

A Comparison of Penalized Regressions for Estimating Directed Acyclic Networks

Kyu Min Lee¹

Industrial Management Engineering,
Korea University,
Seoul, Republic of Korea
dlakrb@naver.com

Sung Won Han²

Industrial Management Engineering,
Korea University,
Seoul, Republic of Korea
swhan@korea.ac.kr

Hyunbin Yun²

School of Electrical Engineering,
Korea University,
Seoul, Republic of Korea
Hb_yun@korea.ac.kr

Abstract—

Network models can be classified into two large groups: undirected and directed. Directed network graphs that can represent causal relationships are likely more appropriate in bio-medical data. There have been many studies to estimate DAGs(Directed Acyclic Graphs), of which the two-stage approach using lasso effectively. Find the edges between the nodes in the first step and find the direction in the second step. In this paper, we try to compare which penalized regression is better to find neighborhoods through simulations. We present the result of the simulations that shows which penalized regression is the best.

I. INTRODUCTION

Graphics models are widely used in many areas because they are one of several ways to understand the relationship among variables. Graphic models are often used in biomedical research to investigate gene interaction(Olex, Turkett, Fetrow, & Loeser, 2014; Plouhinec, et al., 2014; Zhou & Zheng, 2013).

Undirected network is a general type of graphical model that describes the relationship between variables. If there is a conditional dependency between variables, the connection and variable is marked as edge and node in graphical models.

Our genes are known to have many specific hub that regulate other genes. So, undirected graphical model is insufficient to explain gene networks. The interactions between the two nodes is represented by directed edges from the parent node to the child node in the directed graphical model. We can represent an whole directed graph with adjacency matrices, unlike undirected graph(Han & Zhong, 2016). Since the directed graph mainly represents the child node as the row node and the parent node as column node, it is almost an asymmetric matrix. Due to the problem of high-dimension gene data with sparse condition no inverse covariance matrix, many studies have proposed a way to overcome that problem and to estimate the network. Meinshausen and Buhlmann (Meinshausen & Bühlmann, 2006) proposed shrinking the neighborhoods of variables by imposing L_1 penalty on the rest of them. Han, Chen and Zhong are lately proposed another hybrid approach for DAGs, which is the NS-DIST. In first stage, neighborhood selection(NS) is performed adaptive Lasso on score-and-search approach used penalty

function, and then Discrete Improving Search with Tabu(DIST) estimates directionalities of them. They showed that the performance of the NS-DIST to estimate DAGs is higher and more robustness than MMHC in dense structures or having some hub nodes with a number of edges. Above all, the reason why NS-DIST algorithm is better than MMHC is that computational time of the estimated DAGs overwhelmingly fast, even if the number of the nodes or estimated DAG complexity increases. On the other hand, computational time of NS-DIST using adaptive lasso stably little increase during that growing complexity(Han, et al., 2016). since NS-DIST is a method to estimate the undirected graph using the penalty function, adaptive lasso, in the first step and to find the directionality within it, it is important how much the edges are estimated in the first step.

A variety of methods were proposed including Meinshausen and Buhlmann's method, graphical lasso and adaptive lasso for estimating undirected networks. Therefore, in this paper, we will compare penalty regressions such as Lasso to find the neighborhood.

II. THE METHODS FOR ESTIMATING DAGS

EXPLANATION OF THE PENALTY REGRESSION FOR ESTIMATING DAGS

Han et al. (Han, et al., 2016) submitted a new approach that uses the same method of using adaptive lasso. This method has greatly improved performance and the computational time.

Suppose that there are p variables and n samples. A data matrix $\mathbf{X} = [\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_p]$ that $\mathbf{X}_j = (x_{1j}, x_{2j}, \dots, x_{nj})^T$ means a vector comprised of n samples with the j -th variables. A DAG illustrates with a structure matrix (Han & Zhong, 2016). A structure matrix has asymmetric and square properties. Clearly, a structure matrix \mathbf{T} can be expressed as

$$\mathbf{T}_{p \times p} \begin{cases} t_{i,j} = 1 & \text{if } j \rightarrow i \\ t_{i,j} = 0 & \text{if } j \nrightarrow i \end{cases} \quad (1)$$

where $(j \rightarrow i)$ means a directed edge with a child variable \mathbf{X}_i and a parent variable \mathbf{X}_j . Estimating a DAG is consistent with estimating the structure matrix because the structure matrix contains the information of the DAGs. In order to find an appropriate estimation method of the undirected graphs, we confirmed the probability of MB and elastic net match the edge between nodes.

A. LASSO

Meinshausen and Bühlmann suggested neighborhood selection method for a high-dimensional sparse graph called LASSO in this paper. MB selects the neighborhood of the node with the lasso regression introduced by Tibshirani.

$$\hat{\theta}^{a,\lambda} = \arg \min_{\theta} \left(\frac{1}{N} \|\mathbf{X}_a^T - \theta \mathbf{X}^T\|_2^2 + \lambda \|\theta\|_1 \right), \quad (2)$$

where $\hat{\theta}^{a,\lambda} = (\hat{\theta}_1^{a,\lambda}, \hat{\theta}_2^{a,\lambda}, \dots, \hat{\theta}_p^{a,\lambda})$, and $\|\theta\|_1$ means L_1 -norm of the coefficient vector. Many other regression based on the L -norm have been proposed to estimate a lot. When $n = 2$ it usually means ridge estimate. When $n = 1$, choose a method that minimize the risk since the variable selection is the only possible value for high-dimensional problems. Finding the oracle parameters to minimize the risk, we need to know the best penalty among the lasso estimates, and λ obtained through cross-validation is optimal (Meinshausen & Bühlmann, 2006). Lasso is an effective solution to estimate directed graph models because of the variable selection property.

B. Elastic net

Because of the Lasso has the convex optimization problem, the number of coefficients selected are smaller than the number of sample size $n < p$. n is the number of the sample size and p is the number of the variables. If the coefficients are not smaller than a certain value, there is a limitation that lasso does not work well. So Zou et al. (Zou, 2005) proposed elastic net to overcome the limit of lasso. naïve elastic net is defined

$$L(\lambda_1, \lambda_2, \theta) = |y - \mathbf{X}\theta|^2 + \lambda_2 \lambda \|\theta\|_2 + \lambda_1 \|\theta\|_1 \quad (3)$$

where $\hat{\theta} = \arg \min_{\theta} \{L(\lambda_1, \lambda_2, \theta)\}$ is the naïve elastic net estimator.

After threshold of lasso, naïve elastic net follows ridge-type shrinkage. When represented the penalty parameter (λ_1, λ_2) , given data (\mathbf{y}, \mathbf{X}) and added data $(\mathbf{y}', \mathbf{X}')$, some of the problem had lasso were supplemented by elastic net.

III. SIMULATION STUDY

In this paper, we use a data generation algorithm that is made by Han *et al* to simulate which method is the most suitable for finding neighborhoods (Han, et al., 2016), random scale-free network model by barabási-albert (Albert, Barabási et al., 2002) and Peng *et al.* (Peng, et al., 2012). Han *et al.* imitates a

directional gene network, assuming it is DAGs(Directed Acyclic Graphs) for simulation. DAGs have a directionality causality relationship should be expressed in a different way from non-directionality gene data, so Han *et al.* asymmetrically set the coefficient matrix to express the DAG. On the other hand, peng applied the Gaussian graphical model.

In this section, we try to simulate hub directed networks made by Han and random undirected networks by Peng. The number of samples is set to be 100 and the number of edges is 50 and 200 for all the simulations. We used `glmnet()` in R package `glmnet` to compare the methods that lasso and elastic net.(Han, et al., 2016). In Elastic net, hyper-parameter alpha were set to $\alpha = 0.2, 0.6$ where alpha is the ratio of Lasso to ridge, one is lasso, zero is same as ridge.

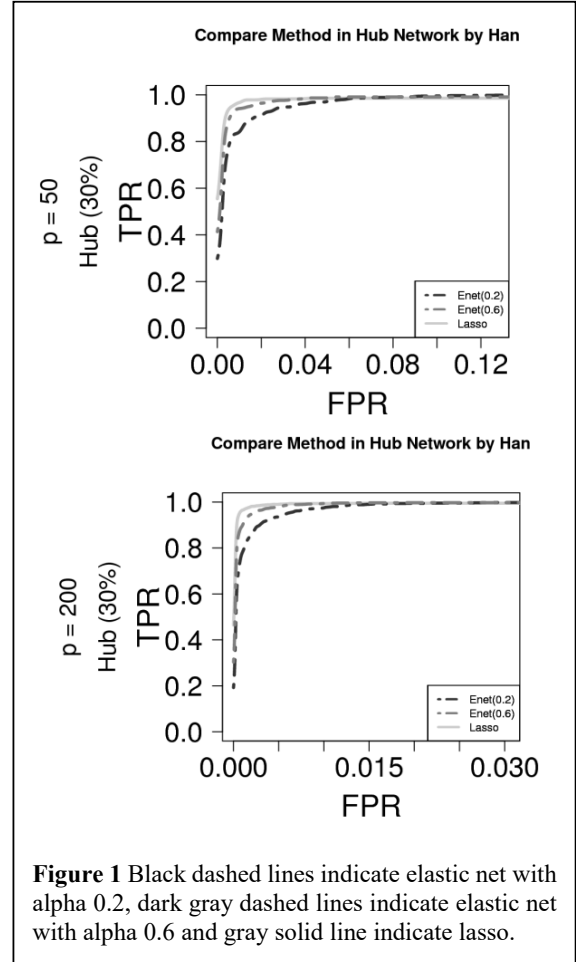


Figure 1 Black dashed lines indicate elastic net with alpha 0.2, dark gray dashed lines indicate elastic net with alpha 0.6 and gray solid line indicate lasso.

As a result of simulations in Han's hub networks where sample size is greater than the number of the edges ($n > p$), lasso was the best and elastic net ($\alpha = 0.2$) was the worst estimator when the variables were selected less because the λ giving the penalty was getting bigger. However, as the penalty decreased and more variables were selected, elastic net ($\alpha = 0.6$)'s true positive ratio (TPR) was a little better than lasso. Finally, elastic

net($\alpha = 0.6$) showed the best appearance, and if set the parameter α well, elastic net seems to be more estimating the edges well network than lasso.

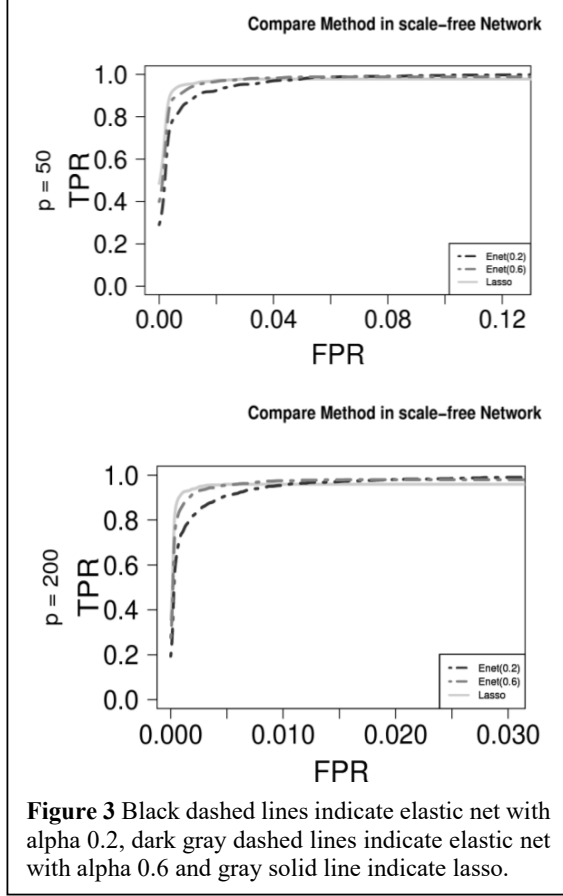


Figure 3 Black dashed lines indicate elastic net with alpha 0.2, dark gray dashed lines indicate elastic net with alpha 0.6 and gray solid line indicate lasso.

According to the simulation in scale-free random network model by barabási-albert, lasso was the best and elastic net $\alpha = 0.2$ was the worst estimator when the variables were selected less. Like the Han's Simulation data, differences of estimating neighborhood selection between elastic net($\alpha = 0.6$) and lasso were not significant in the scale-free model as the penalty decreased and more variables were selected.

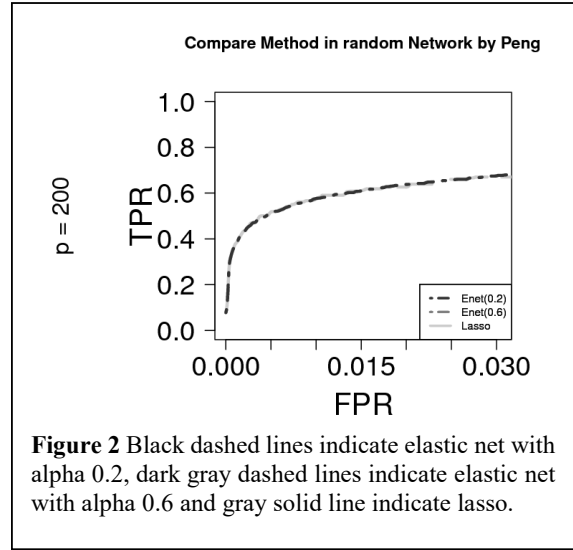
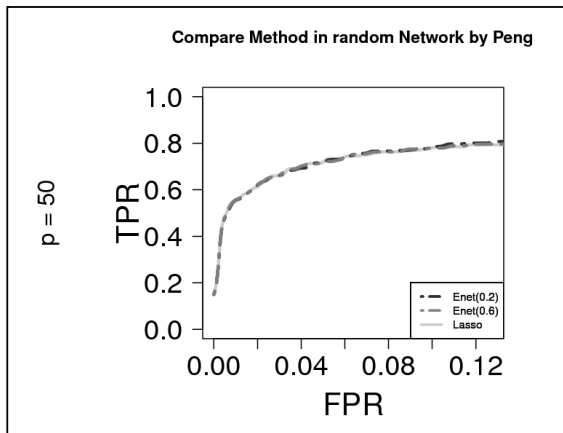


Figure 2 Black dashed lines indicate elastic net with alpha 0.2, dark gray dashed lines indicate elastic net with alpha 0.6 and gray solid line indicate lasso.

In Peng's random network simulation data, which used Gaussian graphical model, both elastic net and lasso methods did not show any differences. Even if the number of nodes changes 50 to 200, did not show any differences

IV. CONCLUSION

To find the most appropriate method for estimating directional graphical model, we compared several methods that have been penalized regression. Except for Peng simulation data, we can see that lasso and elastic net show a similar tendency. And when we looked at the simulation, we were able to confirm that elastic net revise the lasso through ridge-type. This could be expected to lead to a better approach to neighborhood selection. A comparison of penalty regressions can be used to determine a more suitable method for graphical model in gene dataset, and This can be expected that the development of penalty regression in graphical models will lead to a better approach to find neighborhoods selection. Neighborhood selection with a similar pattern is also shown in the further study to find the directionality between nodes. Later, it is necessary to extend to estimate the DAGs in ultra-high dimensional data.

References

- [1] Olex, A. L., Turkett, W. H., Fetrow, J. S., & Loeser, R. F. (2014). Integration of gene expression data with network-based analysis to identify signaling and metabolic pathways regulated during the development of osteoarthritis. *Gene*, 542, 38-45.
- [2] Han, S. W., & Zhong, H. (2016). Estimation of sparse directed acyclic graphs for multivariate counts data. *Biometrics*.
- [3] Han, S. W., Chen, G., Cheon, M.-S., & Zhong, H. (2016). Estimation of Directed Acyclic Graphs Through Two-stage Adaptive Lasso for Gene Network Inference. *Journal of the American Statistical Association*, 1-48.
- [4] Meinshausen, N., & Bühlmann, P. (2006). High-dimensional graphs and variable selection with the lasso. *The annals of statistics*, 1436-1462.
- [5] Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101, 1418-1429.
- [6] Zou, H. (2005). Regularization and variable selection via the elastic net. *ROYAL STATISTICAL SOCIETY*, 67, 301-320.
- [7] Albert ,Réka, Baraba'si , Albert-La'szlo'(2002) Statistical mechanics of complex networks. *Review of Modern Physics*, 74, 47-94.
- [8] Peng, J., Wang, P., Zhou, N., & Zhu, J. (2012). Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*.