Biostatistics (2019), **0**, 0, *pp.* 1–10
doi:10.1093/biostatistics/output

Report: Network-constrained regularization and variable selection for analysis of genomic data

Yang Yang

Department of Statistics, Kansas State University, Manhattan, KS, US

sheepyang@ksu.edu

Summary

The Lasso algorithm is proposed as a variable selection method for sparse linear regression problems with a larger number of predictors p which exceeds the number of observation n. The excellent ability in prediction performance makes LASSO widely applied in many computational biology areas. With the evolution of computing skills, high-dimensional data analysis arises from various scientific fields which encourages the development of LASSO variates. Graph theory and networks analysis are commonly used to model pairwise relationships between interested objects. As more biological networks depicting information of data are identified and documented in databases, networks could provide a useful supplement for prediction. In this context, network-constrained regularization is proposed in order to incorporate the prior information of the data.

Key words:

*

1. Introduction

First introduced in the context of least squares by Tibshirani (1996), the Lasso (least absolute shrinkage and selection operator) improves the prediction accuracy by estimating linear regression coefficient through L_1-constrained least square. As a variable selection method, the Lasso is potent for sparse problems with a larger number of predictors p which exceeds the number of observation n. In cases where observation n is larger than p, even the lasso criterion is not strictly convex, but the solution is still unique regardless of numbers of n and p if the variables are from a continuous probability distribution (tibshirani2013lasso) leading to a wide range of applications of LASSO.

In LASSO algorithm, variable selection is achieved effectively by the constraint in which the sum of the absolute value of the coefficients is less than a fixed value, which is the tuning parameter, to control the amount of shrinkage toward zero in the coefficients. While in another commonly used algorithm, the ridge regression also shrinks the size of the coefficients through the constraint where the sum of the squares of the coefficients is less than a fixed value. However, coefficients are not shrink to zero in ridge regression, so ridge regression is not able to select variables. By contrast, LASSO has shown outstanding performance in both variable selection and regularization.

For the last decade, high-dimensional data from almost all scientific fields encourage the application of LASSO penalized estimation. The excellent ability in prediction performance makes LASSO popular in many computational biology areas, such as genome-wide association analysis. But the various properties of data from different disciplines highlights the need to develop LASSO to remedy certain limitations. A variety of lasso-derived methods are proposed, which include but not limit to Adaptive LASSO (Zou (2006)), smoothly clipped absolute deviation (SCAD) penalty (Fan and Li (2001)), MCP (minimax concave penalty) (zhang2010nearly), Fused lasso (Tibshirani *and others* (2005)), Group lasso (Yuan and Lin (2006)), and Elastic net (Zou and Hastie (2005)).

Graph theory and networks analysis are commonly used to model pairwise relations between interested objects. As more biological networks depicting information of data are identified and documented in databases, networks function as a useful supplement for prediction. In this context, network-constrained regularization is proposed by Li and Li (2008) in order to incorporate the prior information of the data.

2. Theory

Tibshirani (1996) proposed LASSO by minimizing the residual sum of squares which subjects to a constraint on the sum of absolute values of the regression coefficients. This is equivalent to minimize the sum of squares of residuals plus an L_1 penalty on the regression coefficients.

Considering a linear regression model, $\mathbf{y} = (y_1, \ldots, y_n)^T$ is the $n \times 1$ vector of response with n observations and p predictors x_{ij}, $i = 1, \ldots, n$, $j = 1, \ldots, p$. The response y is predicted by

$$\hat{\mathbf{y}} = \hat{\beta}_0 + \mathbf{x}_1\hat{\beta}_1 + \cdots + \mathbf{x}_p\hat{\beta}_p \tag{2.1}$$

where $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \ldots, \hat{\beta}_p)^T$ is the vector of predicted coefficient. LASSO finds the solution by

$$\hat{\boldsymbol{\beta}} = arg\min\{|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2\}, \quad subject\, to \sum_{j=1} |\beta_j| \leqslant t \tag{2.2}$$

which is equivalent to the optimization problem

$$L(\lambda, \boldsymbol{\beta}) = |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda|\boldsymbol{\beta}|_1, \qquad |\boldsymbol{\beta}|_1 = \sum_{j=1}^{p} |\beta_j| \tag{2.3}$$

and

$$\hat{\boldsymbol{\beta}} = arg\min_{\boldsymbol{\beta}}\{L(\lambda, \boldsymbol{\beta})\} \tag{2.4}$$

2.1 *Ridge regression*

Ridge regression performs $L2$ norm regularization by add a penalty on the square of the magnitude of coefficients

$$L(\lambda, \boldsymbol{\beta}) = |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda|\boldsymbol{\beta}|^2, \qquad |\boldsymbol{\beta}|^2 = \sum_{j=1}^{p} \beta_j^2 \tag{2.5}$$

which could be written as

$$L(\lambda, \boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \mathbf{x_i}\boldsymbol{\beta})^{\mathbf{2}} + \lambda \sum_{\mathbf{j=1}}^{\mathbf{p}} \beta_{\mathbf{j}}^{\mathbf{2}} \tag{2.6}$$

and can be rewritten as

$$L(\lambda, \boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \mathbf{x_i}\boldsymbol{\beta})^{\mathbf{2}} + \sum_{\mathbf{j=1}}^{\mathbf{p}} (\mathbf{0} - \sqrt{\lambda}\beta_{\mathbf{j}})^{\mathbf{2}} \tag{2.7}$$

Then the ridge regression can be transformed to a least square regression problem using an augmented data.

$$\mathbf{X}^*_{\mathbf{(n+p)\times p}} = \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I_p} \end{pmatrix}, \quad \mathbf{y}^*_{\mathbf{n+p}} = \begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix} \tag{2.8}$$

So the least square estimate from the augmented data is

$$\begin{aligned} &(\mathbf{X}_*^{\mathbf{T}}\mathbf{X}_*)^{-\mathbf{1}}\mathbf{X}_*^{\mathbf{T}}\mathbf{Y}_* \\ =&\left(\begin{pmatrix} \mathbf{X^T} & \sqrt{\lambda}\mathbf{I_p} \end{pmatrix} \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda}\mathbf{I_p} \end{pmatrix}\right)^{-1} \begin{pmatrix} \mathbf{X^T} & \sqrt{\lambda}\mathbf{I_p} \end{pmatrix} \begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix} \\ =&(\mathbf{X^T}\mathbf{X} + \lambda\mathbf{I_p})^{-\mathbf{1}}\mathbf{X^T}\mathbf{y} \end{aligned} \tag{2.9}$$

2.2 *Elastic net*

The elastic net is a penalized least square method and combines both $L1$ norm and $L2$ norm penalty, thus having the characteristics of both LASSO and ridge regression.

Assume the data set with n observations and p predictors, has the response vector $\mathbf{y} = (y_1, \ldots, y_n)^T$, and design matrix $\mathbf{X}$ with $\mathbf{x}_j = (x_{1j}, \ldots, y_{nj})^T$, $j = 1, \ldots, p$.

Then the design matrix is standardized and the response vector is centered to satisfy

$$\sum_{i=1}^{n} y_i = 0,\ \sum_{i=1}^{n} x_{ij} = 0\ and\ \sum_{i=1}^{n} x_{ij}^2 = 1\ for\ j = 1, \ldots, p. \tag{2.10}$$

The elastic net is described as

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda_1|\boldsymbol{\beta}|_1 + \lambda_2|\boldsymbol{\beta}|^2 \tag{2.11}$$

Similar to the alternative solution of ridge regression, the solution of elastic net is also achieved by an augmented data set,

$$\mathbf{X}^*_{(\mathbf{n+p})\times\mathbf{p}} = (\mathbf{1} + \lambda_\mathbf{2})^{-\mathbf{1/2}} \begin{pmatrix} \mathbf{X} \\ \sqrt{\lambda_2}\mathbf{I_p} \end{pmatrix}, \quad \mathbf{y}^*_{\mathbf{n+p}} = \begin{pmatrix} \mathbf{y} \\ 0 \end{pmatrix} \tag{2.12}$$

Set $\boldsymbol{\beta}^* = \sqrt{1+\lambda_2}\boldsymbol{\beta}$, $\gamma = \lambda_1/\sqrt{1+\lambda_2}$, so that the elastic net is also transformed to an equivalent LASSO problem on augmented data,

$$L(\gamma, \boldsymbol{\beta}) = L(\gamma, \boldsymbol{\beta}^*) = |\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*|^2 + \gamma|\boldsymbol{\beta}^*|_1 \tag{2.13}$$

Since the solution for this LASSO problem is

$$\hat{\boldsymbol{\beta}}^* = arg\min_{\boldsymbol{\beta}^*}\{L(\gamma, \boldsymbol{\beta}^*)\} \tag{2.14}$$

then the solution for elastic net is

$$\hat{\boldsymbol{\beta}} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\boldsymbol{\beta}}^* \tag{2.15}$$

2.3 *Network-constrained regularization*

Similarly, in network-constrained regularization, the design matrix is standardized and the response vector is centered.

The network containing prior information in the data is depicted and represented as $G = (\mathbf{V}, \mathbf{E}, \mathbf{W})$, where $\mathbf{V}$ is the set of vertices corresponding to p predictors, $\mathbf{E} = (\mathbf{u} \sim \mathbf{v})$ is the set of edges between linked predictors u and v and $\mathbf{W} = (\mathbf{u} \sim \mathbf{v})$ is the corresponding weight of the edge. The degree of vertex v is defined as $d_v = \sum_{u \sim v} w(u, v)$ and $d_u = 0$ indicates that predictor u is an isolated vertex.

In general, the Laplacian of graph G is defined by

$$\mathbf{L}(u,v) = \begin{cases} d_u - w(u,v) & if\ u = v\ and\ d_u \neq 0 \\ -w(u,v) & if\ u\ and\ v\ adjacent \\ 0 & otherwise \end{cases} \tag{2.16}$$

But the Laplacian matrix is not always non-negative definite and can not always be spectral decomposed. So a normalized Laplacian matrix is introduced as

$$\mathbf{L}(u,v) = \begin{cases} 1 - w(u,v)/d_u & if\ u = v\ and\ d_u \neq 0 \\ -w(u,v)/\sqrt{d_u d_v} & if\ u\ and\ v\ adjacent \\ 0 & otherwise \end{cases} \tag{2.17}$$

In this context, the network-constrained regularization criterion is defined as

$$L(\lambda_1, \lambda_2, \boldsymbol{\beta}) = |\mathbf{y} - \mathbf{X}\boldsymbol{\beta}|^2 + \lambda_1 |\boldsymbol{\beta}|_1 + \lambda_2 \boldsymbol{\beta}^T \mathbf{L} \boldsymbol{\beta} \tag{2.18}$$

where

$$\boldsymbol{\beta}^T \mathbf{L} \boldsymbol{\beta} = \sum_{\mathbf{u} \sim \mathbf{v}} \left(\frac{\beta_{\mathbf{u}}}{\sqrt{\mathbf{d_u}}} - \frac{\beta_{\mathbf{v}}}{\sqrt{\mathbf{d_v}}}\right)^{\mathbf{2}} \mathbf{w}(\mathbf{u}, \mathbf{v}) \tag{2.19}$$

By spectral decomposition, the Laplacian matrix can be decomposed as $\mathbf{L} = \mathbf{U\Gamma U^T}$, $\mathbf{S} = \mathbf{U\Gamma^{1/2}}$ so that $\mathbf{L} = \mathbf{SS^T}$.

If an augmented data is defined by

$$\mathbf{X}^*_{\mathbf{(n+p)\times p}} = (\mathbf{1}+\lambda_\mathbf{2})^{-\mathbf{1/2}}\begin{pmatrix}\mathbf{X}\\ \sqrt{\lambda_2}\mathbf{S^T}\end{pmatrix}, \quad \mathbf{y}^*_{\mathbf{n+p}} = \begin{pmatrix}\mathbf{y}\\ 0\end{pmatrix} \tag{2.20}$$

Let $\boldsymbol{\beta}^* = \sqrt{1+\lambda_2}\boldsymbol{\beta}$, $\gamma = \lambda_1/\sqrt{1+\lambda_2}$, then the solution is achieved by

$$L(\gamma, \boldsymbol{\beta}) = L(\gamma, \boldsymbol{\beta}^*) = |\mathbf{y}^* - \mathbf{X}^*\boldsymbol{\beta}^*|^2 + \gamma|\boldsymbol{\beta}^*|_1 \tag{2.21}$$

$$\hat{\boldsymbol{\beta}}^* = arg\min_{\boldsymbol{\beta}^*}\{L(\gamma, \boldsymbol{\beta}^*)\} \tag{2.22}$$

$$\hat{\boldsymbol{\beta}} = \frac{1}{\sqrt{1+\lambda_2}}\hat{\boldsymbol{\beta}}^* \tag{2.23}$$

Compared with elastic net, the network-constrained regularization shares the same solution pathway which is by transformation to a LASSO problem on an artificial data set. However, the design of the augmented data is different, because the network-constrained regularization considers the graph structure in the design matrix.

3. Simulation

Suppose we have 10 transcription factors (TFs) and each TF regulates 10 genes, which yields a total number of $10 + 10 \times 10 = 110$ predictors.

The basic model is set up as $\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \epsilon$, and $\epsilon \sim N(0, 1)$. Then four models were designed by four different coefficients $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$, $\boldsymbol{\beta}_3$, $\boldsymbol{\beta}_4$.

In the first model, data is simulated with $\boldsymbol{\beta}_1$ assigned with

$$\begin{aligned}\boldsymbol{\beta}_1 = (5, \frac{5}{\sqrt{10}}, \ldots, \frac{5}{\sqrt{10}}, -5, \frac{-5}{\sqrt{10}}, \ldots, \frac{-5}{\sqrt{10}},\\ 3, \frac{3}{\sqrt{10}}, \ldots, \frac{3}{\sqrt{10}}, -3, \frac{-3}{\sqrt{10}}, \ldots, \frac{-3}{\sqrt{10}}, 0 \ldots 0)\end{aligned} \tag{3.24}$$

The predictors are aligned in order of TF with its ten genes. The first four TFs and corresponding genes are assumed to contribute to the response and the absolute value of coefficients

in each TF is the same. Then the coefficients of the last 66 predictors are set to zero.

The second model assumes that genes regulated by the same TF have both positive and negative effect on response. The effect of the first three genes are opposite to that of the last seven genes, so the $\boldsymbol{\beta}_2$ is given as

$$\begin{aligned}\boldsymbol{\beta}_2 = (5, &\frac{-5}{\sqrt{10}}, \frac{-5}{\sqrt{10}}, \frac{-5}{\sqrt{10}}, \frac{5}{\sqrt{10}}, \ldots, \frac{5}{\sqrt{10}},\\ -5, &\frac{5}{\sqrt{10}}, \frac{5}{\sqrt{10}}, \frac{5}{\sqrt{10}}, \frac{-5}{\sqrt{10}}, \ldots, \frac{-5}{\sqrt{10}},\\ 3, &\frac{-3}{\sqrt{10}}, \frac{-3}{\sqrt{10}}, \frac{-3}{\sqrt{10}}, \frac{3}{\sqrt{10}}, \ldots, \frac{3}{10},\\ -3, &\frac{3}{\sqrt{10}}, \frac{3}{\sqrt{10}}, \frac{3}{\sqrt{10}}, \frac{-3}{\sqrt{10}}, \ldots, \frac{-3}{\sqrt{10}}, 0 \ldots 0)\end{aligned} \tag{3.25}$$

The third model is modified from the first model by replacing the denominators with 10.

$\boldsymbol{\beta}_3$ assigned with

$$\begin{aligned}\boldsymbol{\beta}_3 = (5, &\frac{5}{10}, \ldots, \frac{5}{10}, -5, \frac{-5}{10}, \ldots, \frac{-5}{10},\\ 3, &\frac{3}{10}, \ldots, \frac{3}{10}, -3, \frac{-3}{10}, \ldots, \frac{-3}{10}, 0 \ldots 0)\end{aligned} \tag{3.26}$$

The forth model is modified from the second model by replacing the denominators with 10.

$$\begin{aligned}\boldsymbol{\beta}_4 = (5, &\frac{-5}{10}, \frac{-5}{10}, \frac{-5}{10}, \frac{5}{10}, \ldots, \frac{5}{10},\\ -5, &\frac{5}{10}, \frac{5}{10}, \frac{5}{10}, \frac{-5}{10}, \ldots, \frac{-5}{10},\\ 3, &\frac{-3}{10}, \frac{-3}{10}, \frac{-3}{10}, \frac{3}{10}, \ldots, \frac{3}{10},\\ -3, &\frac{3}{10}, \frac{3}{10}, \frac{3}{10}, \frac{-3}{10}, \ldots, \frac{-3}{10}, 0 \ldots 0)\end{aligned} \tag{3.27}$$

The training set and test set were simulated independently with a sample size of 100 for each. The training set was randomly split at ratio 0.5 to select the best combination of λ_1 and λ_2 based on the least mean square error (MSE). Then the optimized tune parameters were used to fit the model with training set and the performance was assessed on the test set. The sensitivity (true positive), specificity (true negative) and prediction mean-squared-error (PMSE) are adapted to evaluate the prediction accuracy. The results based on 50 replicates were shown in Table 1.

Model	Sensitivity			Specificity			PMSE		
	Lasso	Enet	Net	Lasso	Enet	Net	Lasso	Enet	Net
1	0.987	0.998	1.00	0.828	0.924	0.941	2.959	2.471	1.467
	(0.01)	(0.00)	(0.00)	(0.01)	(0.00)	(0.00)	(0.24)	(0.09)	(0.03)
2	0.754	0.816	0.835	0.812	0.762	0.721	9.595	8.161	7.788
	(0.02)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.61)	(0.21)	(0.17)
3	0.759	0.797	0.914	0.845	0.920	0.908	2.178	2.032	2.177
	(0.02)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)	(0.10)	(0.05)	(0.06)
4	0.560	0.602	0.634	0.865	0.896	0.895	2.563	2.299	2.333
	(0.02)	(0.01)	(0.01)	(0.02)	(0.00)	(0.01)	(0.09)	(0.06)	(0.07)

Table 1: Result of the simulation study. The sensitivity, specificity and prediction mean square errors (PMSE) are calculated based on 50 simulations, where the standard errors are given in parentheses. Enet: elastic net; Net: network-constrained regulation

In this simulation study, compared with LASSO and elastic net, the network-constrained regulation procedure performs relatively better by producing less or comparable PMSE and by obtaining more accuracy in aspects of sensitivity and specificity.

4. Discussion

The network-constrained regularization is introduce to analyze data with prior information depicting the data structure. Both the L_1 and L_2 penalty are employed for coefficient estimate. The L_1 norm functions for variable selection by shrinking the coefficient to zero while the L_2 norm promote the effect of group feature. Therefore, the elastic net and network-constrained regularization share the characteristics of both LASSO and ridge regression. Compared with elastic net, the network-constrained regularization is more specific based on the known network of data. As the development in fields of genomics and computing algorithm, the gene-expression database would be a potent resource to improve data analysis. So the network-constrained regularization is proposed to deal with high-dimensional data with network or graphs from established database.

In this simulation study, the data was assumed to have a genetic network structured as groups. Considering the clustering effect in the gene regulation, the prediction accuracy from network-constrained regularization is much higher than those of LASSO and elastic net in terms of

sensitivity, specificity and PMSE. It also has good performance in models when the predictors have both positive and negative effects on the response, while LASSO and elastic net are not adept. Although the accuracy decreases as the magnitude of coefficient reduces, the overall performance is still relatively better than LASSO and elastic net.

References

Fan, Jianqing and Li, Runze. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American statistical Association* **96**(456), 1348–1360.

Li, Caiyan and Li, Hongzhe. (2008). Network-constrained regularization and variable selection for analysis of genomic data. *Bioinformatics* **24**(9), 1175–1182.

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

Tibshirani, Robert, Saunders, Michael, Rosset, Saharon, Zhu, Ji and Knight, Keith. (2005). Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(1), 91–108.

Yuan, Ming and Lin, Yi. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(1), 49–67.

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

Zou, Hui and Hastie, Trevor. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67**(2), 301–320.

[*Received Dec 1, 2019; revised Dec 1, 2019; accepted for publication December 1, 2019*]