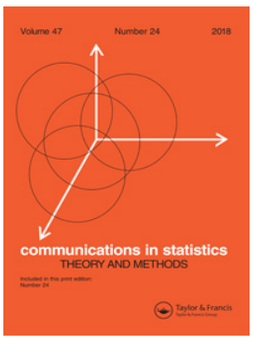


Taylor & Francis
Taylor & Francis Group

Communications in Statistics - Theory and Methods

ISSN: (Print) (Online) Journal homepage: https://www.tandfonline.com/loi/lsta20

A network Lasso model for regression

Meihong Su & Wenjian Wang

To cite this article: Meihong Su & Wenjian Wang (2021): A network Lasso model for regression, Communications in Statistics - Theory and Methods, DOI: 10.1080/03610926.2021.1938125

To link to this article: https://doi.org/10.1080/03610926.2021.1938125

Published online: 14 Jun 2021.

COMMUNICATIONS IN STATISTICS—THEORY AND METHODS
https://doi.org/10.1080/03610926.2021.1938125

A network Lasso model for regression

Meihong Su[a] and Wenjian Wang[b]

[a]School of Computer and Information Technology, Shanxi University, Taiyuan, China; [b]Ministry of Education, Key Laboratory of Computational Intelligence and Chinese Information Processing (Shanxi University), Taiyuan, China

ABSTRACT

Samples often are collected by a network in many modern applications, and the network structure information is potentially helpful in making regression predictions. However, most regression models assume the samples are independent, such as Lasso. Motivated by this, taking the network information into account, we propose a Network Lasso model for regression prediction in this paper. Specially, we consider the effect of the neighborhoods to each response and model each y_i as a linear combination of the covariates x_i, the connected neighbors y_j, and an error term ϵ_i. The corresponding coefficients are referred to effect of node and neighborhoods, respectively. The consistency of the estimators are also established under the regimes where the neighborhoods effect coefficients are known and unknown, respectively. Finally, we evaluate the performance of the proposed model through a series of simulations and a latitude data example.

ARTICLE HISTORY
Received 1 August 2019
Accepted 29 May 2021

KEYWORDS
Network data; lasso; consistency; high dimensionality

1. Introduction

Advances in data collection and social have resulted in network data being collected in many applications, such as communication networks (e.g., telephone networks or the Internet), transportation networks (e.g., networks of roads or rails, or networks of airline routes), and energy networks (e.g., networks for delivery of electricity or gas, or electrical circuits) (Eric and Kolaczy 2009), and so on. The network data records some relational information between units of analysis, at the same times, this information is often collected along with more traditional covariates on each unit of analysis. For example, in the studies of predicting housing prices, the empirically observed phenomenon shows that nearby houses have similar prices and the information on the subjects' adjacent is often available. When predict the price of a house, apart from its own features, the information of its adjacent houses also play an important role through a network link. Thus, it can improve the performances of the model to take advantage of the network structure information in regressive prediction. However, most classical regression models assume the samples are independent, such as bridge penalty (Frank and Friedman 1993); lasso (Tibshirani 1996); SCAD (Fan and Li 2001); elastic net (Zou and Hastie 2005); adaptive lasso (Zou 2006), and so on. The above models generally do not

CONTACT Wenjian Wang wjwang@sxu.edu.cn Key Laboratory of Computational Intelligence and Chinese Information Processing (Shanxi University), Ministry of Education, Taiyuan 030006, China.

take the information of samples into account, but actually and more importantly, this information is potentially helpful in making predictions, since it suggests pooling information from neighboring samples (nodes). Our goal in this paper is to take advantage of the network structure knowledge between samples in order to better do variable selection and parameter estimation simultaneously for high-dimensional linear regression.

In the past decades, there have been a large body of work focusing on analyzing the network structure implied by the relational data alone. For example, Cabreros and Tsirigos (2015) studied the community structures in Hi-C genomic data; Brian, Brian and Mark (2011) proposed an efficient method for detecting communities in network; and Abbe (2017) studied the detection community in stochastic. There are also much work considering regression with dependent observations in certain contexts. For example, following the concepts initially discussed by Manski (1993) in econometrics, Djebbari and Fortin (2009) proposed an auto-regressive model, which assuming the responses are auto-regressive. Based on these work, Lee (2007) further considered the identification and estimation of structural interaction effects in a social interaction model. Such models assumed specific forms of different types of network effects, namely, endogenous effects, exogenous effects and correlated effects, and most of these literatures were focused on identifiability of such effects. On the other hand, these methods had mainly been used to identify social effects defined within a very special and difficult to verify model, without a focus on interpretability or good prediction performance. Recently, there are some focus on developing a general statistical framework for using network data in prediction. Zhu, Li and Levina (2018) proposed a prediction model for network-linked data. Specially, they introduced the network-based penalty in individual mode effects to encourage similarity between predictors for linked nodes and showed that incorporating it into prediction leads to improvement over traditional models both theoretically and empirically. Their approach can be viewed as a regression version of point estimation problem discussed in Sharpnack and Krishnamurthy (2013); and Wang and Tibshirani (2016). Alternatively, it can be view in a Bayesian framework, as regression with a Gaussian Marlov random filed prior over the network. However, including naerby responses as covariates in linear regression has not yet been well studied. Hence, in this paper, we take into the network structure between samples account to linear regression for high-dimensional data. Specially, we propose the Network Lasso, which use nearby responses to build the network information and to better predict response. We assume each response is a linear function of its covariates and the neighbors' responses. In the high-dimensional regime, we use the l_1 penalized approach to induce the sparsity of parameters which in turn can select important covariates.

The rest of this paper is organized as follows. In Section 2, we first briefly review the notations of this paper and introduce the linear regression with independent samples. Then using network structure information, we propose the Network Lasso model and further give its optimization approach. The theoretical analysis of the proposed model are established in Section 3. In Section 4, we numerically evaluate the selection performance of our proposed method and gains in selection accuracy compared with Lasso without using network information. Section 5 includes the application to real data analysis. We conclude the paper in Section 6.

2. Methodology

In this section, we firstly give some notations. Then we introduce our method in the setting of linear regression.

2.1. Set-up and notation

We start from setting up notation. The data consist of n observations (y_1, x_1), $..., (y_n, x_n)$, where $y_i \in R$ is the response variable and $x_i \in R^p$ is the vector of covariates for observation i. We write $Y = (y_1, y_2, ..., y_n)^T$ for the response vector, and $X = (x_1, x_2, ..., x_n)^T$ for the $n \times p$ design matrix. We treat X as fixed and assume its columns have been standardized to have mean 0 and variance 1. We also observe the network connecting the observations, and denote it $G = (V, E)$, where $V = \{1, 2, ..., n\}$ is the node set of the graph, and $E \subset V \times V$ is the edge set. We represent the graph by its adjacency matrix $A \in R^{n \times n}$, where $A_{vu} = 1$ if $(v, u) \in E$ and 0 otherwise.

2.2. Linear regression with network information

The general linear regression model for each sample is formulated as follows:

$$y_i = x_i'\beta + \epsilon_i, i = 1, 2, .., n.$$

where $\{y_i\}_{i=1}^n$ are responses, $\{x_i\}_{i=1}^n$ are covariates, $\{\epsilon\}_{i=1}^n$ are error terms. $\beta = (\beta_1, \beta_2, ..., \beta_p) \in R^p$ is the set of unknown parameter that need to be estimated. Different from the traditional setting, in this paper, we assume the n samples are linked by a network $G = (V, E)$. Therefore, considering the network structure information, we model each y_i as a linear combination of the covariates x_i, the connected neighbors $y_j (j \in M, M = \{Ve_i\})$, and an error term ϵ_i. Specifically, we propose the following model:

$$y_i = x_i'\beta + \frac{\alpha}{n_i}\sum_{j \in M} y_j + \epsilon_i, i = 1, 2, ..., n, \tag{1}$$

where α is the neighborhoods effect coefficient and $n_i = |M|$. Accordingly, the parameters can be estimated as the following optimization problem,

$$\min_\beta \frac{1}{2n}\sum_{i=1}^n \left(y_i - \left(x_i'\beta + \frac{\alpha}{n_i}\sum_{j \in M} y_j\right)\right)^2 + \lambda\|\beta\|_1. \tag{2}$$

Rewriting the (1) as following form:

$$Y = X\beta + AY\alpha + \epsilon$$

where $Y = (y_1, y_2, ..., y_n)^T$ is the response vector. $X = (x_1, x_2, ..., x_n)^T$ is the $n \times p$ covariates matrix. $A = (a_{ij}) \in R^{n \times n}$ is the adjacency matrix including the network information, which represents the information between samples and the element a_{ij} is neither 1 or 0. If $a_{ij} = 1$, Y_i is a neighbor of Y_j; otherwise Y_i is not a neighbor of Y_j. $\epsilon = (\epsilon_1, \epsilon_2, ..., \epsilon_n)$ is the error term, and $\epsilon_i \sim_{i.i.d} N(0, \sigma^2), i = 1, 2, ..., n$.

Furthermore, the parameter β can be estimated as

$$\hat{\beta} \in \arg\min_{\beta} \frac{1}{2n} \|Y - AY\alpha - X\beta\|_2^2 + \lambda\|\beta\|_1, \tag{3}$$

For model (3), it can be solved efficiently by the coordinate descent algorithm (see Guo and Zhang 2020).

3. Theoretical analysis

We first derive the bound of the proposed estimator (3) and the Lasso estimator to show the gain in estimation accuracy. For simplicity, we consider the oracle setting where the neighborhoods effect coefficient α is known, denoted by α^0. Hence we only need to estimate β. Denote our estimator and the lasso estimator as

$$\hat{\beta} \in \arg\min_{\beta} \frac{1}{2n} \|Y - AY\alpha^0 - X\beta\|_2^2 + \lambda\|\beta\|_1$$

and

$$\hat{\beta}^{lasso} \in \arg\min_{\beta} \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda_{lasso}\|\beta\|_1$$

respectively. As in most high-dimensional problems, to ensure the model identifiability and to enhance the model fitting accuracy and interpretability, the true regression coefficient vector β^0 is commonly imposed to be sparse with only a small proportion (see Belloni and Chernozhukov 2009; Fan, Fan and Baryt 2012). Denoting the number of nonzero elements of the true regression coefficients by s, and we write $S = \{j : \beta_j \neq 0, 1 \leq j \leq p\}$, then $s = |S|$ and we can write $\beta = (\beta_S, \beta_{S^c})$.

And in order to establish the estimation consistency, we assume the restricted eigenvalue condition hold, i.e.,

Assumption 1. For a given p-dimension vector u, the design matrix X satisfies restricted eigenvalue condition with parameter $\nu > 0$ if

$$\inf\left\{\frac{u^{\mathrm{T}}\left(\frac{X^{\mathrm{T}}X}{n}\right)u}{u^{\mathrm{T}}u} : 3\|u_S\|_1 \geq \|u_{S^c}\|_1\right\} \geq \nu > 0. \tag{A1}$$

Under assumption $A1$, we have the main results as follows,

Theorem 1. *Suppose the design matrix is normalized such that* $\|X_j\|_2^2 = n$. *Based on* (Negahban et al. 2012), *let*

$$\lambda_0 = 2\sigma\sqrt{\frac{\log p}{n}}.$$

When $\lambda \geq 2\lambda_0$, then with probability at least $1 - \frac{2}{p}$,

$$\|\hat{\beta} - \beta^0\|_2 \leq \frac{3\lambda\sqrt{s}}{\nu},$$

and when $\lambda_{\text{lasso}} \geq 2(\|(\alpha^0 AY)^{\mathrm{T}}X\|_\infty + \lambda_0)$, then with probability at least $1 - \frac{2}{p}$,

$$\|\hat{\beta}^{\text{lasso}} - \beta^0\|_2 \leq \frac{3\lambda_{\text{lasso}}\sqrt{s}}{\nu}.$$

Proof. The proof consists of two parts, i.e., the derivation for the bound of the $\hat{\beta}$ and $\hat{\beta}^{\text{lasso}}$, respectively.

Part I: The bound of our estimator $\hat{\beta}$.

By definition, we have

$$\frac{1}{2n}\|Y - X\hat{\beta} - \alpha^0 AY\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|Y - X\beta^0 - \alpha^0 AY\|_2^2 + \lambda\|\beta^0\|_1.$$

Using the fact that $Y = X\beta^0 + AY\alpha^0 + \epsilon$, we further have,

$$\frac{1}{2n}\|X\hat{\beta} - X\beta^0\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{n}|\epsilon^{\mathrm{T}}X(\hat{\beta} - \beta^0)| + \lambda\|\beta^0\|_1, \tag{5}$$

which is exactly the same with the display of the lasso under the null model ($A = 0$). For the sake of completeness, we next follow the lines of the standard lasso to prove. For any two vectors a and b, $|a^{\mathrm{T}}b| \leq \|a\|_\infty\|b\|_1$, thus (5) implies

$$\frac{1}{2n}\|X\hat{\beta} - X\beta^0\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{n}\|\epsilon^{\mathrm{T}}X\|_\infty\|\hat{\beta} - \beta^0\|_1 + \lambda\|\beta^0\|_1. \tag{6}$$

Note that $\epsilon^{\mathrm{T}}X_j \sim N(0, \sigma^2\|X_j\|_2^2)$, and $\|X_j\|_2^2 = n$, and for standard gaussian random variable z and any nonnegative t,

$$P(|z| > t) \leq 2\exp\left(-\frac{t^2}{2}\right),$$

so bring together these pieces and use the union bound, we have

$$P\left(\frac{1}{n}\|\epsilon^{\mathrm{T}}X\|_\infty \geq \lambda_0\right) \leq \frac{2}{p},$$

with $\lambda_0 = 2\sigma\sqrt{\frac{\log p}{n}}$. Then if $\lambda \geq 2\lambda_0$, then with probability larger than $1 - \frac{2}{p}$,

$$\frac{1}{2n}\|X\hat{\beta} - X\beta^0\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{\lambda}{2}\|\hat{\beta} - \beta^0\|_1 + \lambda\|\beta^0\|_1. \tag{7}$$

By the following two facts,

$$\|\hat{\beta}\|_1 = \|\hat{\beta}_S\|_1 + \|\hat{\beta}_{S^c}\|_1 \geq \|\beta_S^0\|_1 - \|\hat{\beta}_S - \beta_S^0\|_1 + \|\hat{\beta}_{S^c}\|_1,$$

and

$$\|\hat{\beta} - \beta^0\|_1 = \|\hat{\beta}_S - \beta_S^0\|_1 + \|\hat{\beta}_{S^c} - \beta_{S^c}^0\|_1,$$

we then have

$$\begin{aligned}\frac{1}{2n}\|X\hat{\beta} - X\beta^0\|_2^2 &+ \lambda\left(\|\beta_S^0\|_1 - \|\hat{\beta}_S - \beta_S^0\|_1 + \|\hat{\beta}_{S^c}\|_1\right) \\ &\leq \frac{\lambda}{2}\left(\|\hat{\beta}_S - \beta_S^0\|_1 + \|\hat{\beta}_{S^c} - \beta_{S^c}^0\|_1\right) + \lambda\|\beta^0\|_1,\end{aligned} \tag{8}$$

which can be simplified as

$$\frac{1}{2n}\|X\hat{\beta} - X\beta^0\|_2^2 + \frac{\lambda}{2}\|\hat{\beta}_{S^c}\|_1 \leq \frac{3\lambda}{2}\|\hat{\beta}_S - \beta_S^0\|_1. \tag{9}$$

From (9), we have two findings. On the one hand,

$$\|\hat{\beta}_{S^c}\|_1 \leq 3\|\hat{\beta}_S - \beta_S^0\|_1, \text{i.e.,} \|\hat{\beta}_{S^c} - \beta_{S^c}^0\|_1 \leq 3\|\hat{\beta}_S - \beta_S^0\|_1, \tag{10}$$

which implies

$$\|\hat{\beta} - \beta^0\|_1 \leq 4\|\hat{\beta}_S - \beta_S^0\|_1 \leq 4\sqrt{s}\|\hat{\beta}_S - \beta_S^0\|_2 \leq 4\sqrt{s}\|\hat{\beta} - \beta^0\|_2. \tag{11}$$

On the other hand,

$$\frac{1}{n}\|X\hat{\beta} - X\beta^0\|_2^2 \leq 3\lambda\|\hat{\beta}_S - \beta_S^0\|_1 \leq 3\lambda\sqrt{s}\|\hat{\beta} - \beta^0\|_2, \tag{12}$$

where the last inequality follows from (11). In addition,

$$\begin{aligned}\frac{1}{n}\|X\hat{\beta} - X\beta^0\|_2^2 &= \frac{1}{n}\|X(\hat{\beta} - \beta^0)\|_2^2 \\ &= \frac{1}{n}(\hat{\beta} - \beta^0)^T X^T X(\hat{\beta} - \beta^0) \\ &= (\hat{\beta} - \beta^0)^T \frac{X^T X}{n}(\hat{\beta} - \beta^0)\end{aligned}$$

As (10) and (A1), then

$$\inf\left\{\frac{(\hat{\beta} - \beta^0)^T \frac{X^T X}{n}(\hat{\beta} - \beta^0)}{(\hat{\beta} - \beta^0)^T(\hat{\beta} - \beta^0)} : 3\|\hat{\beta}_S - \beta_S^0\| \geq \|\hat{\beta}_{S^c} - \beta_{S^c}^0\|\right\} \geq \nu$$

Hence,

$$\frac{1}{n}\|X\hat{\beta} - X\beta^0\|_2^2 \geq \nu\|\hat{\beta} - \beta^0\|_2^2. \tag{13}$$

Finally, combine (12) with (13), we arrive the result that

$$\|\hat{\beta} - \beta^0\|_2 \leq \frac{3\lambda\sqrt{s}}{\nu},$$

with probability larger than $1 - \frac{2}{p}$.

Part II: The bound of the lasso estimator $\hat{\beta}^{\text{lasso}}$.

By definition, we have

$$\frac{1}{2n}\left\|Y - X\hat{\beta}^{\text{lasso}}\right\|_2^2 + \lambda_{\text{lasso}}\left\|\hat{\beta}^{\text{lasso}}\right\|_1 \leq \frac{1}{2n}\|Y - X\beta^0\|_2^2 + \lambda_{\text{lasso}}\|\beta^0\|_1.$$

Using the fact that $Y = X\beta^0 + AY\alpha^0 + \epsilon$, we have,

$$\frac{1}{2n}\left\|X\beta^0 + AY\alpha^0 + \epsilon - X\hat{\beta}^{\text{lasso}}\right\|_2^2 + \lambda_{\text{lasso}}\left\|\hat{\beta}^{\text{lasso}}\right\|_1 \leq \frac{1}{2n}\|AY\alpha^0 + \epsilon\|_2^2 + \lambda\|\beta^0\|_1.$$

Applying the similar arguments as in **Part I**, we then have

$$\frac{1}{2n}\left\|X\beta^0 - X\hat{\beta}^{\text{lasso}}\right\|_2^2 + \lambda_{\text{lasso}}\left\|\hat{\beta}^{\text{lasso}}\right\|_1 \leq \frac{1}{n}\|(AY\alpha^0+\epsilon)^{\mathrm{T}}X\|_\infty\left\|\hat{\beta}^{\text{lasso}} - \beta_0\right\|_1 + \lambda_{\text{lasso}}\|\beta_0\|_1.$$

Note that with probability larger than $1-\frac{2}{p}$,

$$\|(AY\alpha^0+\epsilon)^{\mathrm{T}}X\|_\infty \leq \|AY\alpha^0 X\|_\infty + \lambda_0,$$

where $\lambda_0 = 2\sigma\sqrt{\frac{\log p}{n}}$. Follow the same lines as in **Part I**, if

$$\lambda_{\text{lasso}} \geq \frac{2}{n}\left(\|(\alpha^0 AY)^{\mathrm{T}}X\|_\infty + \lambda_0\right),$$

then with probability at least $1-\frac{2}{p}$,

$$\left\|\hat{\beta}^{\text{lasso}} - \beta^0\right\|_2 \leq \frac{3\lambda_{\text{lasso}}\sqrt{s}}{\nu}.$$

Now we attempt to derive an upper bound for $\|(AY)^{\mathrm{T}}X\|_\infty$. By the basic inequalities,

$$\|(AY)^{\mathrm{T}}X\|_\infty = \max_j|(AY)^{\mathrm{T}}X_j| \leq \max_j\|AY\|_2\|X_j\|_2 = \sqrt{n}\|AY\|_2 \leq \sqrt{n}\sqrt{\sigma_{\max(A^{\mathrm{T}}A)}}\|Y\|_2,$$

where $\sigma_{\max}(\cdot)$ denotes the maximum eigenvalue of a matrix. Next we focus on bounding $\|Y\|_2$. Recall that

$$Y = X\beta^0 + AY\alpha^0 + \epsilon, \epsilon \sim N_n(0,\sigma^2 I),$$

where n is the number of samples. Hence after some calculations we have,

$$Y = (I-\alpha^0 A)^{-1}X\beta^0 + (I-\alpha^0 A)^{-1}\epsilon.$$

As a result, conditioning on X,

$$Y \sim N_n\left((I-\alpha^0 A)^{-1}X\beta^0, \sigma^2(I-\alpha^0 A)^{-2}\right).$$

Then,

$$Y - (I-\alpha^0 A)^{-1}X\beta^0 \sim N_n\left(0, \sigma^2(I-\alpha^0 A)^{-2}\right),$$

and then,

$$\frac{(I-\alpha^0 A)Y}{\sigma} - \frac{X\beta^0}{\sigma} \sim N_n(0, I).$$

And thus,

$$\left\|\frac{(I-\alpha^0 A)Y}{\sigma} - \frac{X\beta^0}{\sigma}\right\|_2^2 \sim \mathcal{X}_n^2,$$

where $\mathcal{X}_n^2$ denotes the chi-squared distribution with degree n. By the usual chi-squared Chernoff bound,

$$P(\mathcal{X}_n^2 \geq tn) \leq \exp\left\{-\frac{n}{2}(t - \log t - 1)\right\},$$

for any $t > 0$. Thus with probability larger than $1 - \exp\{-\frac{n}{2}(t - \log t - 1)\}$,

$$\left\| \frac{(I-\alpha^0 A)Y}{\sigma} - \frac{X\beta^0}{\sigma} \right\|_2^2 \le tn.$$

By the triangle inequality,

$$\left\| \frac{(I-\alpha^0 A)Y}{\sigma} \right\|_2^2 \le \frac{\|X\beta^0\|_2^2}{\sigma^2} + tn. \tag{14}$$

Use the definition of the minimum eigenvalue of a matrix, we then have,

$$\left\| \frac{(I-\alpha^0 A)Y}{\sigma} \right\|_2^2 \ge \frac{\|Y\|_2^2}{\sigma^2} \sigma_{\min}((I-\alpha^0 A)^{\mathrm{T}}(I-\alpha^0 A)), \tag{15}$$

where $\sigma_{\min}(\cdot)$ denotes the minimum eigenvalue of a matrix (here $I-\alpha^0 A$ is invertible with carefully chosen α^0). Combine (14) with (15), we obtain,

$$\|Y\|_2 \le \sqrt{\frac{\|X\beta^0\|_2^2 + tn\sigma^2}{\sigma_{\min}((I-\alpha^0 A)^{\mathrm{T}}(I-\alpha^0 A))}}. \tag{16}$$

If we choose $t = \frac{2\log p}{n}$, then (16) holds with probability larger than $1 - \frac{1}{p} \cdot (\frac{2e\log p}{n})^{n/2}$, which converges to 1 as p and n increase if say $p = O_p(e^{n^\kappa})$ for $0 < \kappa < 1$. Recall that if

$$\lambda_{\text{lasso}} \ge 2(\|(\alpha^0 AY)^{\mathrm{T}} X\|_\infty + \lambda_0),$$

then with probability at least $1 - \frac{2}{p}$,

$$\left\| \hat{\beta}^{\text{lasso}} - \beta^0 \right\|_2 \le \frac{3\lambda_{\text{lasso}}\sqrt{s}}{\nu}.$$

Consequently,

$$\left\| \hat{\beta}^{\text{lasso}} - \beta^0 \right\|_2 \le 6 \left(\alpha^0 \sqrt{n\sigma_{\max}(A^{\mathrm{T}}A) \left(\frac{\|X\beta^0\|_2^2 + (2\log p)\sigma^2}{\sigma_{\min}\left((I-\alpha^0 A)^{\mathrm{T}}(I-\alpha^0 A)\right)} \right)} + 2\sigma\sqrt{\frac{\log p}{n}} \right) \frac{\sqrt{s}}{\nu}, \tag{17}$$

and (17) holds with probability at least $1 - \frac{1}{p} \cdot (\frac{2e\log p}{n})^{n/2} - \frac{2}{p}$, which converges to 1.

In the preceding paragraphs, we show the gain in estimation accuracy of our estimator compared with the lasso under the oracle setting where the neighborhood effect coefficient is known. The main reason that we consider this setting is to make sure the assumption $(A1)$ is the same for the two estimators. Now we turn to the general setting where α^0 is unknown to justify the estimation bound of our estimator. We firstly provide the modified restricted eigenvalue condition and then give the estimation bound of our proposed estimator. We assume the modified restricted eigenvalue condition hold, i.e.,

$$\liminf_{n\to\infty} \left\{ \frac{u^T \left(X^T \left(I - \frac{AY(AY)^T}{\|AY\|_2^2} \right) X \right) u}{nu^T u} : 3\|u_S\|_1 + o_p(1) \ge \|u_{S^c}\|_1 \right\} \ge \nu > 0, \tag{B1}$$

then the result follows.

Theorem 2. *Suppose the design matrix is normalized such that* $\|X_j\|_2^2 = n$, *and let*

$$\lambda_0 = 2\sigma\sqrt{\frac{\log p}{n}}.$$

When $\lambda \geq 2\lambda_0$, then with probability at least $1 - \frac{2}{p} - \frac{(e)^{\frac{u}{2}}(\frac{2\log p}{n})^{\frac{u}{2}}}{p}$ and sufficiently large n,

$$\|\hat{\beta} - \beta^0\|_2 = O_p\left(\frac{\lambda\sqrt{s}}{\nu}\right).$$

Remark. The condition $(B1)$ is stronger than condition $A1$ in the sense that, for any vector u,

$$u^{\mathrm{T}}\left(X^{\mathrm{T}}\left(I - \frac{AY(AY)^{\mathrm{T}}}{\|AY\|_2^2}\right)X\right)u = (Xu)^{\mathrm{T}}\left(I - \frac{AY(AY)^{\mathrm{T}}}{\|AY\|_2^2}\right)Xu \leq (Xu)^{\mathrm{T}}Xu = u^{\mathrm{T}}X^{\mathrm{T}}Xu,$$

where we use the fact that $I - \frac{AY(AY)^{\mathrm{T}}}{\|AY\|_2^2}$ is an idempotent matrix with eigenvalues 1 ($n-1$ multiplicities) and 0 (1 multiplicity).

Proof. By definition, we have

$$\frac{1}{2n}\|Y - X\hat{\beta} - \hat{\alpha}AY\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|Y - X\beta^0 - \alpha^0 AY\|_2^2 + \lambda\|\beta^0\|_1.$$

Using the fact that $Y = X\beta^0 + AY\alpha^0 + \epsilon$, we further have,

$$\frac{1}{2n}\|X\beta^0 + \alpha^0 AY - X\hat{\beta} - \hat{\alpha}AY + \epsilon\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|\epsilon\|_2^2 + \lambda\|\beta^0\|_1, \tag{18}$$

which is equivalent to

$$\frac{1}{2n}\|X\beta^0 + \alpha^0 AY - X\hat{\beta} - \hat{\alpha}AY\|_2^2 + \frac{1}{n}\epsilon^{\mathrm{T}}(X\beta^0 + \alpha^0 AY - X\hat{\beta} - \hat{\alpha}AY) + \lambda\|\hat{\beta}\|_1 \leq \lambda\|\beta^0\|_1. \tag{19}$$

Now we use $\hat{\alpha} = ((AY)^{\mathrm{T}}AY)^{-1}(AY)^{\mathrm{T}}(Y - X\hat{\beta})$ to simplify (19). Note that

$$X\beta^0 + \alpha^0 AY - X\hat{\beta} - \hat{\alpha}AY = \left(I - \frac{AY(AY)^{\mathrm{T}}}{\|AY\|_2^2}\right)X(\beta^0 - \hat{\beta}) - \frac{AY(AY)^{\mathrm{T}}}{\|AY\|_2^2}\epsilon,$$

then (19) implies

$$\frac{1}{2n}\left\|(I - C)X\left(\beta^0 - \hat{\beta}\right)\right\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\epsilon^{\mathrm{T}}C\epsilon + \frac{1}{n}\|\epsilon^{\mathrm{T}}(I - C)X\|_\infty\|\beta^0 - \hat{\beta}\|_1 + \lambda\|\beta^0\|_1, \tag{20}$$

where we denote $\frac{AY(AY)^{\mathrm{T}}}{\|AY\|_2^2}$ as C for the sake of simplicity. Noting that the eigenvalues of C and I – C are both 0 and 1 (but with different multiplicities), and using the tail bound of gaussian variable, (20) implies

$$\frac{1}{2n}\|(I - C)X(\beta^0 - \hat{\beta})\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|\epsilon\|_2^2 + \frac{1}{n}\lambda_0\|\beta^0 - \hat{\beta}\|_1 + \lambda\|\beta^0\|_1, \tag{21}$$

holds with probability larger than $1-\frac{2}{p}$, where $\lambda_0 = 2\sigma\sqrt{\frac{\log p}{n}}$. Different from the derivation of the standard lasso, an extra term $\frac{1}{2n}\|\epsilon\|_2^2$ appears in the right hand side of (21). So before we continue, we first bound $\frac{1}{2n}\|\epsilon\|_2^2$. Note that

$$\frac{\|\epsilon\|_2^2}{\sigma^2} \sim \mathcal{X}_n^2,$$

where $\mathcal{X}_n^2$ denotes the chi-squared distribution with degree n. By the usual chi-squared Chernoff bound,

$$P(\mathcal{X}_n^2 \geq tn) \leq \exp\left\{-\frac{n}{2}(t - \log t - 1)\right\},$$

for any $t > 0$. Therefore, if we choose $t = \frac{2\log p}{n}$, then

$$\frac{1}{2n}\|\epsilon\|_2^2 \leq \sigma^2 \frac{\log p}{n},$$

with probability larger than $1 - \frac{(e)^{\frac{n}{2}}(\frac{2\log p}{n})^{\frac{n}{2}}}{p}$, which converges to 1 if p increases exponentially with n, and $\frac{\log p}{n}$ converges to 0, i.e., $\frac{1}{2n}\|\epsilon\|_2^2 = O_p\left(\frac{\log p}{n}\right)$. Hence with large probability, (21) implies

$$\frac{1}{2n}\|(I-C)X(\beta^0 - \hat{\beta})\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq O_p\left(\frac{\log p}{n}\right) + \frac{1}{n}\lambda_0\|\beta^0 - \hat{\beta}\|_1 + \lambda\|\beta^0\|_1.$$

Follow the same lines as the standard lasso, when $\lambda \geq 2\lambda_0$,

$$\frac{1}{2n}\|(I-C)X(\beta^0 - \hat{\beta})\|_2^2 + \frac{\lambda}{2}\|\hat{\beta}_{S^c}\|_1 \leq O_p\left(\frac{\log p}{n}\right) + \frac{3\lambda}{2}\|\hat{\beta}_S - \beta_S^0\|_1. \tag{22}$$

From (22), we have two findings. On the one hand,

$$\|\hat{\beta}_{S^c}\|_1 \leq 3\|\hat{\beta}_S - \beta_S^0\|_1 + O_p\left(\frac{\log p}{n}\right), \text{i.e.,} \|\hat{\beta}_{S^c} - \beta_{S^c}^0\|_1 \leq 3\|\hat{\beta}_S - \beta_S^0\|_1 + O_p\left(\frac{\log p}{n}\right), \tag{23}$$

which implies

$$\begin{aligned}\|\hat{\beta} - \beta^0\|_1 &\leq 4\|\hat{\beta}_S - \beta_S^0\|_1 + O_p\left(\frac{\log p}{n}\right) \leq 4\sqrt{s}\|\hat{\beta}_S - \beta_S^0\|_2 + O_p\left(\frac{\log p}{n}\right) \\ &\leq 4\sqrt{s}\|\hat{\beta} - \beta^0\|_2 + O_p\left(\frac{\log p}{n}\right).\end{aligned} \tag{24}$$

On the other hand,

$$\frac{1}{n}\|(I-C)(X\hat{\beta} - X\beta^0)\|_2^2 \leq 3\lambda\|\hat{\beta}_S - \beta_S^0\|_1 + O_p\left(\frac{\log p}{n}\right) \leq 3\lambda\sqrt{s}\|\hat{\beta} - \beta^0\|_2 + O_p\left(\frac{\log p}{n}\right), \tag{25}$$

where the last inequality follows from (24). In addition, use the restricted eigenvalue condition $(B1)$ by noting (24), for sufficiently large n, we have

$$\frac{1}{n}\|(I-C)(X\hat{\beta}-X\beta^0)\|_2^2 \geq \frac{\nu}{2}\|\hat{\beta}-\beta^0\|_2^2. \tag{26}$$

Combine (25) with (26), we have

$$\frac{\nu}{2}\|\hat{\beta}-\beta^0\|_2 - 3\lambda\sqrt{s}\|\hat{\beta}-\beta^0\|_2^2 - O_p\left(\frac{\log p}{n}\right) \leq 0. \tag{27}$$

The left hand side of (27) can be regarded as a quadratic function of $\|\hat{\beta}-\beta^0\|_2$, and it is easy to see that $\|\hat{\beta}-\beta^0\|_2$ is smaller than the bigger root of the quadratic function, that is,

$$\|\hat{\beta}_j-\beta_j^0\|_2 \leq \frac{3\lambda\sqrt{s}+\sqrt{9\lambda^2 s+O_p\left(\frac{\log p}{n}\right)}}{\nu}.$$

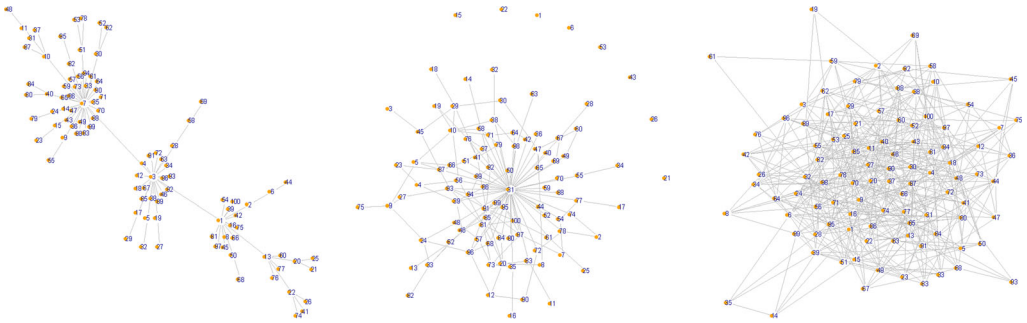
Finally, as $\lambda \geq 2\lambda_0 = 2\sqrt{\frac{\sigma\log p}{n}}$, we arrive the conclusion that

$$\|\hat{\beta}-\beta^0\|_2 = O_p\frac{\lambda\sqrt{s}}{\nu},$$

with probability at least $1-\frac{2}{p}-\frac{(e)^{\frac{n}{2}}(\frac{2\log p}{n})^{\frac{n}{2}}}{p}$.

4. Simulation

To demonstrate the finite sample performance of the proposed model, we present three examples in this section. The main difference is the generating mechanism of the adjacency matrix A. Specially, A is Scale-free network, Hub network and Erdős-Renyi network, respectively, as shown in Figure 1. And for each example, the parameters are basically set $n=100$ and $p=50$, 100, 150 respectively. And the random error ϵ_i is simulated from a standard normal distribution $N(0,1)$, and the covariate x_i is generated from a multivariate normal distribution $N(0,\Sigma)$, where $\Sigma=(\sigma_{ij}), \sigma_{ij}=0.5^{|i-j|}$. The following two model parameters are considered: (1) $\beta=(0.5,1,0.8,0.2,0.3,\underbrace{0.5,\ldots 0.5}_{(20)},\underbrace{0,\ldots 0}_{p-25})$, (2) $\beta=(U(0,2,20),\underbrace{0,\ldots 0}_{p-20})$. Then Y can be generated according to



(a) (b) (c)

Figure 1. Three network graph with $n=100$. (a) Scale-Free network. (b) Hub network. (c) Erdős-Renyi.

$$Y = X\beta + AY\alpha + \epsilon,$$

where $\alpha = 2$. For comparison, the performance of our method and Lasso are evaluated.

T: the total number of true $\beta_0 \neq 0$; P: the total number of estimated $\hat{\beta} \neq 0$.

TP: the number of $(\beta_0 \neq 0)$ equals $(\hat{\beta} \neq 0)$; FP: the number of $(\beta_0 = 0)$ equals $(\hat{\beta} \neq 0)$.

TN: the number of $(\beta_0 = 0)$ equals $(\hat{\beta} = 0)$; FN: the number of $(\beta_0 \neq 0)$ equals $(\hat{\beta} = 0)$.

- TPR = TP/T; FPR = FP/F
- L_2 norm/loss, which is defined as $||\hat{\beta} - \beta_0||^2$.
- F_1=2*P*R/(P + R), where P = TP/(TP + FP), R = TP/(TP + FN).

Each curve is smoothed over 5 replications.

(a) In this part, we generate the model parameter β as (1), and the corresponding simulation results and analysis are summarized as follows.

Figures 2, 4 and 6 show the effect of the number of covariates p by ROC curves of the Scare-free network, Hub network and Erdős-Renyi network, respectively. As one can expect, the performance of the proposed model is superior to Lasso. Specially, Figure 2 shows that Network Lasso performs best when $p = 100$, and the results of Lasso are almost same as p increased. More importantly, we can see that the proposed model performs much better than Lasso no matter whether $p = 50$, 100 or $p = 150$. For Hub and Erdős-Renyi network, from Figures 4 and 6, it is easy to find that the performances both of Lasso and Network Lasso have no obvious changes with p increased. But as a whole, the proposed model outperforms Lasso in terms of ROC.

Figures 3, 5 and 7 show the effect of the number of covariates p by L_2 norm with varying FPR for the Scare-free network, Hub network and Erdős-Renyi network, respectively. For Scale-free network, Figure 3 shows that the results of Network Lasso is almost invariable with p increased. Lasso has the best performance when $p = 50$, which the value of L_2 is much larger than Network Lasso. For Hub and Erdős-Renyi network, we have similar found. Hence, from these three Figures, we can conclude that the proposed model performs much better than Lasso in term of L_2 norm.

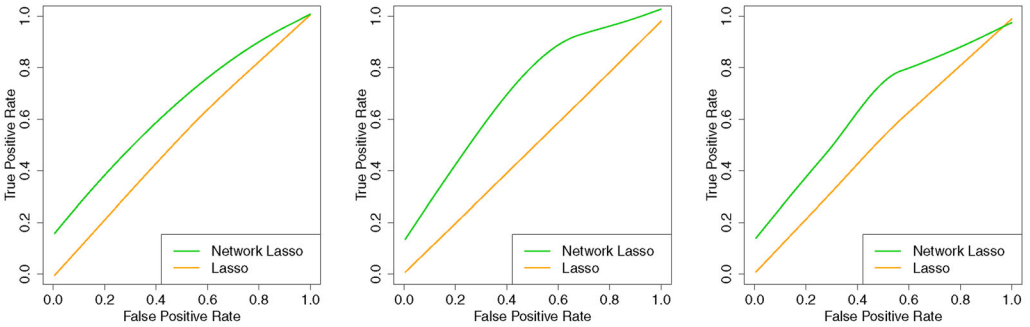


Figure 2. ROC curves of Lasso and Network Lasso for Scale-free network. (a) $p = 50$. (b) $p = 100$. (c) p = 150.

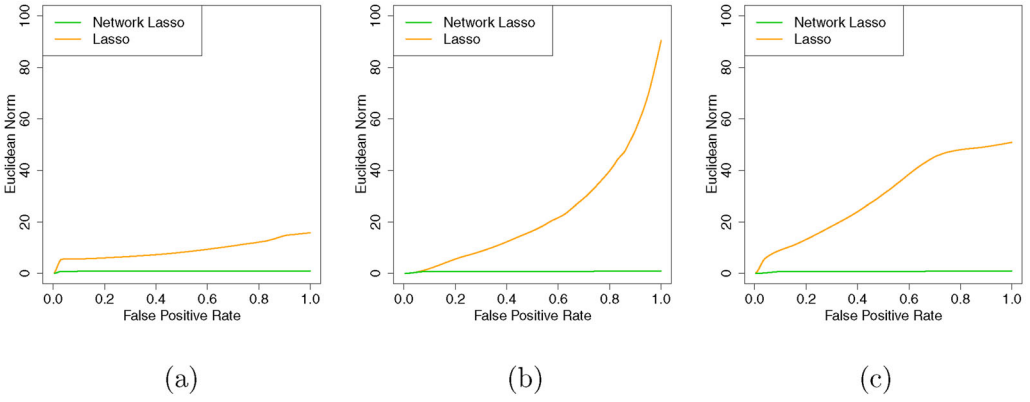


Figure 3. Performance is evaluated by the L_2-norm of β with varying FPR for Scale-free network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

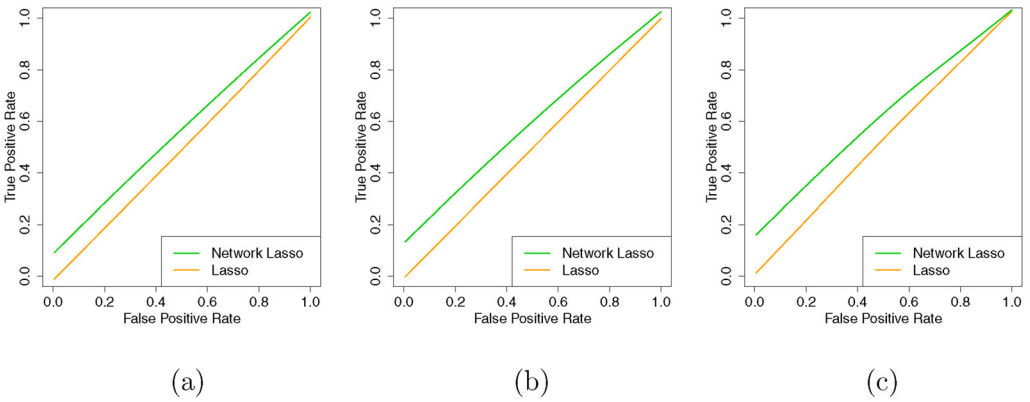


Figure 4. ROC curves of Lasso and Network Lasso for Hub network. (a) $p = 50$. (b) $p = 100$. (*c*) p = 150.

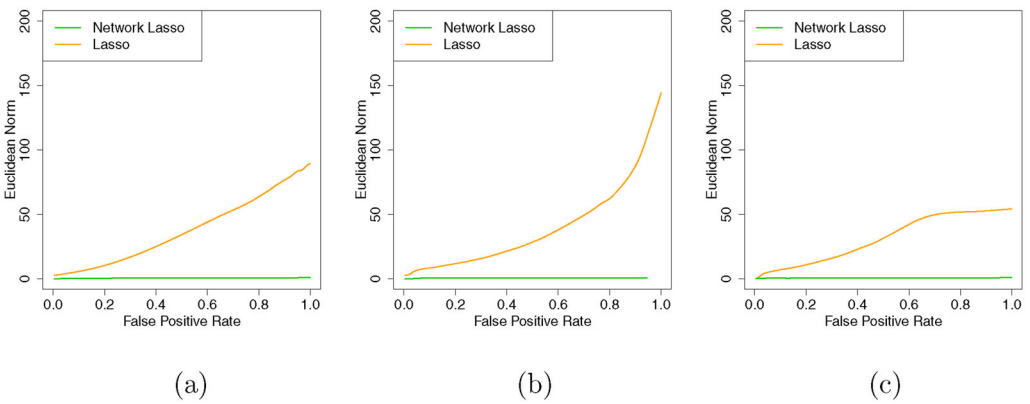


Figure 5. Performance is evaluated by the L_2-norm of β with varying FPR for Hub network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

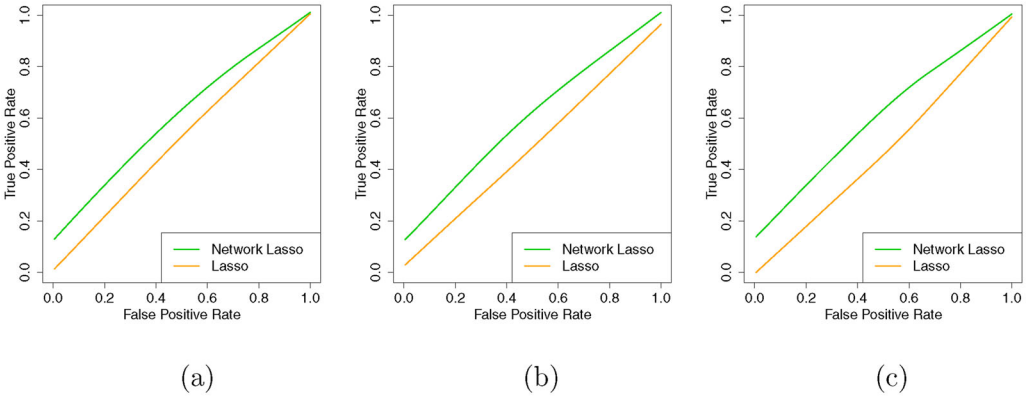


Figure 6. ROC curves of Lasso and Network Lasso for Erdős-Renyi network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

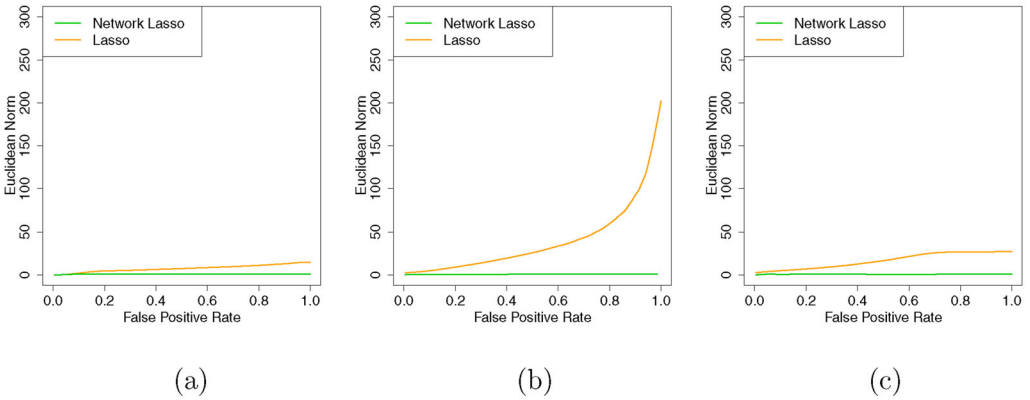


Figure 7. Performance is evaluated by the L_2-norm of β with varying FPR for Erdős-Renyi. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

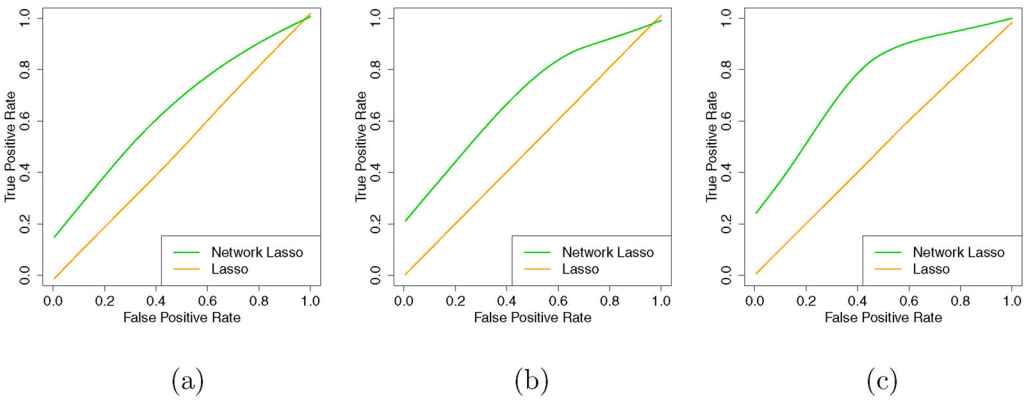


Figure 8. ROC curves of Lasso and Network Lasso for Scale-free network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

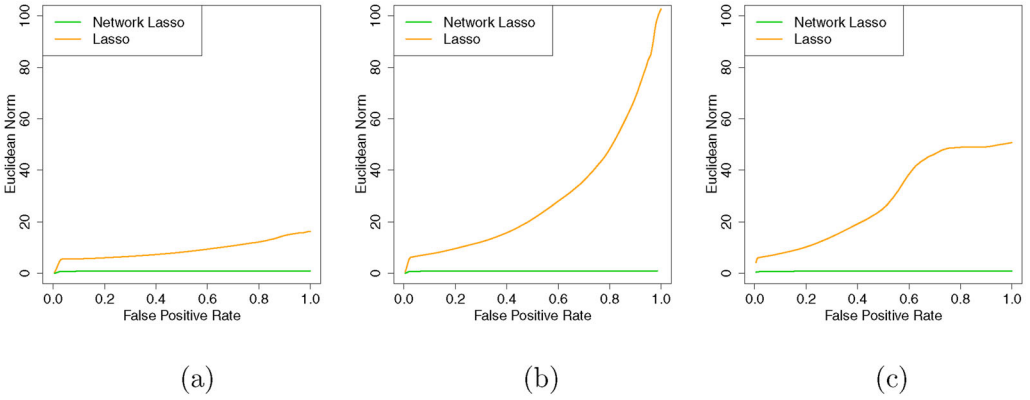


Figure 9. Performance is evaluated by the L_2-norm of β with varying FPR for Scale-free network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

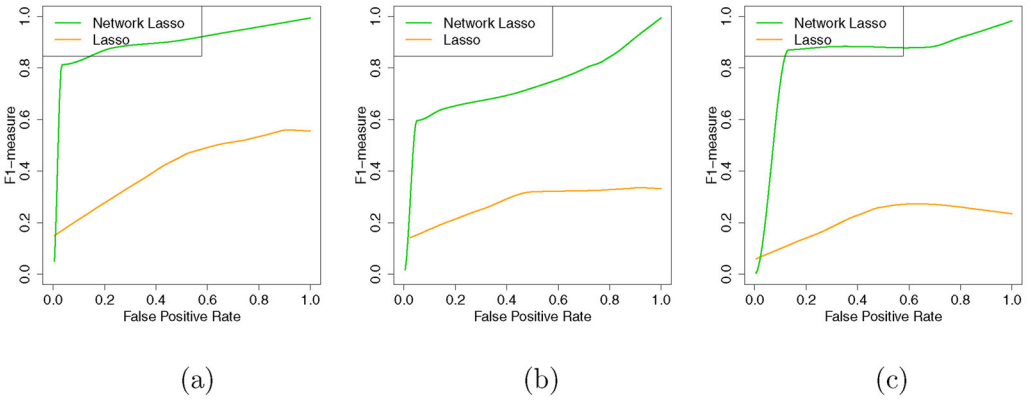


Figure 10. Performance is evaluated by the F1 with varying FPR for Scale-free network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

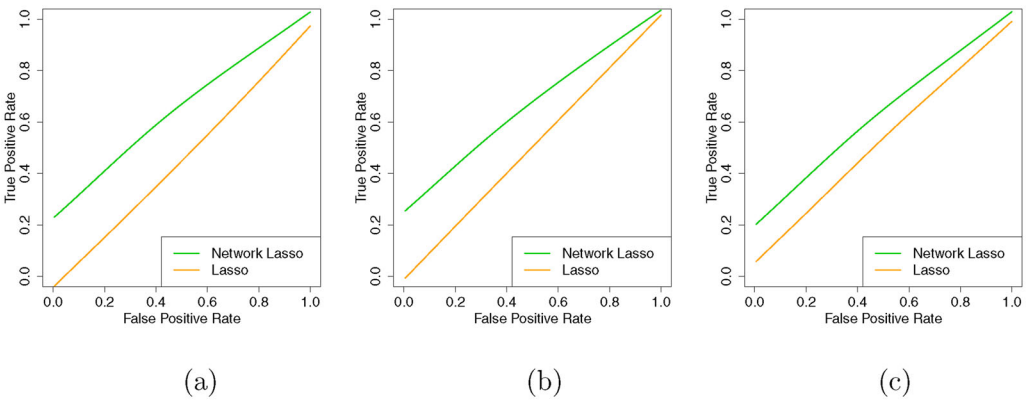


Figure 11. ROC curves of Lasso and Network Lasso for Hub network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

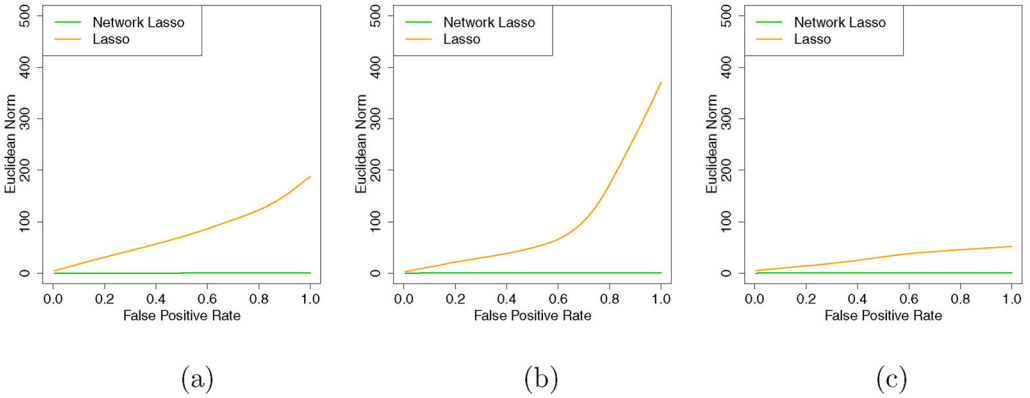


Figure 12. Performance is evaluated by the L_2-norm of β with varying FPR for Hub network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

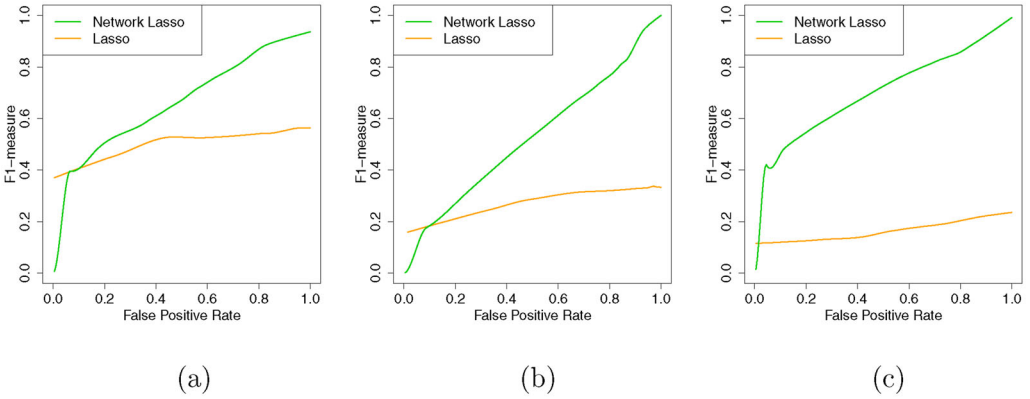


Figure 13. Performance is evaluated by the F1 with varying FPR for Hub network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

(b) In this part, the model parameter β generated as (2), and the corresponding simulation results and analysis of are reported as follows. Besides, we add F_1-measure to evaluate the performance of models and the results are summarized in Figures 8-16.

Figures 8, 11 and 14 show the ROC curve of Scale-free, Hub and Erdős-Renyi network graph, respectively. For Scale-free network, the performances of Network Lasso is much better than Lasso, and the differences between theirs become more and more obvious as p increased. For Hub network graph, the differences between Network Lasso and Lasso become more and more smaller as p increased. While the proposed model show significant performances over the Lasso. For Erdős-Renyi network, Network Lasso performs similar to Lasso when $p = 50$, and Network Lasso has much better performances than Lasso when $p = 100$.

From Figures 9, 12 and 15, we can also find that the L_2-norm of Lasso is much larger than Network Lasso in these three network graph. For Scale-free and Hub network, the results of Network Lasso almost equal as p increased, while Lasso performs best when

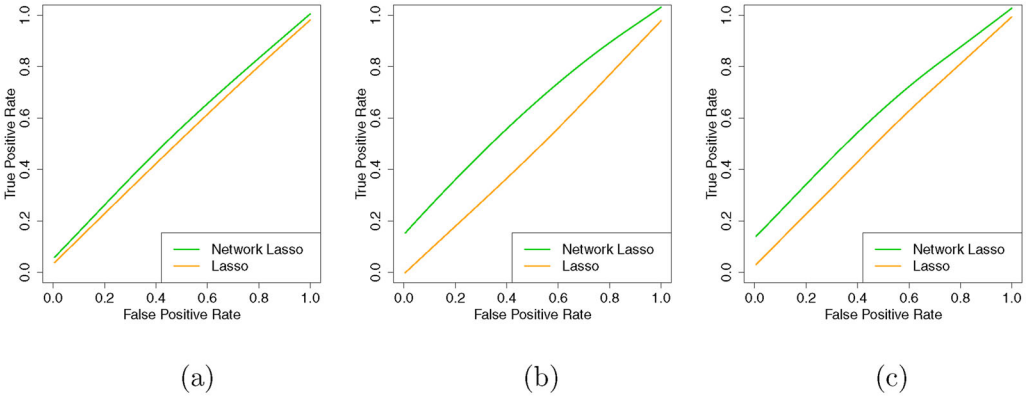


Figure 14. ROC curves of Lasso and Network Lasso for Erdős-Renyi. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

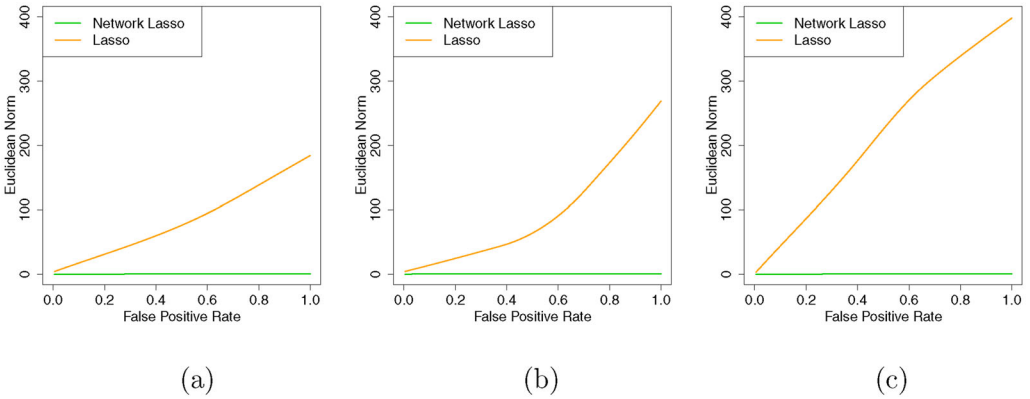


Figure 15. Performance is evaluated by the L_2-norm of β with varying FPR for Erdős-Renyi. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

$p = 50$. For Erdős-Renyi network, Network Lasso performs similar as p increased, while Lasso performs best when $p = 150$.

Figures 10, 13 and 16 show the F_1-measure of Scale-free, Hub and Erdős-Renyi network graph, respectively. For Scale-free network, the performances of Lasso is getting worse as p increased, while the Network Lasso performs best while $p = 150$. Figure 13 shows that the Network Lasso has the similar performances when $p = 50$ and $p = 150$. On the other hand, the results of Lasso become more and more worse as p increased. For Erdős-Renyi, the proposed model becomes more and more better while the performances of Lasso are decreasing as p increased. As a whole, all these Figures of F_1 further verifies that the proposed model perform much better than Lasso.

Furthermore, in order to get a general result, we increase this replication number to 10 and the simulation results are showed in Appendix 1. And increase the replication number to 100 and the corresponding results are listed in Appendix 1. Both of results of above two added simulation study show similar performances with the simulation (b), which further demonstrate the merit and availability of the proposed method.

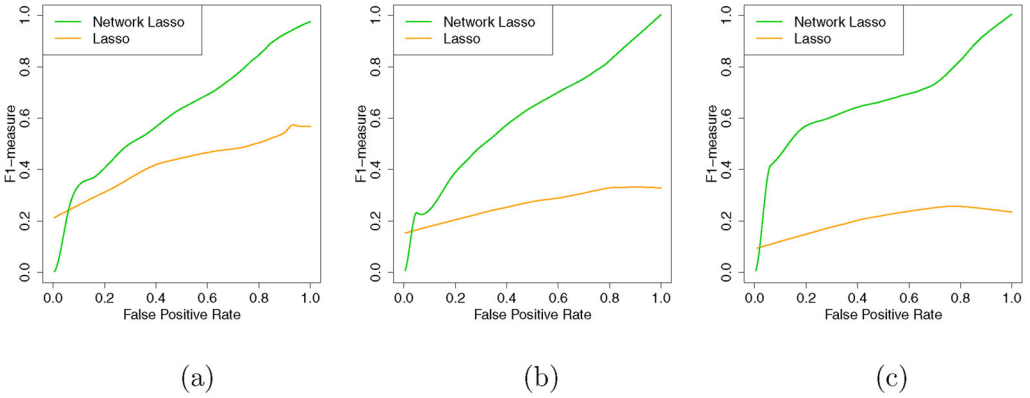


Figure 16. Performance is evaluated by the F1 with varying FPR for Erdős-Renyi. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

Therefore, due to the space concern, we only show the study results of above two simulations and do not report the repetition of the results analysis.

5. Real data analysis

In this section, in order to evaluate the proposed model performance, we apply it to estimate the price of homes based on latitude/longitude data and a set of features, as in Boyd, Leskovec and Hallac (2015). This dataset is a list of real estate transactions over a one-week period in May 2008 in the Greater Sacramento ares. It contains information on 985 sales, including latitude, longitude, number of bedrooms, number of bathrooms, square feet, and sales price. However, as often happens with real data, we are missing some of the values. Some of the bedroom/bathroom/size data is not provided. The prices and all attributes are standardized to zero mean and unit variance, so any missing features are ignored by setting the value to zero, the average. To verify our results, we use a random subset of 200 houses as our test set.

We build the graph by using the latitude/longitude coordinates of each house. After removing the test set, we connect every remaining house to the g(g = 3,5,7,10,all) nearest home with an edge weight inversely proportional to the distance between the houses. If house j is in the set of nearest neighbors of i, there is an undirected edge regardless of whether or not house i is one of j' nearest neighbors.

We plot the in-sample and out-sample mean squared prediction errors over the grid of λ' to evaluate the performance of our model. Figure 17 shows that the Lasso performances worst, while when the number of neighbor is 10, the Network Lasso has the best performance.

6. Summary

In this paper, we have proposed the Network Lasso by assuming the samples are linked by a network. In the high-dimensional setting, we have fitted the model in the l_1 penalized maximum likelihood and used the coordinate descent based algorithm to solve the problem. Theoretically, we have proved the consistency of the estimator under the

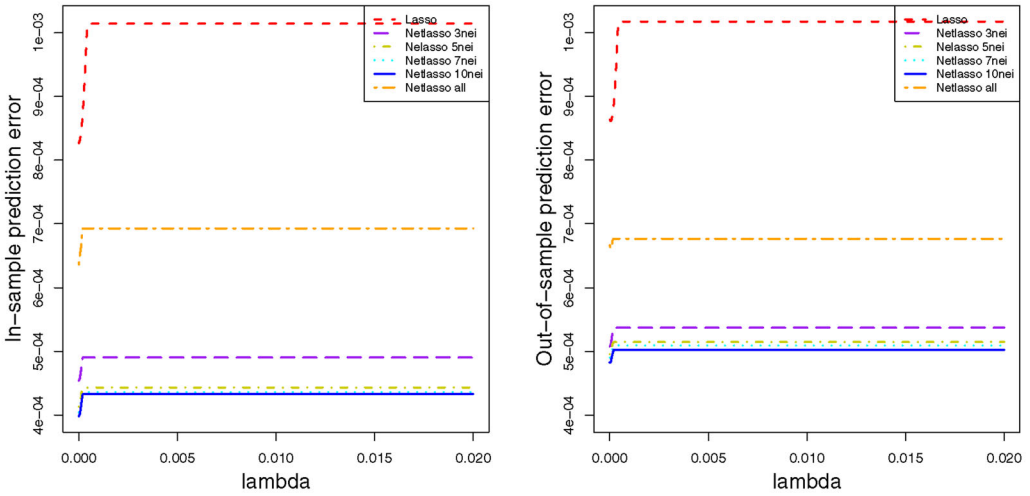


Figure 17. Performance is evaluated by in-sample and out-sample mean squared prediction errors with varying λ of each model.

restricted eigenvalue condition. Numerically, the simulation studies and the latitude data example have shown the merits and efficacy of the proposed method.

It is worth noting that the neighborhoods effect coefficients is same in the proposed model, thus it would be beneficial to consider the coefficients are different. In additional, it would be interesting to extend the framework to robust regression in order to solve the problem of heavy-tailed distributions.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Nos. 62076154, 61673249, U1805263), Key R&D program of Shanxi Province (International Cooperation, 201903D421050) and The central government guides local science and technology innovation projects#YDZX20201400001224).

References

Abbe, E. 2017. Community detection and stochastic block models: Recent developments. *arXiv Preprint arXiv* 1703:10146.

Belloni, A., and V. Chernozhukov. 2009. l_1-penalized quantile regression in high-dimensional sparse models. *Annals of Statistics* 39:82–130.

Boyd, S., J. Leskovec, and D. Hallac. 2015. Network Lasso: Clustering and optimization in large graphs. *arXiv Preprint arXiv* 1507:00280.

Brian, B., K. Brian, and E. Mark. 2011. Efficient and principled method for detecting communities in networks. *Physical Review E Statistical Nonlinear and Soft Matter Physics* 84. doi:10.1103/physreve.84.036103.

Cabreros, I., and A. Tsirigos. 2015. Detecting community structures in Hi-C genomic data, conference on information science and systems, Princeton University. *arXiv Preprint arXiv* 1509:05121.

Djebbari, H., and B. Fortin. 2009. Identification of peer effects through social network. *Journal of Econometrics* 150:41–55.

Eric, D., and K. Kolaczy. 2009. *Statistical analysis of network data: Methods and models*. Springer Series in Statistic.
Fan, J., Y. Fan, and E. Baryt. 2012. Adaptive robust variable selection. *Annals of Statistics* 42: 324–51.
Fan, J., and R. Li. 2001. Variable selection via nonconvave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96 (456):1348–60. doi:10.1198/016214501753382273.
Frank, L., and J. Friedman. 1993. A statistical view of some chemometrics regression tools. *Technometrics* 35 (2):109–35. [Database] doi:10.1080/00401706.1993.10485033.
Guo, X., and H. Zhang. 2020. Sparse directed acycic graphs incroporating the covariates. *Statistical Papers* 61 (5):2119–48. doi:10.1007/s00362-018-1027-8.
Lee, L. 2007. Identification and estimation of econometric models with group interactions, contextual factors and fixed effects. *Journal of Econometrics* 140 (2):333–74. doi:10.1016/j.jeconom.2006.07.001.
Manski, C. 1993. Identification of endogenous social effects: The reflection problem. *The Review of Economic Studies* 60 (3):531–42. doi:10.2307/2298123.
Negahban, S. N., P. Ravikumar, M. J. Wainwright, and B. Yu. 2012. A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. *Statistical Science* 27 (4):538–57. doi:10.1214/12-STS400.
Sharpnack, J., and A. Krishnamurthy. 2013. Detecting activations over graphs using spanning tree wavelet bases. *In Artificial intelligence and statistics*, 536–44.
Tibshirani, R. 1996. Regression shrinkage selection via the lasso. *Journal of the Royal Statistical Society Series B* 58:267–88.
Wang, Y., and R. Tibshirani. 2016. Trend filtering on graphs. *Journal of Machine Learning Research* 17: 1–41.
Zhu, J., T. Li, and E. Levina. 2018. Prediction models for network-linked data. *arXiv Preprint arXiv* 1602:01192.
Zou, H. 2006. The adaptive lasso and its oracle properties. *Journal of the American Statistical Association* 101 (476):1418–29. doi:10.1198/016214506000000735.
Zou, H., and T. Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67 (2):301–20. doi:10.1111/j.1467-9868.2005.00503.x.

Appendix 1

Figures 18-26.

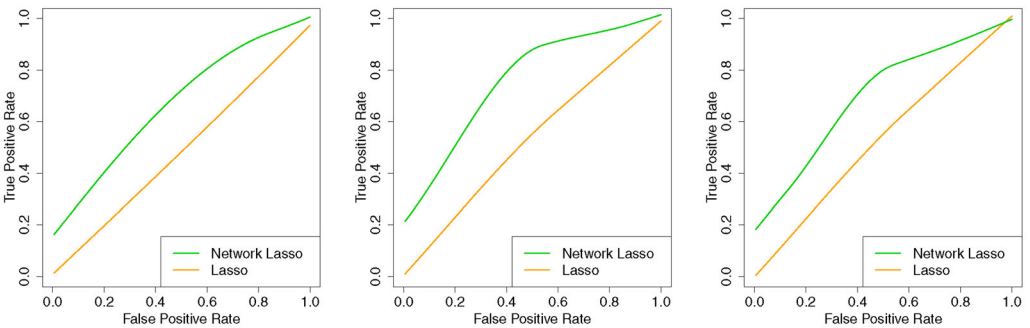


Figure 18. ROC curves of Lasso and Network Lasso for Scale-free network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

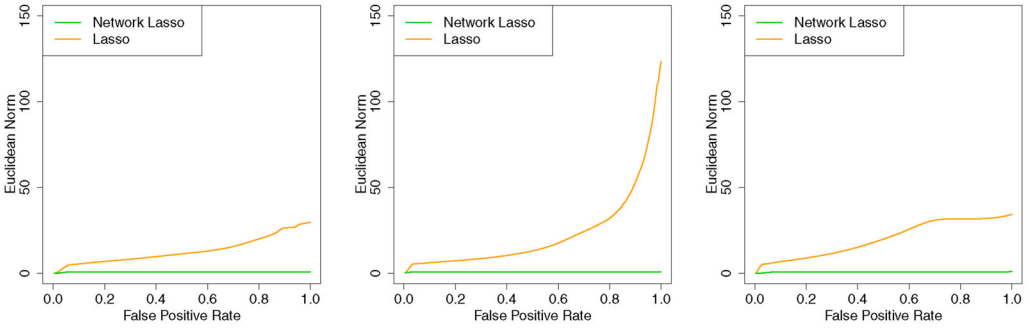


Figure 19. Performance is evaluated by the L_2-norm of β with varying FPR for Scale-free network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

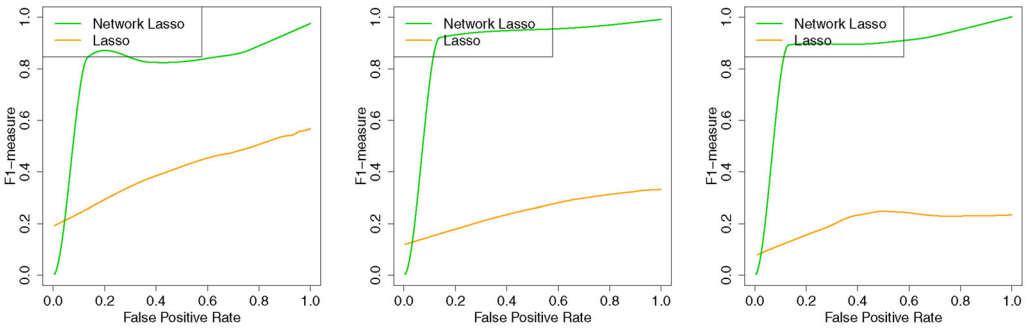


Figure 20. Performance is evaluated by the F1 with varying FPR for Scale-free network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

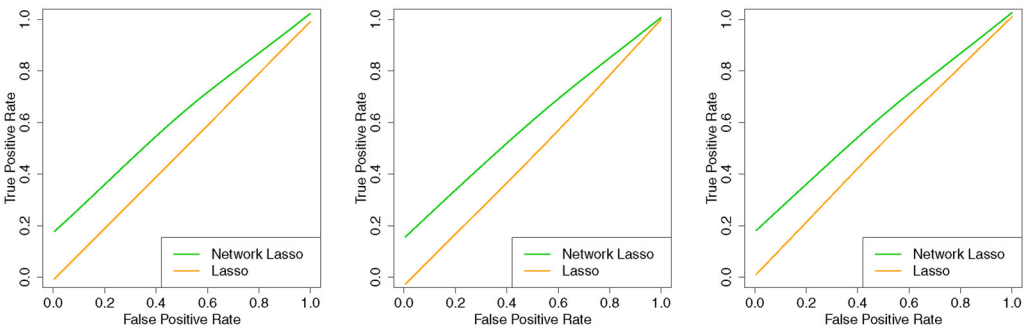


Figure 21. ROC curves of Lasso and Network Lasso for Hub network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

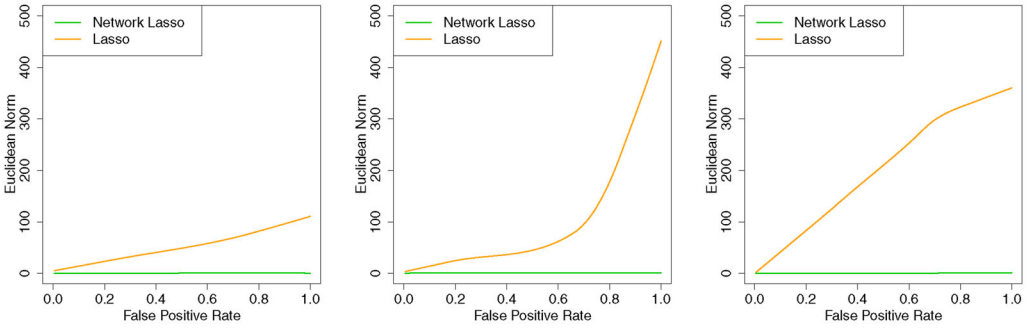


Figure 22. Performance is evaluated by the L_2-norm of β with varying FPR for Hub network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

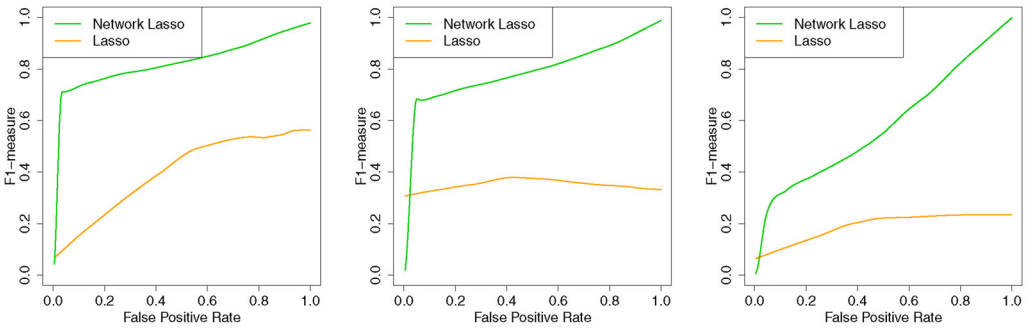


Figure 23. Performance is evaluated by the F1 with varying FPR for Hub network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

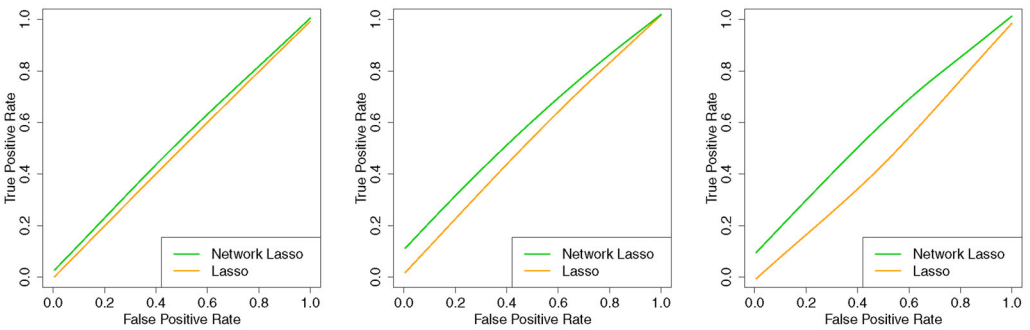


Figure 24. ROC curves of Lasso and Network Lasso for Erdős-Renyi network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

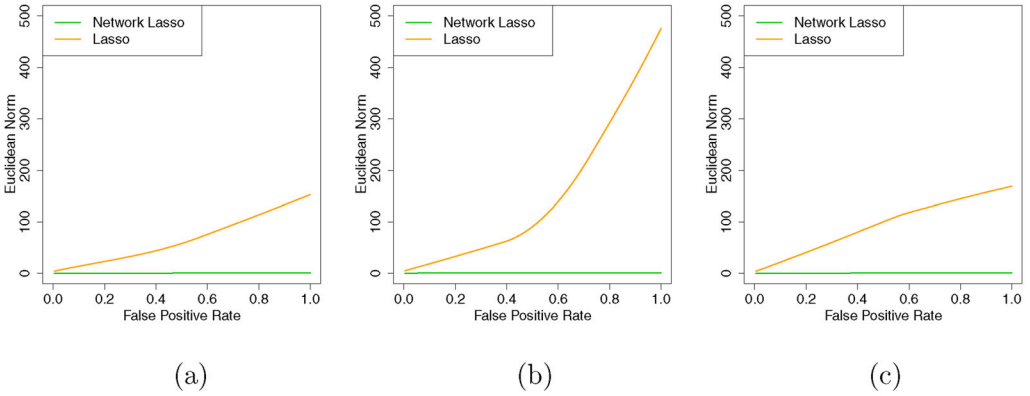


Figure 25. Performance is evaluated by the L_2-norm of β with varying FPR for Erdős-Renyi network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

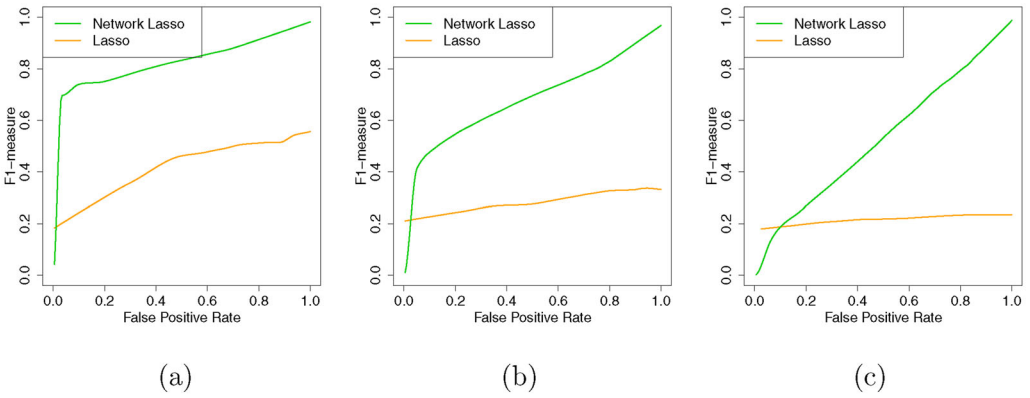


Figure 26. Performance is evaluated by the F1 with varying FPR for Erdős-Renyi network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

Appendix 2

Figures 27-35.

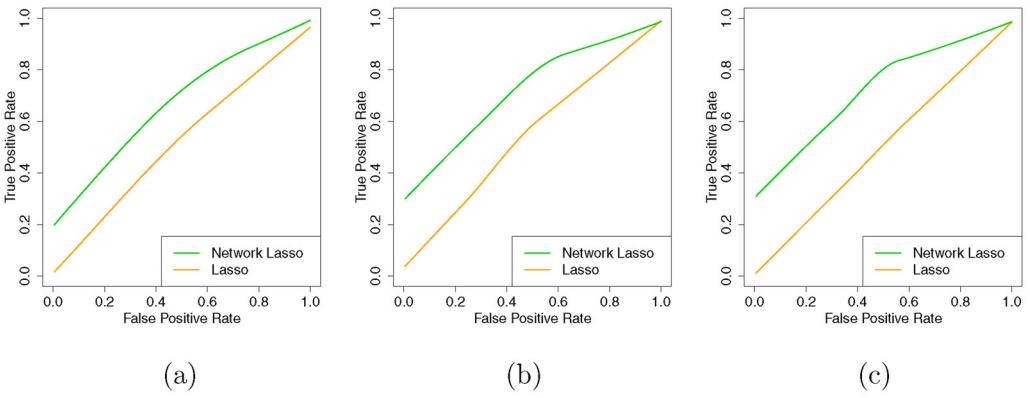


Figure 27. ROC curves of Lasso and Network Lasso for Scale-free network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

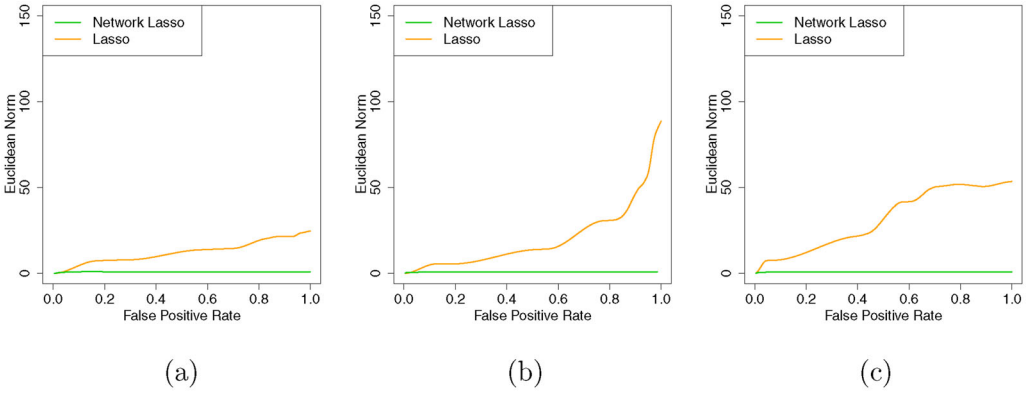


Figure 28. Performance is evaluated by the L_2-norm of β with varying FPR for Scale-free network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

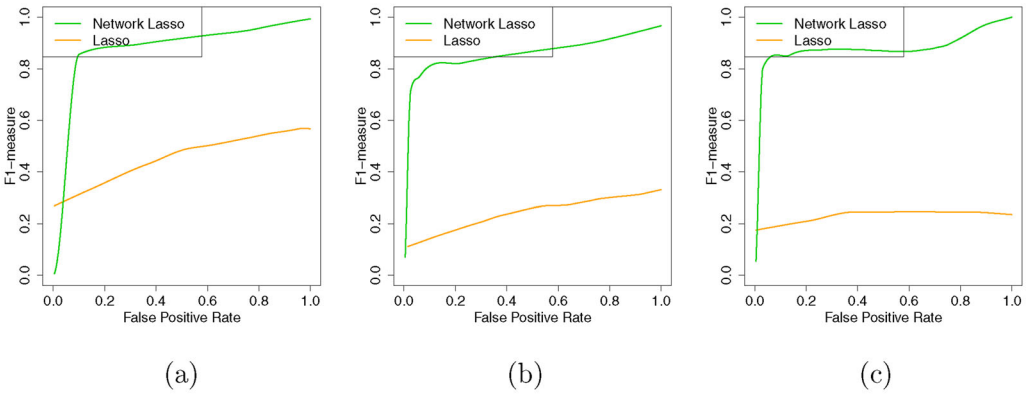


Figure 29. Performance is evaluated by the F1 with varying FPR for Scale-free network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

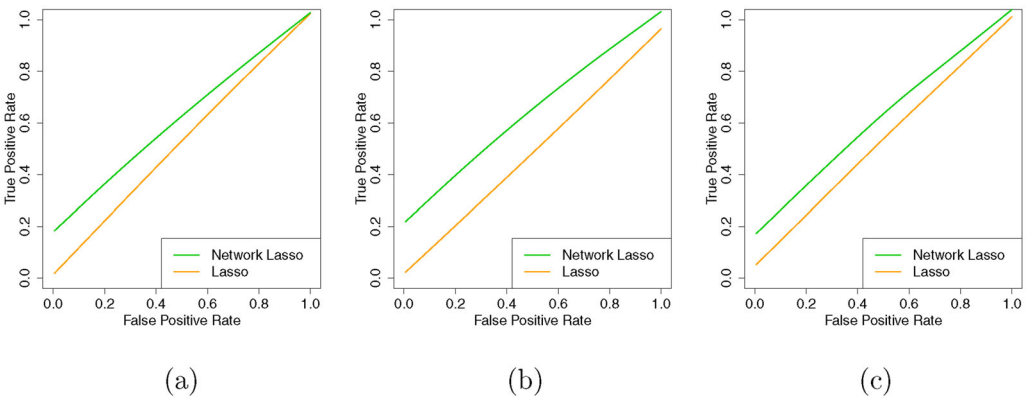


Figure 30. ROC curves of Lasso and Network Lasso for Hub network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

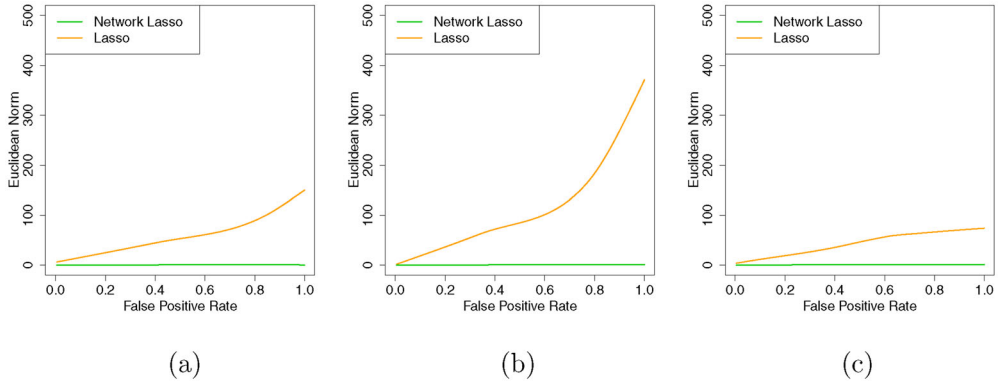


Figure 31. Performance is evaluated by the L_2-norm of β with varying FPR for Hub network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

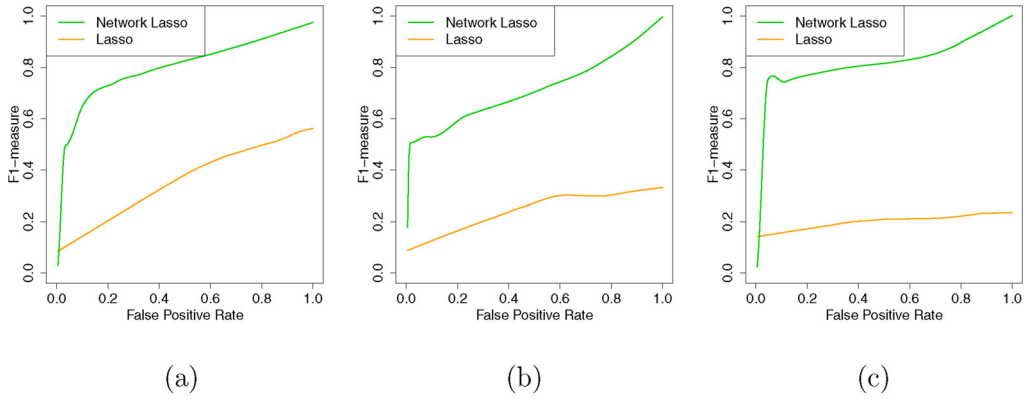


Figure 32. Performance is evaluated by the F1 with varying FPR for Hub network.
(a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

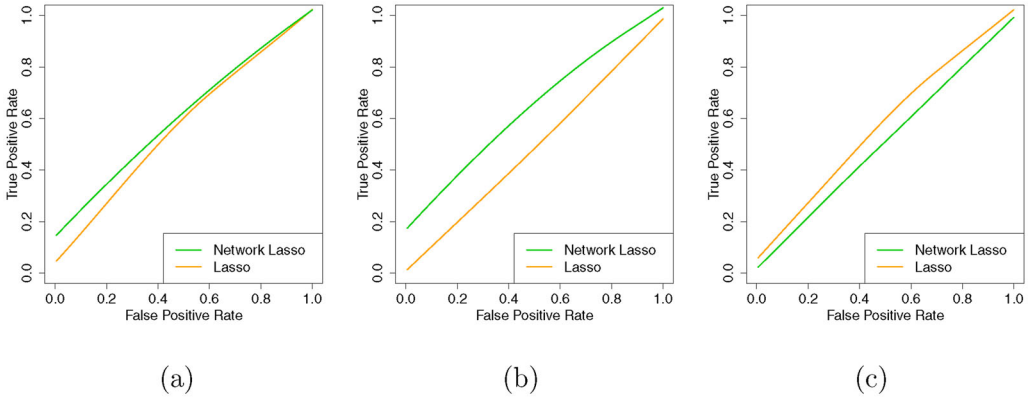


Figure 33. ROC curves of Lasso and Network Lasso for Erdős-Renyi network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

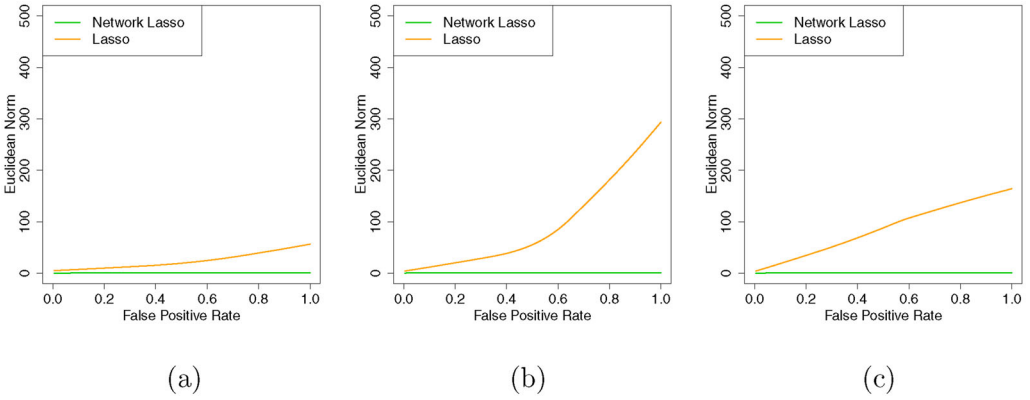


Figure 34. Performance is evaluated by the L_2-norm of β with varying FPR for Erdős-Renyi network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.

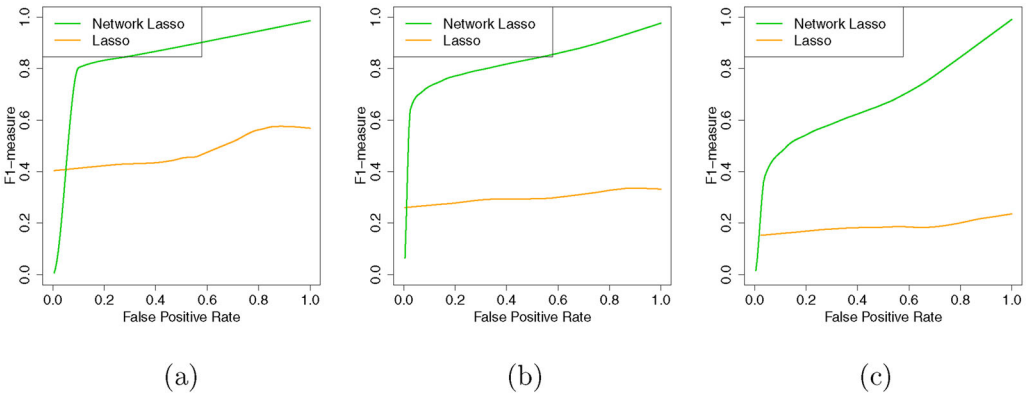


Figure 35. Performance is evaluated by the F1 with varying FPR for Erdős-Renyi network. (a) $p = 50$. (b) $p = 100$. (c) $p = 150$.