Science Journal of Applied Mathematics and Statistics
2015; 3(6): 293-297
Published online December 30, 2015 (http://www.sciencepublishinggroup.com/j/sjams)
doi: 10.11648/j.sjams.20150306.19
ISSN: 2376-9491 (Print); ISSN: 2376-9513 (Online)

The Optimal Estimation of Lasso

Huiyi Xia

Department of Mathematics and Computer Science, Chizhou University, Anhui, China

Email address:
xiayz_88@163.com

To cite this article:
Huiyi Xia. The Optimal Estimation of Lasso. *Science Journal of Applied Mathematics and Statistics.* Vol. 3, No. 6, 2015, pp. 293-297.
doi: 10.11648/j.sjams.20150306.19

Abstract: The estimation of lasso is important problem of high dimensional data; the optimal estimation's formula of lasso is unsolved riddle of high dimensional data. In order to solve this problem, we give the structure of lasso estimation by using mathematical method in the orthogonal design. The optimal estimation's formula of lasso is solved in the orthogonal design, it is pointed out that there is a gradual process of dimension reduction by using method of lasso.

Keywords: Lasso, Estimation, Solution

1. Introduction

Tibshirani (1996) propose a new technique, called lasso. It shrinks coefficients and set others to 0, and hence tries to retain the good features of both ridge regression and subset selection. The lasso estimate has 'soft threshold' estimator by Donoho and Johnstone (1994). Fan and Li (2001) propose SCAD that the penalty functions is the smoothly clipped absolute deviation. Knight and Fu (2000) research asymptotic for lasso-type estimation. Efron et al. (2004) propose a new model selection algorithm, called Least Angle Regression (LARS); a simple modification of the LARS algorithm may implement the lasso. Because algorithm of LARS is very fast, making the method of lasso is popular in the world. Zou and Hastie (2005) propose the elastic net, real world data and a simulation study show the elastic net often outperforms the lasso. An algorithm called LARS-EN is proposed for computing elastic net regularization paths efficiently, much like algorithm LARS does for the lasso. Tibshiani et al. (2005) proposed the 'fused lasso', the fused lasso penalizes the L_1-norm of both the coefficients and their successive differences. The technique is also extended to the 'hinge' loss function that underlies the support vector classifier. Wasserman and Roeder (2009) doing variable selection in the high-dimensional models, and consider three screening methods: the lasso, marginal regression, and forward stepwise regression. Zou and Zhang (2009) research the adaptive elastic-net with diverging number of parameters. Austin et al. (2013) study penalized regression and risk prediction in genome-wide association studies by lasso. Wu et al. (2014) proposes the nonnegative-lasso method for variable selection in high dimensional sparse linear models with the nonnegative constraints on the coefficients. This method is an extension of lasso. Bunea et al. (2013) introduce and study the Group Square-Root Lasso (GSRL) method for estimation in high dimensional sparse regression models with group structure. Ahrens and Bhattacharjee (2015) exploit the lasso estimator and mimics two-step least squares to account for endogeneity of the spatial lag.

Related research of lasso is very much, it is inconvenient one by one in this narrative. The optimal estimation of lasso is unsolved riddle of lasso.

For example, if we use the lasso method to select five variables from ten variables, because of the tuning parameter is not unique. How to choose the tuning parameters to get the best estimate of lasso? We refer to the literature and found that this problem has not been solved. After careful deliberation, we solved this problem.

2. Some Definition

Suppose that we have data $\left(x^i, y_i\right)$, where y_i are the responses, $i = 1, 2, \cdots, N$ and $x^i = \left(x_{i1}, \cdots, x_{ip}\right)^T$ are the predictor variables. We assume the observations are independent, the x_{ij} are standardized so that:

$$\sum_i x_{ij} / N = 0, \ \sum_i x_{ij}^2 / N = 1.$$

Let $\hat{\beta}=\left(\hat{\beta}_1,\cdots,\hat{\beta}_p\right)^T$, we can loss of generality that $\bar{y}=0$, and the tuning parameter $t\geq 0$.

The lasso estimate $\hat{\beta}$ is defined by

$$\hat{\beta}=\arg\min\left\{\sum_{i=1}^{N}\left(y_i-\sum_j\beta_j x_{ij}\right)^2\right\},\ \sum_j\left|\beta_j\right|\leq t \qquad (1)$$

Ridge regression estimate $\hat{\beta}$ is defined by

$$\hat{\beta}=\arg\min\left\{\sum_{i=1}^{N}\left(y_i-\sum_j\beta_j x_{ij}\right)^2\right\},\ \sum_j\beta_j^2\leq t \qquad (2)$$

Let X be the $n\times p$ design matrix with ijth entry x_{ij}, and suppose that $X^TX=I$, I denotes the identity matrix. Let $\hat{\beta}^0=\left(\hat{\beta}_1^0,\cdots,\hat{\beta}_p^0\right)^T$ be the full least squares estimate.

The solution to equation (1) are easily shown to be

$$\hat{\beta}_j=sign\left(\hat{\beta}_j^0\right)\left(\left|\hat{\beta}_j^0\right|-\gamma\right)^+ \qquad (3)$$

Where γ is determined by the condition $\sum\hat{\beta}_j=t$.

(3) is the soft threshold estimator.

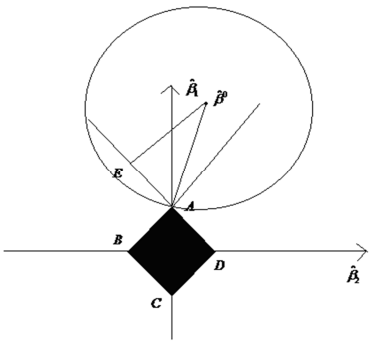

Figure 1. *The picture 1 of lasso.*

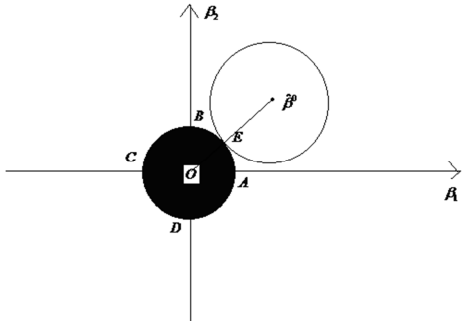

Figure 2. *The picture of ridge regression.*

When $t\geq\sqrt{\left(\hat{\beta}_1^0\right)^2+\cdots+\left(\hat{\beta}_p^0\right)^2}$, the solution to equation (2) are easily shown to be:

$$\hat{\beta}=\hat{\beta}^0=\left(\hat{\beta}_1^0,\cdots,\hat{\beta}_p^0\right)^T$$

When $t<\sqrt{\left(\hat{\beta}_1^0\right)^2+\cdots+\left(\hat{\beta}_p^0\right)^2}$, the solution to equation (2) are easily shown to be

$$\hat{\beta}_j=\frac{1}{1+t}\hat{\beta}_j^0 \qquad (4)$$

Figure 1 and Figure 2 provides some insight for the case $p=2$.

When $X^TX=I$, the criterion $\sum_{i=1}^{N}\left(y_i-\sum_j\beta_j x_{ij}\right)^2$ equals the quadratic function $\left(\beta-\hat{\beta}^0\right)^T\left(\beta-\hat{\beta}^0\right)$. The circular contours of this function are shown by the full curves in Figure.2; they are centered at the OLS estimates $\hat{\beta}^0$, the constraint region is the rotated square. The lasso solution is the first place that the contours touch the square, and this will sometimes occur at a corner, corresponding to a zero coefficient.

3. Some Result

Lemma 1: Let $\hat{\beta}^0=\left(\hat{\beta}_1^0,\cdots,\hat{\beta}_p^0\right)^T$, $\beta=\left(\beta_1,\cdots,\beta_p\right)^T$, $p>2$, and

$$\hat{\beta}_p^0>\cdots>\hat{\beta}_1^0>0 \qquad (5)$$

There exist j, make

$$\begin{cases}\hat{\beta}_j^0-\dfrac{1}{p-j+1}\left(\hat{\beta}_j^0+\cdots+\hat{\beta}_p^0-t\right)\geq 0\\ \hat{\beta}_{j-1}^0-\dfrac{1}{p-j+2}\left(\hat{\beta}_{j-1}^0+\cdots+\hat{\beta}_p^0-t\right)\leq 0\end{cases} \qquad (6)$$

$$\hat{\beta}=\arg\min\left(\beta-\hat{\beta}^0\right)^T\left(\beta-\hat{\beta}^0\right),\ \sum_{j=1}^{p}\left|\beta_j\right|\leq t \qquad (7)$$

Let $\hat{\beta}=\left(\hat{\beta}_1,\cdots,\hat{\beta}_p\right)^T$ is solution of (7), then

$$\begin{cases}\hat{\beta}_1=0,\cdots,\hat{\beta}_{j-1}=0,\\ \hat{\beta}_j=\hat{\beta}_j^0+\dfrac{1}{p-j+1}\left(t-\hat{\beta}_j^0-\cdots-\hat{\beta}_p^0\right),\cdots,\\ \hat{\beta}_p=\hat{\beta}_p^0+\dfrac{1}{p-j+1}\left(t-\hat{\beta}_j^0-\cdots-\hat{\beta}_p^0\right)\end{cases} \qquad (8)$$

Proof: According to (5), (7) is equivalent to

$$\hat{\beta} = \arg\min\left(\beta - \hat{\beta}^0\right)^T \left(\beta - \hat{\beta}^0\right), \sum_{j=1}^{p} \left|\beta_j\right| \le t \qquad (9)$$

We obtain solution of (9):

$$\begin{cases} \hat{\beta}_1 = 0, \cdots, \hat{\beta}_{j-1} = 0, \\ \hat{\beta}_j = \hat{\beta}_j^0 + \dfrac{1}{p-j+1}\left(t - \hat{\beta}_j^0 - \cdots - \hat{\beta}_p^0\right), \cdots, \\ \hat{\beta}_p = \hat{\beta}_p^0 + \dfrac{1}{p-j+1}\left(t - \hat{\beta}_j^0 - \cdots - \hat{\beta}_p^0\right) \end{cases}$$

Theorem 1: Let $\beta = \left(\beta_1, \cdots, \beta_p\right)^T$, we can loss of generality that $\bar{y} = 0$, $X^T X = I$ and the tuning parameter $t \ge 0$. Let $\hat{\beta}^0 = \left(\hat{\beta}_1^0, \cdots, \hat{\beta}_p^0\right)^T$ be the least squares estimate $\left(\hat{\beta}_1^0 \ne 0, \cdots, \hat{\beta}_p^0 \ne 0\right)$. The lasso estimate $\hat{\beta}$ is defined by (1). Then, we proved that some coefficients become 0 by method of lasso.

Proof: when $X^T X = I$, (1) equivalent to (7)

1. when $p = 2$,

(1) When $\left|\hat{\beta}_2^0\right| = \left|\hat{\beta}_1^0\right|$, the coefficient can not become 0 by method of lasso.

(2) When $\hat{\beta}_2^0 > \hat{\beta}_1^0 > 0$, $\beta = (\beta_1, \beta_2)^T$, $t \le \hat{\beta}_2^0 - \hat{\beta}_1^0$.

As figure.1, the line AB denotes $\beta_1 + \beta_2 = t$, the point $\hat{\beta}^0$ denotes the least squares estimate, the point $\hat{\beta}^0$ above the line AB, the rotated square $ABCD$ denotes the constraint region, $\hat{\beta} = \arg\min\left(\beta - \hat{\beta}^0\right)^T \left(\beta - \hat{\beta}^0\right)$ equivalent to the shortest distance point of from the rotated square $ABCD$ to the point $\hat{\beta}^0$.

$$\left|\hat{\beta}^0 A\right| = \sqrt{\left|\hat{\beta}_0 E\right|^2 + \left|EA\right|^2}$$

Obviously $\left|\hat{\beta}^0 A\right|$ is the shortest distance between the point $\hat{\beta}^0$ and the rotated square $ABCD$. The point A of the rotated square $ABCD$ is the nearest point of the point $\hat{\beta}^0$.

We may assume that the coordinates of the point A is $\left(0, \hat{\beta}_2\right)$, the lasso estimate is $\hat{\beta} = \left(0, \hat{\beta}_2\right)^T$.

We proved that a coefficient become 0 by method of lasso.

(3) Suppose $\hat{\beta}_1^0 > 0$, $\hat{\beta}_2^0 < 0$, $\left|\hat{\beta}_2^0\right| > \left|\hat{\beta}_1^0\right|$, $\beta = (\beta_1, \beta_2)^T$, $t < \left|\hat{\beta}_2^0\right| - \left|\hat{\beta}_1^0\right| = -\hat{\beta}_2^0 - \hat{\beta}_1^0$.

As shown figure 3, the point $\hat{\beta}^0$ denotes the least squares estimate, the point $\hat{\beta}^0$ below the line BC, the rotated square $ABCD$ denotes the constraint region, the line BC denotes $\left|\beta_1\right| + \left|\beta_2\right| = \beta_1 - \beta_2 = t$.

$\hat{\beta} = \arg\min\left(\beta - \hat{\beta}^0\right)^T \left(\beta - \hat{\beta}^0\right)$ equivalent to the shortest distance point of from the rotated square $ABCD$ to the point $\hat{\beta}^0$.

$$\left|\hat{\beta}^0 C\right| = \sqrt{\left|\hat{\beta}_0 E\right|^2 + \left|EA\right|^2}$$

Obviously, $\left|\hat{\beta}_0 C\right|$ is the shortest distance between the point $\hat{\beta}^0$ and the rotated square $ABCD$, the point C of the rotated square $ABCD$ is the nearest point of the point $\hat{\beta}^0$.

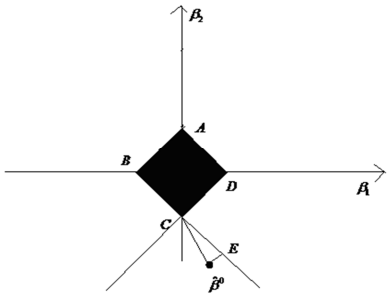

Figure 3. *The picture 2 of lasso.*

We may assume that the coordinates of the point C is $\left(0, \hat{\beta}_2\right)$. The lasso estimate is $\hat{\beta} = \left(0, \hat{\beta}_2\right)^T$. We proved that some coefficients become 0 by method of lasso.

On the other two cases: $\hat{\beta}_1^0 < 0$, $\hat{\beta}_2^0 > 0$; $\hat{\beta}_1^0 < 0$, $\hat{\beta}_2^0 < 0$. Similarly the two cases can be proved.

Thus, when $p = 2$, we proved that some coefficient become 0 by method of lasso.

2. When $p > 2$

(1) There are equal numbers in $\left|\hat{\beta}_1^0\right|, \cdots, \left|\hat{\beta}_p^0\right|$, equal number of $\left|\hat{\beta}_1^0\right|, \cdots, \left|\hat{\beta}_p^0\right|$ as one factor; we proved that some coefficients become 0 by method of lasso.

(2) $\hat{\beta}_p^0 > \cdots > \hat{\beta}_1^0 > 0$, $\hat{\beta} = \left(\hat{\beta}_1, \cdots, \hat{\beta}_p\right)^T$,

$$\begin{cases} \hat{\beta}_j^0 - \dfrac{1}{p-j+1}\left(\hat{\beta}_j^0 + \cdots + \hat{\beta}_p^0 - t\right) \ge 0 \\ \hat{\beta}_{j-1}^0 - \dfrac{1}{p-j+2}\left(\hat{\beta}_{j-1}^0 + \cdots + \hat{\beta}_p^0 - t\right) \le 0 \end{cases}$$

According to Lemma 1, solution of (1) is

$$\begin{cases} \hat{\beta}_1 = 0, \cdots, \hat{\beta}_{j-1} = 0, \\ \hat{\beta}_j = \hat{\beta}_j^0 + \dfrac{1}{p-j+1}\left(t - \hat{\beta}_j^0 - \cdots - \hat{\beta}_p^0\right), \cdots, \\ \hat{\beta}_p = \hat{\beta}_p^0 + \dfrac{1}{p-j+1}\left(t - \hat{\beta}_j^0 - \cdots - \hat{\beta}_p^0\right) \end{cases}$$

In the case of $\hat{\beta}_p^0 > \beta_{p-1}^0 > \cdots > \hat{\beta}_1^0 > 0$, we proved that some coefficients become 0 by method of lasso.

(3) $\left|\hat{\beta}_p^0\right| > \cdots > \left|\hat{\beta}_1^0\right| > 0$, we suppose $\hat{\beta}_j^0 < 0$, other parameters are bigger than 0. Let $\hat{\beta}_j'^0 = -\hat{\beta}_j^0$, we consider $\hat{\beta}_p^0 > \cdots > \hat{\beta}_j'^0 > \cdots > \hat{\beta}_1^0 > 0$ by symmetry of $\beta_1, \cdots, \beta_p$, we proved that some coefficients become 0 by method of lasso. The other conditions of $\hat{\beta}_1^0, \cdots, \hat{\beta}_p^0$ can be proved similarly. Thus theorem 1 is proven.

Inference 1: Let $\hat{\beta}^0 = \left(\hat{\beta}_1^0, \cdots, \hat{\beta}_p^0\right)^T$, $\hat{\beta}_p^0 > \cdots > \hat{\beta}_1^0 > 0$, $p \geq 2$, There exist t, make

$$\hat{\beta}_j^0 + \cdots + \hat{\beta}_p^0 - (p-j+1)\hat{\beta}_j^0 < t$$
$$t \geq \hat{\beta}_{j-1}^0 + \cdots + \hat{\beta}_p^0 - (p-j+2)\hat{\beta}_{j-1}^0$$

$\hat{\beta} = \left(\hat{\beta}_1, \cdots, \hat{\beta}_p\right)^T$ is solution of (7)

Then, when $t = \hat{\beta}_{j-1}^0 + \cdots + \hat{\beta}_p^0 - (p-j+2)\hat{\beta}_{j-1}^0$,

$\hat{\beta} = \left(\hat{\beta}_1, \cdots, \hat{\beta}_p\right)^T$ is the optimal estimation of lasso.

Example let $y_1 = 0.5, y_2 = 1, y_3 = 5, y_4 = 6$, $X = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix}$. What is the optimal estimation of lasso?

Proof: the least squares estimate $\hat{\beta}^0 = (0.5, 1, 5, 6)^T$, let $10.5 < t \leq 12.5$, then the lasso estimate:

$$\begin{cases} \hat{\beta}_1 = \hat{\beta}_1^0 - \left(\hat{\beta}_1^0 + \hat{\beta}_2^0 + \hat{\beta}_3^0 + \hat{\beta}_4^0 - t\right)/4 \\ \hat{\beta}_2 = \hat{\beta}_2^0 - \left(\hat{\beta}_1^0 + \hat{\beta}_2^0 + \hat{\beta}_3^0 + \hat{\beta}_4^0 - t\right)/4 \\ \hat{\beta}_3 = \hat{\beta}_3^0 - \left(\hat{\beta}_1^0 + \hat{\beta}_2^0 + \hat{\beta}_3^0 + \hat{\beta}_4^0 - t\right)/4 \\ \hat{\beta}_4 = \hat{\beta}_4^0 - \left(\hat{\beta}_1^0 + \hat{\beta}_2^0 + \hat{\beta}_3^0 + \hat{\beta}_4^0 - t\right)/4 \end{cases}$$

$$\begin{cases} \hat{\beta}_1 = t/4 - 2.625 \\ \hat{\beta}_2 = t/4 - 2.125 \\ \hat{\beta}_3 = t/4 + 1.875 \\ \hat{\beta}_4 = t/4 + 2.875 \end{cases}$$

When $t = 10.5$, then the optimal estimation of lasso:

$$\begin{cases} \hat{\beta}_1 = 0 \\ \hat{\beta}_2 = 0.5 \\ \hat{\beta}_3 = 4.5 \\ \hat{\beta}_4 = 5.5 \end{cases}$$

When $\beta_1 = 0$, let $9 < t \leq 10.5$, then the lasso estimate:

$$\begin{cases} \hat{\beta}_2 = \hat{\beta}_2^0 - \left(\hat{\beta}_2^0 + \hat{\beta}_3^0 + \hat{\beta}_4^0 - t\right)/3 \\ \hat{\beta}_3 = \hat{\beta}_3^0 - \left(\hat{\beta}_2^0 + \hat{\beta}_3^0 + \hat{\beta}_4^0 - t\right)/3 \\ \hat{\beta}_4 = \hat{\beta}_4^0 - \left(\hat{\beta}_2^0 + \hat{\beta}_3^0 + \hat{\beta}_4^0 - t\right)/3 \end{cases}, \begin{cases} \hat{\beta}_2 = t/3 - 3 \\ \hat{\beta}_3 = t/3 + 1 \\ \hat{\beta}_4 = t/3 + 2 \end{cases},$$

When $\beta_1 = \beta_2 = 0$, let $1 < t \leq 9$, then the lasso estimate

$$\begin{cases} \hat{\beta}_3 = \hat{\beta}_3^0 - \left(\hat{\beta}_3^0 + \hat{\beta}_4^0 - t\right)/2 \\ \hat{\beta}_4 = \hat{\beta}_4^0 - \left(\hat{\beta}_3^0 + \hat{\beta}_4^0 - t\right)/2 \end{cases}, \begin{cases} \hat{\beta}_3 = t/2\text{-}0.5 \\ \hat{\beta}_4 = t/2 + 0.5 \end{cases}$$

When $t = 9$, then the optimal estimation of lasso:

$$\begin{cases} \hat{\beta}_1 = 0 \\ \hat{\beta}_2 = 0 \\ \hat{\beta}_3 = 4 \\ \hat{\beta}_4 = 5 \end{cases}$$

When $\beta_1 = \beta_2 = \beta_3 = 0$, let $0 < t \leq 1$, then the lasso estimate $\beta_4 = t$

$t = 1$, then the optimal estimation of lasso:

$$\begin{cases} \hat{\beta}_1 = 0 \\ \hat{\beta}_2 = 0 \\ \hat{\beta}_3 = 0 \\ \hat{\beta}_4 = 1 \end{cases}$$

4. Conclusion

The lasso estimate has 'soft threshold' estimator by Donoho and Johnstone, We give a new estimate of the lasso estimation, and we obtained the following conclusions with the new estimates and examples:

1. There is a gradual process of dimension reduction by using method of lasso, p variables of lasso can only get

rid of one variable using p-dimensional data of lasso, If you want to get rid of the second variables, and you must use p-1 dimensional data of lasso, Present algorithm of lasso must be modified.
2. The calculation formula of the optimal Lasso is found.
3. Making a historic contribution to the computation of high dimensional data.

Acknowledgements

I would like to express my gratitude to all those who helped me during the writing of this paper.

References

[1] R. Tibshirani. "Regression Shrinkage and Selection Via the Lasso," Journal of the Royal Statistical Society. 1996, Series B, 58(1), pp: 267-288.

[2] B. Efron, T. Hastie, I. "Johnstone and R. Tibshirani. Least Angle Regression," The Annals of Statistics 2004, Vol. 32. No. 2, pp: 407-499.

[3] J. Fan and R. Li. "Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties," Journal of the American Statistical Association. 2001, Vol. 96, No. 456, pp: 1348-1360.

[4] K. Knight, W. Fu. "Asymptotics for Lasso-Type Estimators," The Annals of Statistics. 2000, Vol.28, No. 5, pp: 1356-1378.

[5] H. Zou, H. Zhang, "On the Adaptive Elastic-net with a Diverging Number of Parameters" The Annals of Statistics. 2009, 37(4), pp: 1733–1751.

[6] D. Donohu, I. Johnstone, "Ideal spatial adaption by wavelet shinkage," Biometrica. 1994, 81, pp: 425-455.

[7] H. Zou, T. Hastie, "Regularization and variable selection via the elastic net," Journal of the Royal Statistical Society. 2005, Series B, 67(2), pp: 301-320.

[8] R. Tibshirani, M. Saunders, S. Rossrt, J. Zhu, K. Knight, "Sparsity and smoothness via the fused lasso," Journal of the Royal Statistical Society. 2005, Series B, 67(1), pp: 91-108.

[9] L. Wasserman, K. Roeder, "HIGH-DIMENSIONAL VARIABLE SELECTION," The Annals of Statistics. 2009, 37(5A), pp: 2718–2201.

[10] E. Austin, W. Pan, X. Shen, "Penalized Regression and Risk Prediction in Genome-Wide Association Studies," Stat Anal Data Min. 2013, 6(4), pp: 1: 23.

[11] L. Wu, Y. Yang, H. Liu, "Nonnegative-lasso and in index tracking," Computational Statistics and Data Analysis. 2014, 70, pp: 116-126.

[12] F. Bunea, J. Leder, Y. She, "The Group Square-Root Lasso: Theoretical Properties and Fast Algorithms," Information Theory IEEE Transactions on. 2013, 60(2), pp: 1313-1325.

[13] A. Ahrens, A. Bhattacharjee, "Two-Step Lasso Estimation of the Spatial weighs Matrix," Econometrics. 2015, 3, pp: 128-155.