{Lasso}} = \arg\min_{\boldsymbol{\beta}} \{\sum_{i=1}^n (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \lambda \sum_{j=1}^p |\beta_j|\}, \tag{2}$$

where λ is a tuning parameter controlling model complexity. By imposing the penalty term as sum of absolute values of regression coefficients, the lasso can perform simultaneous parameter estimation and variable selection. A recent work, however, suggested that the lasso may suffer from the following limitations (Zou and Hastie, 2005):

- In $p > n$ situation, the lasso can select at most n variables, because of the convex optimization problem.
- The lasso cannot take account of a grouping effect of predictor variables, and thus tends to select only one variable from a group.

It implies that the lasso cannot perform well for high dimensional data analysis.

To overcome the drawbacks, various L_1-type regularization methods have been proposed. The elastic net (Zou and Hastie, 2005) especially has been draw a large amount of attention in various fields of research,

$$\hat{\boldsymbol{\beta}}^{\text{Elastic net}} = \arg\min_{\boldsymbol{\beta}} \{\sum_{i=1}^n (y_i - \boldsymbol{x}_i^T \boldsymbol{\beta})^2 + \lambda_1 \sum_{j=1}^p |\beta_j| + \lambda_2 \sum_{j=1}^p \beta_j^2\}. \tag{3}$$

By imposing the additional L_2-penalty, called as a ridge (Hoerl and Kennard, 1970), to lasso, the elastic net performs effectively feature selection in high dimensional data analysis, i.e., there is no limitation of subset size. Furthermore, the elastic net can enjoy the following grouping effect,

$$D_{\lambda_1,\lambda_2}(j,k) = \frac{1}{|\boldsymbol{y}|_1} |\hat{\beta}_j(\lambda_1,\lambda_2) - \hat{\beta}_k(\lambda_1,\lambda_2)| \le \frac{1}{\lambda_2}\sqrt{2(1-\rho)}, \tag{4}$$

where $\rho = \boldsymbol{x}_j^T \boldsymbol{x}_k$ is sample correlation (Zou and Hastie, 2005).

Although the elastic net performs well for high dimensional data analysis, Wang et al (2011) demonstrated the demerit of elastic net as follows,

- The property of "grouping effect" leads to erroneous estimation results when coefficients of highly correlated variables with different magnitudes, especially with different sign. However, coefficients of highly correlated variables with different magnitudes can be easily observed in various fields of research (e.g., genes in common biological pathway are usually highly correlated, and their regression coefficient can be given as different magnitudes or different sign).

The adaptive L_1-type penalties have been also proposed and widely used for regression modeling,

- adaptive lasso:

$$P_\lambda^{\text{ad.Lasso}}(|\boldsymbol{\beta}|) = \lambda \sum_{j=1}^p w_j |\beta_j|\}, \tag{5}$$

- adaptive elastic net:

$$P_{\lambda}^{\text{ad.Elastic net}}(|\boldsymbol{\beta}|) = \lambda\{(1-\alpha)\sum_{j=1}^{p} w_j|\beta_j| + \alpha\sum_{j=1}^{p}\beta_j^2\}, \tag{6}$$

where $w_j = 1/|\hat{\beta}_j^{OLS}|^{\gamma}$ is an adaptive data driven weight for $\gamma > 0$. By using the weight, we can discriminately impose a penalty to each features depending on significance of features, and thus perform effectively features selection. However, the performances of adaptive regularization methods heavily depend on the OLS estimator, and thus suffer from multicollinearity. Furthermore, the adaptive L_1-type regularization methods also suffer from the drawbacks of ordinary lasso (i.e., selected subset size) and elastic net (i.e., biased estimation results).

3. Random lasso

Wang et al. (2011) focused on the drawbacks of the existing L_1-type approaches, and proposed a random lasso based on bootstrap strategy with the random forest method. In random lasso procedure, randomly selected q variables are considered as candidate variables in regression modeling for each bootstrap sample. Thus, the results do not suffer from the drawbacks of highly correlated variables, since each bootstrap sample may include only subset of the highly correlated variables. Furthermore, the random lasso can overcome the limitation of subset size, since variable selection is based on bootstrap regression modeling results with randomly selected q variables.

Wang et al. (2011) proposed an algorithm to implement the random lasso based on two-step bootstrap procedures.

ALGORITHM 1 *Random lasso*

- Step 1: Generating importance measure for predictor variables.
  1. Draw B bootstrap samples with size n by sampling with replacement from the original dataset.
  2. For the b_1^{th} bootstrap sample, $b_1 \in \{1, 2, ..., B\}$, q_1 candidate variables are randomly selected and lasso is applied for regression modeling and we obtain estimators $\hat{\beta}_j^{(b_1)}$ for $j = 1, ..., p$.
  3. The importance measure of x_j is calculated as $I_j = |B^{-1}\sum_{b_1=1}^{B}\hat{\beta}_j^{(b_1)}|$.
- Step 2: Variable selection
  1. Draw B bootstrap samples with size n by sampling with replacement from original dataset.
  2. For the b_2^{th} bootstrap sample, $b_2 \in \{1, 2, ..., B\}$, q_2 candidate variables are randomly selected with selection probability of x_j proportional to I_j, and adaptive lasso is applied for regression modeling and we obtain estimator $\hat{\beta}_j^{(b_2)}$ for $j = 1, ..., p$.
  3. Compute final estimator $\hat{\beta}_j$ as $\hat{\beta}_j = B^{-1}\sum_{b_2=1}^{B}\hat{\beta}_j^{(b_2)}$ for $j = 1, ..., p$.

For noise predictor variables, the coefficients in respective bootstrap samples are estimated in small or even have different sign, thus the absolute value of averaging coefficients (i.e., I_j) will be small or close to zero. On the other hand, the coefficients of crucial predictor variables may consistently large in difference bootstrap samples, thus a crucial gene has a large value of $|I_j|$. It implies that the selection probability I_j properly operates to effective features selection. Wang et al. (2011) considered q_1 and q_2 as tuning parameters, and the importance measure I_j is also used as the weight for adaptive lasso in their study.

Wang et al. (2011) pointed out that the variable selection result of the random lasso is little unfair, since some of final non-zero coefficients may results from only any particular bootstrap sample. And, they considered a threshold $t_n = 1/n$ for variable selection, and predictor variables with $|\hat{\beta}_j| \leq t_n$ were deleted in final model.

4. Recursive random lasso (RRLasso)

The random lasso can overcome the drawbacks of existing L_1-type regularization by using random forest method with bootstrap regression modeling. Although the random lasso performs well for high dimensional regression modeling with highly correlated predictors, the method also suffers the following drawbacks,

- The random lasso is computationally intensive method, due to two bootstrap procedures with respective B replications. The computational complexity is significantly increased in high dimensional data analysis.
- There are too many tuning parameters, i.e., λ, α and γ, q_1 and q_2. The large number of tuning parameters also leads to time consuming, since the random lasso procedures should be repeatedly implemented to select the optimal combination of the parameters.
- Wang et al. (2011) arbitrarily selected the threshold without any statistical background.

We propose a recursive bootstrap procedures for generating the importance measure and regression coefficients estimation. We also propose a novel threshold to effectively select predictor variables based on parametric statistical test. Furthermore, a number of candidate predictors q in each bootstrap sample is also randomly selected in our strategy (i.e., we consider q as not a tuning parameter).

ALGORITHM 2 *Recursive random lasso (or elastic net)*

1. Draw B bootstrap samples with size n by sampling with replacement from the original dataset.
2. For the first bootstrap sample (i.e., $b = 1$), q candidate variables are randomly selected and lasso (or elastic net) is applied to regression modeling. We then obtain estimators $\hat{\beta}_j^{(1)}$ for $j = 1, ..., p$.
3. For the $b \in \{2, ..., B\}$, the importance measure of x_j is calculated as $I_j = |(b-1)^{-1} \sum_{b=1}^{b-1} \hat{\beta}_j^{(b)}|$. The q candidate variables are randomly selected with selection probability I_j, and adaptive lasso (or adaptive elastic net) with $w_j = 1/|I_j|$ is applied to regression modeling. We obtain estimators $\hat{\beta}_j^{(b)}$ for $j = 1, ..., p$.
4. Final estimators are computed as $\hat{\beta}_j = B^{-1} \sum_{b=1}^{B} \hat{\beta}_j^{(b)}$.
5. We finally perform variable selection based on a threshold t^* and a parametric test.

A novel threshold: In order to effectively select crucial variables, we proposed parametric statistical test based on the above bootstrap regression modeling results. We first consider a $B \times p$ binary matrix $\mathbf{D}$ obtained from the above recursive bootstrap procedures. We set an element of the binary matrix as $D_{bj} = 1$ for a non-zero $\hat{\beta}_j$ in b^{th} bootstrap sample; otherwise $D_{bj} = 0$. In other word, we consider that the binary matrix is obtained from Bernoulli experiments, and let $\boldsymbol{D}_j$ be a random variable associated with Bernoulli trials as follows,

$$D_{bj}(\hat{\beta}_j^b \neq 0) = 1 \quad \text{and} \quad D_{bj}(\hat{\beta}_j^b = 0) = 0. \tag{7}$$

The Bernoulli random variable has the following probability density function,

$$f(d_j) = \pi^{d_j}(1-\pi)^{1-d_j}, \quad d_j = 0, 1, \tag{8}$$

where the probability π can be estimated as follows,

$$\hat{\pi} = \frac{1}{p \times B} \sum_{j=1}^{p} \sum_{b=1}^{B} D_{bj}, \tag{9}$$

which indicates an average of selection ratio of predictor variables. To reasonable variable selection, we then consider a statistics,

$$C_j = \sum_{b=1}^{B} D_{bj}, \quad j = 1, ..., p, \tag{10}$$

which indicates the number of non-zero $\hat{\beta}_j^{(b)}$ in B Bernoulli trials (i.e. B bootstrap samples). Since the Bernoulli trials are independent and the probability of $\hat{\beta}_j^{(b)} \neq 0$ and $\hat{\beta}_j^{(b)} = 0$ on each trials are, π and $1 - \pi$, respectively, the statistics C_j follows the Binomial distribution $b(B, \hat{\pi})$,

$$f(c) = \frac{B!}{c!(B-c)!} \hat{\pi}^c (1-\hat{\pi})^{B-c}, \quad c = 0, 1, ..., B. \tag{11}$$

We then calculate a p-value for each predictor variable as follows,

$$\begin{aligned} p\text{-value}_j &= p(c \geq C_j|\hat{\pi}) \\ &= \sum_{c=C_j}^{B} \frac{B!}{c!(B-c)!}\hat{\pi}^c(1-\hat{\pi})^{B-c}, \end{aligned} \tag{12}$$

and finally perform variable selection based on the p-value with a threshold $t^* = 0.05$ as follows,

$$\hat{\beta}_j^* = \hat{\beta}_j I(p\text{-value}_j < 0.05), \tag{13}$$

where $I(\cdot)$ is an indicator function. By using the parametric statistical test, we can effectively perform variable selection in bootstrap regression modeling without false positive results.

5. Simulation studies

Monte Carlo simulations are conducted to investigate the effectiveness of the proposed modeling strategy. We simulate 100 datasets from the following linear regression model,

$$y_i = \boldsymbol{x}_i^T\boldsymbol{\beta} + \varepsilon_i, \qquad i = 1, ..., n, \tag{14}$$

where ε_i are $N(0, \sigma^2)$, and the correlation between x_l and x_m is $0.5^{|l-m|}$.
We consider the following situations for simulation study,

- Type1: $n = 100$ and $p = 1000$ as $\beta_j = 3$ for randomly selected 50 variables, otherwise $\beta_j = 0$,
- Type2: $n = 100$ and $p = 1000$ as $\beta_j = 3$ for randomly selected 25 variables, $\beta_j = -3$ for randomly selected 25 variables, otherwise $\beta_j = 0$,
- Type3: $n = 100$ and $p = 1000$ as $\beta_j = 3$ for randomly selected 150 variables, otherwise $\beta_j = 0$.
- Type4: $n = 100$ and $p = 1000$ as $\beta_j = 3$ for randomly selected 75 variables, $\beta_j = -3$ for randomly selected 75 variables, otherwise $\beta_j = 0$,

We consider a larger number of crucial predictor variables than sample size (i.e, $n = 100$ and 150 truly non-zero coefficients) in Types 3 and 4. In order to evaluate the proposed recursive random lasso and elastic net, we compare performances of the lasso, adaptive lasso, elastic net and existing random lasso. In numerical studies, we use the ridge estimator for weight in the existing adaptive lasso, and consider the threshold of the existing random lasso as s/n, and select s based on mean squares error in validation dataset. The tuning parameters are selected by 5-fold cross validation based on training dataset.
We compare regression modeling results based on prediction accuracy (i.e., P.error) in Table 1. Mean squared error is given as prediction error and average of T.P and T.N (i.e., T.P&T.N) indicates average of true positive rate (i.e., average number of the true non-zero coefficients, incorrectly set to zero) and true negative

Table 1: Simulation result: prediction error and variable selection result in regression modeling

		EFF.EL	EFF.LA	RD.LA	AD.LA	ELA	LASSO
P.error	Type1	**20.01**	20.12	20.14	21.17	20.35	21.12
	Type2	**19.37**	19.49	19.77	20.62	19.80	20.51
	Type3	**37.84**	38.31	38.42	41.00	39.18	40.95
	Type4	**34.28**	34.69	34.67	36.44	35.56	36.38
T.P&T.N	Type1	**0.70**	**0.70**	0.64	0.66	0.59	0.59
	Type2	**0.69**	**0.69**	0.63	0.65	0.57	0.58
	Type3	**0.59**	**0.59**	0.55	0.56	0.52	0.52
	Type4	**0.58**	0.57	0.54	0.55	0.52	0.52

rate (i.e., the average percentage of true zero coefficients, that were correctly set to zero) as variable selection results. From Table 1, we can see that the proposed recursive random elastic net and lasso show outstanding performances for variable selection and prediction accuracy in all simulation situations. It can be also seen that the lasso and adaptive lasso show poor variable selection results, since the methods cannot perform well for feature selection in high dimensional data analysis. In short, the proposed recursive random lasso and elastic net show not only computationally effective performances but also outstanding regression modeling results (i.e., prediction accuracy and variable selection results).

6. Conclusions

We have proposed a novel statistical strategy based on recursive bootstrap approach and parametric statistical test. To effectively perform high dimensional data analysis, we have considered recursive bootstrap strategies in line with the random lasso. Furthermore, we have proposed a parametric statistical test to variable selection based on bootstrap regression results.

We can see through the numerical studies that the proposed methods show outstanding performances for variable selection and prediction accuracies. Furthermore, our methods showed time effective performances compared with existing random lasso. We can expect that our methods based on recursive bootstrap regression modeling and parametric statistical test will be a useful tool for high dimensional genomic data analysis.

Further work remains to be done for real world example to show effectiveness of our method in real research fields.

References

Hoerl, A. E., Kennard, R .W. (1970). Ridge regression: biased estimation for nonorthogonal problems. *Techonometrics*, 12:55-67.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B*, 73:273-282.

Zou, H., Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B*, 67:301-320.

Wang, S., Nan, B., Rosset, S., Zhu, J. (2011). Random lasso. *The Annals of Applied Statistics*, 5:468-485.

Zou, H., Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of Statistics* 37(4):1733-1751.