Effects of underlying gene-regulation network structure on prediction accuracy in high-dimensional regression

Yuichi Okinaga* Daisuke Kyogoku† Satoshi Kondo‡ Atsushi J. Nagano§
Kei Hirose¶

Abstract

Motivation: The least absolute shrinkage and selection operator (lasso) and principal component regression (PCR) are popular methods of estimating traits from high-dimensional omics data, such as transcriptomes. The prediction accuracy of these estimation methods is highly dependent on the covariance structure, which is characterized by gene regulation networks. However, the manner in which the structure of a gene regulation network together with the sample size affects prediction accuracy has not yet been sufficiently investigated. In this study, Monte Carlo simulations are conducted to investigate the prediction accuracy for several network structures under various sample sizes.
Results: When the gene regulation network was random graph, the simulation indicated that models with high estimation accuracy could be achieved with small sample sizes. However, a real gene regulation network is likely to exhibit a scale-free structure. In such cases, the simulation indicated that a relatively large number of observations is required to accurately predict traits from a transcriptome.
Availability and implementation: Source code at `https://github.com/keihirose/simrnet`
Contact: hirose@imi.kyushu-u.ac.jp

1 Introduction

Technological advancements have enabled the collection of highly multidimensional data from biological systems (Gehlenborg *et al.*, 2010; Mochida and Shinozaki, 2011; Li and Sillanpää, 2012; Hasin *et al.*, 2017). For example, RNA sequencing quantifies expression levels of thousands of genes. Such omics data is useful in predicting organismal traits, with empirical applications including diagnosis and classification of diseases and prediction of patient survival (van 't Veer *et al.*, 2002; Bøvelstad *et al.*, 2007; Chan *et al.*, 2016; Nandagopal *et al.*, 2019) and possible future applications in predicting crop yields (Kremling *et al.*, 2018), insecticide resistance (Dermauw *et al.*, 2013), and environmental adaptation (Nagano *et al.*, 2019).

A common challenge in predicting traits from omics data is the dimension of the data far exceeding that of the sample size (known as high-dimensional regression). For example, if one is to apply least-squares estimation in multiple regression (e.g. trait $\approx \beta_0 + \beta_1 \text{gene}_1 + \beta_2 \text{gene}_2 + \cdots$) to predict a trait value from a transcriptome, the sample size needs to be (at least) larger than the number of model parameters. However, because transcriptome studies typically observe thousands of genes, a sample size exceeding the number of genes is not realistic at present. In this case, high-dimensional regression modeling must be considered.

The least absolute shrinkage and selection operator (lasso, Tibshirani, 1996) is one of the most frequently used methods for high-dimensional regression. It simultaneously achieves variable selection and parameter

*Graduate School of Mathematics, Kyushu University, 744 Motooka, Fukuoka 819-0395, Japan,
†The Museum of Nature and Human Activities, 6 Yayoigaoka, Sanda, Hyogo 669-1546, Japan,
‡Agriculture and Biotechnology Business Division, Toyota Motor Corporation, Miyoshi, Aichi 470-0201, Japan,
§Faculty of Agriculture, Ryukoku University, Otsu, Shiga 520-2194, Japan,
¶Institute of Mathematics for Industry, Kyushu University, 744 Motooka, Fukuoka 819-0395, Japan and RIKEN Center for Advanced Intelligence Project, 1-4-1 Nihonbashi, Chuo-ku, Tokyo 103-0027, Japan

estimation. Theoretically, the prediction accuracy of the lasso is highly dependent on the correlation structure among exploratory variables; it is high under certain strong conditions, such as the compatibility condition (van de Geer and Bühlmann, 2009). However, in practice, it is not easy to check whether the compatibility condition holds. Another popular estimation method for high-dimensional regression is principal component regression (PCR, Jolliffe, 1986). PCR is a two-stage procedure: first, principal component analysis is conducted for predictors, following which the regression model on which the principal components are used as predictors is fitted. This method may perform well when the exploratory variables are highly correlated.

It is reasonable to assume that gene regulation networks will result in conditional independence among the levels of gene expression (Wei and Li, 2007; Dobra *et al.*, 2004; Yu *et al.*, 2013). Here, two variables are conditionally independent when they are independent given other variables (e.g. two focal variables are independently influenced by a third variable, Wille and Bühlmann, 2006). When a random vector of exploratory variables follows a multivariate normal distribution, two variables are conditionally independent if and only if the corresponding element of the inverse covariance matrix is nonzero. Essentially, the networks are characterized by the nonzero pattern of the inverse covariance matrix.

One of the most notable characteristics of biological networks is their scale-free nature, that is, the degree distribution of the network follows a power-law expressed as $p(x) \propto x^{-\gamma}$ ($\gamma > 1$) (Barabási and Albert, 1999; Milo *et al.*, 2002). Empirical studies suggest that biological networks are often scale-free (Barabasi and Oltvai, 2004; Albert, 2005; Arita, 2005), although exceptions have also been found Broido and Clauset (2019). Therefore, it is reasonable to consider the problem of high-dimensional regression when the networks of exploratory variables are scale-free. Here, it should be noted that the relative performance of different high-dimensional regression techniques may depend on sample sizes. However, to the best of our knowledge, the effect of the gene regulation network structure together with sample size on prediction accuracy has not yet been sufficiently investigated.

This paper provides a general simulation framework to study the effects of correlation structure in explanatory variables. As an example, the prediction of ambient temperature from the transcriptome, for which good empirical data is available (Nagano *et al.*, 2012, 2019), is considered. It should be noted that the implementation of the proposed procedure is independent of the empirical data in Nagano *et al.* (2012, 2019); the proposed framework may be applied to predict any consequence of gene expression differences (e.g. crop yield). The proposed framework is based on the Monte Carlo simulations. Three datasets of transcriptome and their traits are generated. The datasets are characterized by the covariance structure of exploratory variables; one of the covariance structures corresponds to the scale-free gene regulation network. Both lasso and PCR are applied to these simulated datasets to investigate the prediction accuracy with different types of gene regulation networks. The sample size is also varied to examine its effect on the prediction accuracy.

The remainder of this paper is organized as follows. Section 2 describes prediction methods for high-dimensional regression in the given simulation. Section 3 discusses the proposed simulation framework. Finally, Section 4 presents the concluding remarks.

2 Prediction methods for high-dimensional data

Suppose that we have n observations $\{(\boldsymbol{x}_i, y_i) \mid i = 1, \ldots, n\}$, where $\boldsymbol{x}_i$ are p-dimensional vector of explanatory variables and y_i are responses ($i = 1, \ldots, n$). Let $X = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)^T$ and $Y = (y_1, \cdots, y_n)^T$. Consider the linear regression model:

$$Y = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T$ is a vector of error variables with $E(\boldsymbol{\epsilon}) = \mathbf{0}$ and $\mathrm{var}(\boldsymbol{\epsilon}) = \sigma^2 I_n$.

2.1 Lasso

The lasso minimizes a loss function that consists of quadratic loss with a penalty based on an L_1 norm of a parameter vector:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|Y - X\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{1}$$

where $\lambda > 0$ is a regularization parameter. Because of the nature of the L_1 norm in the penalty term, some of the elements of the coefficients are estimated to be exactly zero. Thus, variable selection is conducted, and only variables that correspond to nonzero coefficients affect the responses.

2.2 PCR

In some cases, the first few largest eigenvalues of the covariance matrix of predictors (i.e., proportional contributions of principle components) can be considerably large (e.g., spiked covariance model, Johnstone *et al.*, 2001). In such a case, the lasso may not function effectively in terms of both prediction accuracy and consistency in model selection, because the conditions for its effective performance (e.g., compatibility condition, Bühlmann and van de Geer, 2011) may not be satisfied. This issue could be addressed using PCR because it transforms data with a large number of highly correlated variables into a few principal components. In the first stage of PCR, principal component analysis is applied to predictors, and the dimension of $\boldsymbol{x}_i$ is reduced to d ($d \ll p$). In this work, d was chosen such that d principle components collectively explain 90% or more variance (and $d-1$ principle components do not). Then, in the second stage, regression analysis is conducted, for which the principal components are used as predictors. Here, the regression coefficients in the second stage are estimated by the lasso.

3 Simulation framework

An overview of the simulation is presented in Fig. 1. First, the model that defines the relationship between the trait and the levels of gene expression was parameterized. This was done using the empirical data in Nagano *et al.* (2019), which quantified the transcriptome of wild *Arabidopsis halleri* subsp. *gemmifera* weekly for two years in their natural habitat as well as bihourly on the equinoxes and solstices ($p = 17205$ genes for $n = 835$ observations). Three types of simulated data were generated using different covariance matrices of genes, denoted as Σ_j ($j = 1, 2, 3$). Σ_1 is the sample covariance matrix of genes. Generally, none of the elements of the inverse of sample covariance matrix are exactly zero, implying that each gene interacts with all the other genes. Such a fully connected network is ineffective in terms of interpretation of the mechanism of gene regulation. Thus, two other covariance matrices were produced to simulate sparse networks based on the sample covariance matrix Σ_1. Σ_2 is generated by the graphical lasso (Yuan and Lin, 2007), which corresponds to the random graph. Although the graphical lasso is widely used because of its computational efficiency, real networks are often scale-free. Therefore, Σ_3, which corresponds to the scale-free network, was generated here. The estimation of scale-free networks is achieved by the reweighted graphical lasso (Liu and Ihler, 2011). Based on these three covariance matrices Σ_j ($j = 1, 2, 3$), the simulated transcriptome data were generated from the multivariate normal distribution. The simulated ambient temperature were generated from simulated transcriptome data. Finally, lasso and PCR were applied to these simulated data to compare their prediction accuracies. The sample sizes of the simulated data were varied to investigate the relationship between prediction accuracy and sample sizes.

3.1 Evaluation of the estimation procedure

The performance of the estimation procedure is investigated by the following expected prediction error:

$$E\left[\left\{Y^* - (X^*)^T\hat{\boldsymbol{\beta}}\right\}^2\right],$$

where X^* and Y^* follow $X^* \sim N(\mathbf{0}, \Sigma_j)$ ($j = 1, 2,$ or 3) and $Y^* \sim N((X^*)^T\boldsymbol{\beta}, \sigma^2)$, respectively. The estimator $\hat{\boldsymbol{\beta}}$ is obtained using current observations, while X^* and Y^* correspond to future observations. The Σ_j ($j = 1, 2, 3$), $\boldsymbol{\beta}$, and σ^2 are true values but unknown. In practice, these parameters are defined by using the actual dataset, (X, Y). Detail of setting of these parameters will be presented in Section 3.2.

To estimate the expected prediction error, the Monte Carlo simulation is conducted. We first randomly generate training and test data, $(\tilde{X}_{train}, \tilde{Y}_{train})$ and $(\tilde{X}_{test}, \tilde{Y}_{test})$, respectively. Here, $\tilde{X}_{train}$ follows a multivariate normal distribution with mean vector μ_X and variance-covariance matrix Σ_j, where μ_X is the

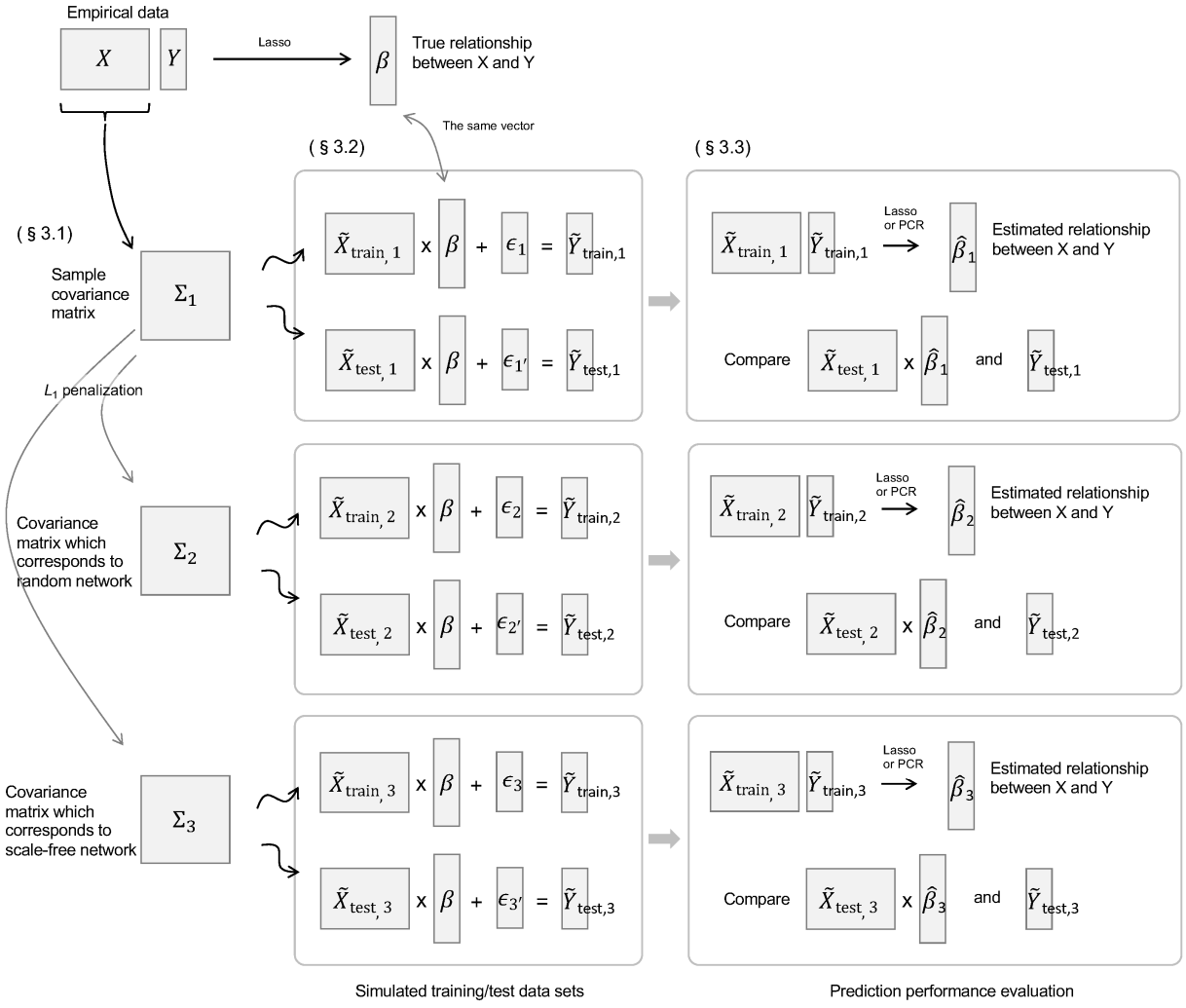


Figure 1: Overview of the simulation.

sample mean of X. Then, $\tilde{Y}_{train}$ is generated by $\tilde{Y}_{train} = \tilde{X}_{train}\boldsymbol{\beta} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon}$ is a random sample from $N(\mathbf{0}, \sigma^2 I)$ with I being an identity matrix. The test data, $(\tilde{X}_{test}, \tilde{Y}_{test})$, are generated by the same procedure as $(\tilde{X}_{train}, \tilde{Y}_{train})$ but independent of $(\tilde{X}_{train}, \tilde{Y}_{train})$. The number of observations for the training and test data are N $(N = 50, 100, 200, 300, 500, 1000)$ and 1000, respectively. The lasso and the PCR described in Section 2 are performed with training data $(\tilde{X}_{train}, \tilde{Y}_{train})$, following which RMSE is calculated in (8). The above process, from random generation of data to RMSE calculation, was performed 100 times.

3.2 Parameter setting

3.2.1 Covariance structures

Here, the characterization of the network structure of predictors by conditional independence is considered. When the predictors follow a multivariate normal distribution, the network structure based on the conditional independence corresponds to the nonzero pattern of the inverse covariance (precision) matrix. In other words, the network structure is characterized by the inverse covariance matrix of predictors.

Let S be the sample covariance matrix of predictors, that is, $S = \sum_{i=1}^{n}(\boldsymbol{x}_i - \bar{\boldsymbol{x}})(\boldsymbol{x}_i - \bar{\boldsymbol{x}})^T/n$ with $\bar{\boldsymbol{x}} = \sum_{i=1}^{n}\boldsymbol{x}_i/n$. Let $\Omega_j = \Sigma_j^{-1}$ $(j = 1, 2, 3)$. Σ_1 is a ridge estimator of the sample variance-covariance matrix, that is, $\Sigma_1 = S + \delta I$. Here δ is a small positive value (in this simulation, $\delta = 10^{-5}$). The term δI allows the existence of Ω_1. Note that because Ω_1 is not sparse, it leads to the complete graph, which is of no use in interpreting gene regulatory networks. To generate a covariance matrix whose inverse matrix is sparse, L_1 penalization is employed for the estimation of Ω_2 and Ω_3 as follows:

$$\hat{\Omega}_j = \arg\min_{\Omega} \{\log|\Omega| - \mathrm{tr}(\Omega S) - P_j(\Omega)\} \quad (j = 2, 3), \tag{2}$$

where $P_j(\Omega)$ $(j = 2, 3)$ are penalty terms which enhance the sparsity of the inverse covariance matrix. To estimate the sparse inverse covariance matrix, the lasso penalty is typically used as follows:

$$P_2(\Omega) = \rho \sum_{i=1}^{p} \|\boldsymbol{\omega}_{-i}\|_1, \tag{3}$$

where $\boldsymbol{\omega}_{-i} = (\omega_{i1}, \omega_{i2}, \cdots, \omega_{i(i-1)}, \omega_{i(i+1)}, \cdots, \omega_{ip})^T \in \mathbb{R}^{p-1}$. The problem (3) is referred to as the graphical lasso (Yuan and Lin, 2007), and there exists several efficient algorithms to obtain the solution (Friedman *et al.*, 2008; Witten *et al.*, 2011; Boyd, 2011). The estimator of (2) with (3) corresponds to Ω_2 and $\Sigma_2 = \Omega_2^{-1}$.

The lasso penalty (3) does not enhance scale-free networks. It penalizes all edges equally so that the estimated graph is likely to be a random graph, that is, the degree distribution becomes a binomial distribution. To enhance scale-free networks (i.e., power-law distribution), the log penalty (Liu and Ihler, 2011) is used as follows:

$$P_3(\Omega) = \rho \sum_{i=1}^{p} \log\left(\|\boldsymbol{\omega}_{-i}\|_1 + a_i\right), \tag{4}$$

where $a_i > 0$ are tuning parameters. From a Bayesian viewpoint, the prior distribution which corresponds to the log penalty becomes the power-law distribution (Liu and Ihler, 2011); thus, the penalty (4) is likely to estimate the scale-free networks. The estimator of (2) with (4) corresponds to Ω_3.

Because the log-penalty (4) is nonconvex, it is not easy to directly optimize (2). To implement the maximization problem (2), Liu and Ihler (2011) constructed the minorize-maximization (MM) algorithm (Hunter and Lange, 2004), in which the weighted lasso penalty $P_M^{(t)}(\Omega)$ with current parameter $\Omega_3^{(t)}$ is used:

$$P_M^{(t)}(\Omega) = \sum_{i=1}^{p}\sum_{j\neq i} \rho_{ij}^{(t)}|\omega_{ij}|, \tag{5}$$

where $\rho_{ij}^{(t)}$ are the weights

$$\rho_{ij}^{(t)} = \frac{\rho}{\|\boldsymbol{\omega}_{-i}^{(t)}\|_1 + a_i}. \tag{6}$$

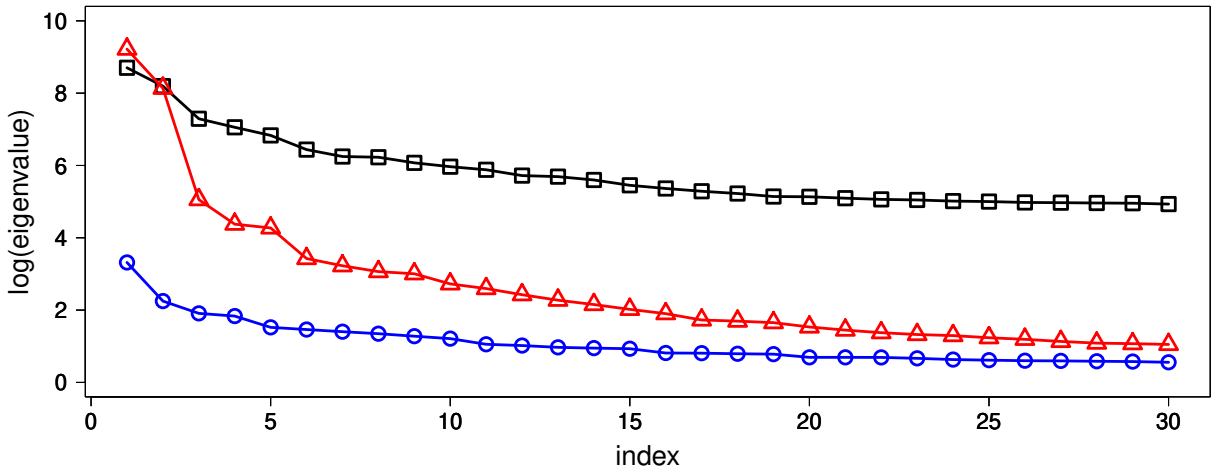


Figure 2: Logarithm graph of the largest 30 eigenvalues of Σ_1 (black square), Σ_2 (blue triangle) and Σ_3 (red circle). The horizontal axis expresses the index of eigenvalues arranged in descending order.

Because the weighted graphical lasso can be implemented by a standard graphical lasso algorithm, the estimator is obtained as the following algorithm:

1. Set $t = 0$. Get $\Omega_3^{(0)}$ via ordinary graphical lasso. Repeat 2 to 4 until convergence.
2. Update weights $\rho_{ij}^{(t)}$ using (6).
3. Get $\Omega_3^{(t+1)}$ via the weighted graphical lasso (2) with penalty (5).
4. $t \leftarrow t + 1$.

To obtain Σ_2 and Σ_3, the tuning parameters a_i $(i = 1 \ldots, p)$ and ρ must be determined. Following the experiments in Liu and Ihler (2011), $a_i = 1$ was set for $i = 1 \ldots, p$. To select the value of the regularization parameter ρ, several candidates were first prepared. In our simulation, the candidates were $\rho = 0.3, 0.4, 0.5, 0.6, 0.7$. From these, the value of ρ was selected such that the extended Bayesian information criterion (EBIC, Chen and Chen, 2008; Foygel and Drton, 2010)

$$\text{EBIC} = -n \left\{ \log |\Omega_2| - \text{tr}(\Omega_2 S) \right\} + q \log n + 4q\delta \log p \tag{7}$$

was minimized. Here, q is the number of nonzero parameters of the upper triangular matrix of $\hat{\Omega}$, and $\delta \in [0, 1)$ is a tuning parameter. As the value of δ increases, a sparser graph is generated. Note that $\delta = 0$ corresponds to the ordinary BIC (Schwarz, 1978). We set $\delta = 0.5$ because Foygel and Drton (2010) showed that $\delta = 0.5$ yielded good performance in both simulated and real data analyses. As a result, the EBIC selected $\rho = 0.5$.

The upper triangular matrix Ω_3 must be estimated with the reweighted graphical lasso problem. A value of $p = 17205$ results in $p(p+1)/2 \approx 148$ million parameters. As a result, with the machine used in this study (Intel Core Xeon 3 GHz, 128 GB memory), it would take several days to conduct the reweighted graphical lasso approach, even with a small number of iterations such as $T = 5$. For this reason, $T = 5$ iterations were employed to produce Σ_3 here.

Fig. 2 depicts the logarithm of the largest 30 eigenvalues of Σ_j $(j = 1, 2, 3)$. The first few largest eigenvalues of Σ_3 are significantly larger than those of Σ_2, implying that the scale-free networks tend to produce predictors with large correlations.

3.2.2 Regression parameters

The values of $\boldsymbol{\beta}$ and σ^2 are determined as follows. First, 10-fold cross-validation is performed as described below, and the regularization parameter λ in (1) is selected. The data (X, Y) are divided into ten datasets, $(X^{(j)}, Y^{(j)})$ $(j = 1, \ldots, 10)$, which consist of almost equal sample sizes. Let $X^{(-j)} = (X^{(1)}, \ldots, X^{(j-1)}, X^{(j+1)}, \ldots, X^{(10)})$, and $Y^{(-j)} = (Y^{(1)}, \ldots, Y^{(j-1)}, Y^{(j+1)}, \ldots, Y^{(10)})$ $(j = 1, \ldots, 10)$. For each j $(j = 1, \ldots, 10)$, the training and test data are defined by $(X^{(-j)}, Y^{(-j)})$ and $(X^{(j)}, Y^{(j)})$, respectively. Then, the parameter $\hat{\boldsymbol{\beta}}^{(j)}$ $(j = 1, \ldots, 10)$ is found by the lasso:

$$\hat{\boldsymbol{\beta}}^{(j)} = \arg\min_{\boldsymbol{\beta}} \left(\|Y^{(-j)} - X^{(-j)}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right).$$

For each j $(j = 1, \ldots, 10)$, the verification error is calculated as follows:

$$\mathrm{CV}^{(j)} = \frac{1}{\#Y^{(j)}} \|Y^{(j)} - X^{(j)}\hat{\boldsymbol{\beta}}^{(j)}\|_2^2.$$

Then, λ is adopted such that it minimizes $\mathrm{CV} = \frac{1}{10}\sum_{j=1}^{10} \mathrm{CV}^{(j)}$, the mean of $\mathrm{CV}^{(j)}$. Following this, the dataset (X, Y) is again randomly divided into two datasets: test data (X_{test}, Y_{test}) and training data (X_{train}, Y_{train}). Lasso estimation (1) is performed using the training data, with λ obtained by the above 10-fold cross-validation. Then, $\boldsymbol{\beta}$ is defined as the lasso estimator, resulting in the number of nonzero parameters of $\boldsymbol{\beta}$ being 259. Fig. 3 shows the histogram of nonzero parameters of $\boldsymbol{\beta}$. It is seen that the majority of the nonzero coefficients were close to zero; only 15 parameters had absolute values larger than 0.1.

In addition, the root mean squared error (RMSE) is calculated as follows:

$$\mathrm{RMSE} = \frac{1}{\sqrt{\#Y_{test}}} \|Y_{test} - X_{test}\hat{\boldsymbol{\beta}}\|_2, \tag{8}$$

and the variance of errors, σ^2, is defined by $\sigma^2 = (\mathrm{RMSE})^2$.

3.3 Results

The box and whisker plot of the RMSE is drawn in Fig. 4. The horizontal axis is N (the number of observations of training data) and the vertical axis is the RMSE based on 1000 observations of test data.

We compared the performance of the lasso with that of the PCR. When Σ_1 and Σ_3 were used, the PCR performed worse than the lasso for small sample sizes. Some predictors associated with small eigenvalues may affect prediction performance. Meanwhile, for Σ_2, the performance of PCR was slightly more stable than that of the lasso for small sample sizes.

The prediction accuracy was compared among the three covariance structures. For both lasso and PCR, when Σ_1 and Σ_3 were used, the values of RMSE decreased as N increased. On the other hand, when Σ_2 was used, the values of RMSE remained almost unchanged, approximately ranging between 2.5–3, as N increased. In the case of scale-free (Σ_3), when the number of observations was large, the estimation accuracy was approximately between 2.5–3, which was almost identical to the accuracy of Σ_2. For example, the mean RMSE at $N = 1000$ in Fig. 4 (e) was nearly equal to that at $N = 50$ in (c). The reason the accuracy of Σ_2 remained high even at $N = 50$ is considered to be that Σ_2 was weakly-correlated (Fig. 2) and the majority of the nonzero parameters of $\boldsymbol{\beta}$ were small (Fig. 3). Such a weakly-correlated covariance matrix implies conditions on Σ_2 that achieve nearly optimal rates may be satisfied (e.g., van de Geer and Bühlmann, 2009).

As described before, Σ_1 was the sample covariance matrix, while Σ_3 (and Σ_2) was estimated using the graphical lasso. As the lasso-type regularization methods shrink parameters toward zero, the correlations among exploratory variables reduce with the graphical lasso. Therefore, Σ_3 resulted in smaller correlations as compared to Σ_1. Consequently, the prediction accuracy reduces with stronger correlations.

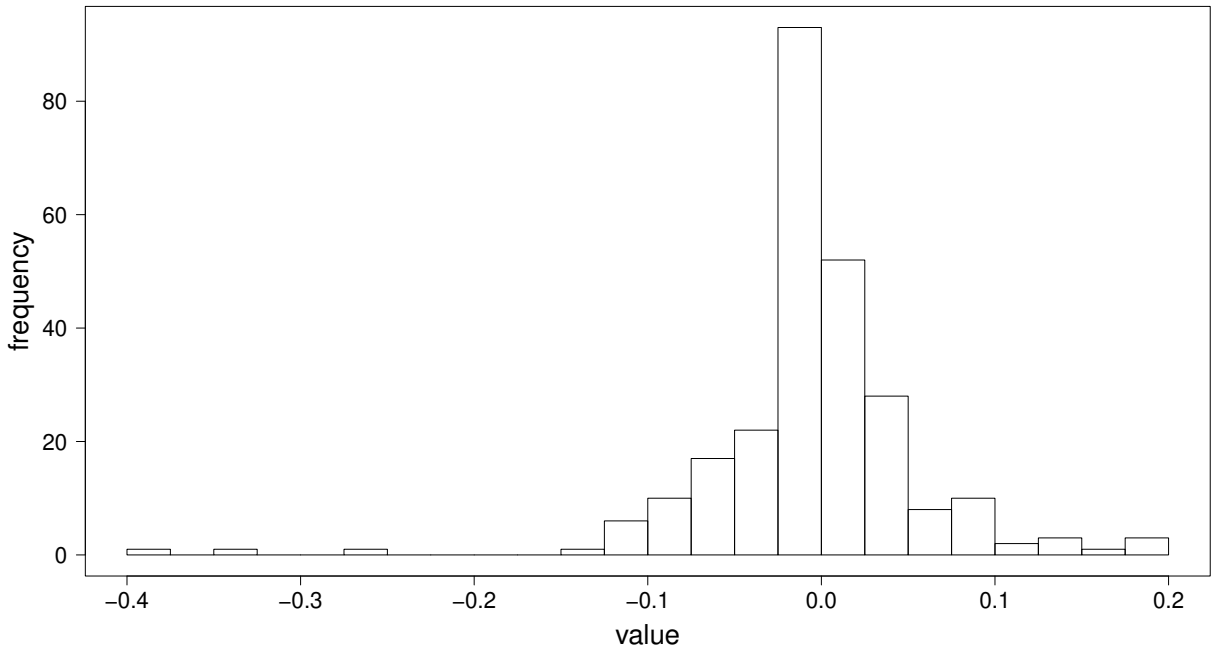


Figure 3: Histogram of 259 nonzero parameters of $\boldsymbol{\beta}$.

3.4 Code availability

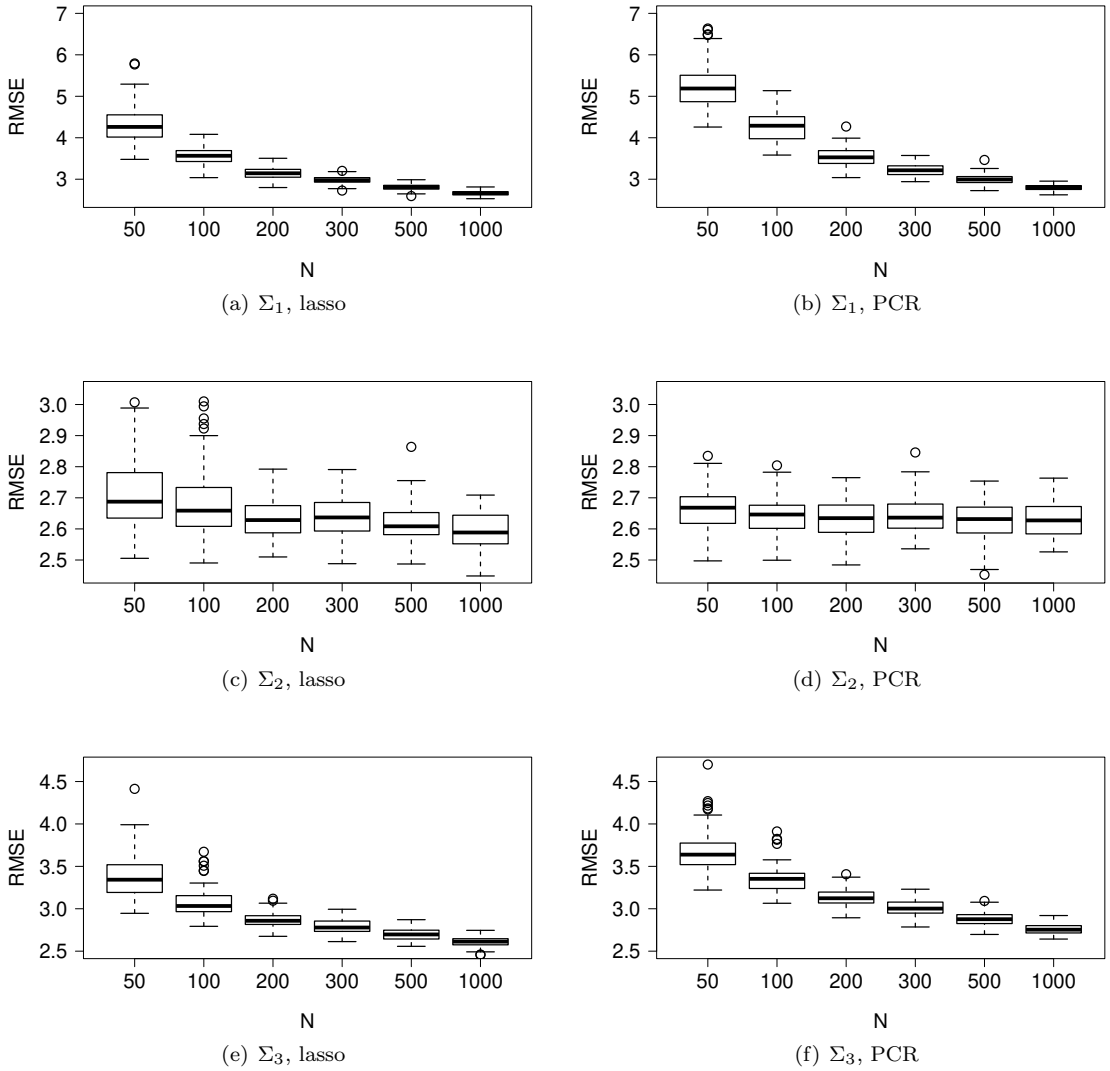
The proposed simulation is implemented in `R` package `simrnet`, which is available at `https://github.com/keihirose/simrnet`. Below is a sample code of the `simrnet` in `R`:

```
library(devtools)
install_github("keihirose/simrnet") #install package
library(simrnet) #load package
data(nagano2019)
attach(nagano2019)
rho <- (1:9) / 10 #tuning parameters for glasso
pars <- genpar(X,Y,rho) #set true parameter
result <- simrnet(pars, times.sim=100) #conduct simulation
plot(result)
```

When $p = 100$, it took less than 12 minutes to conduct the simulation with 100 replications using the machine employed herein (Intel Core Xeon 3GHz, 128GB memory). For high-dimensional data such as $p = 17205$, which was used in the simulation presented in this paper, several days were required to complete the simulation task.

4 Concluding remarks

In a gene regulation network, a gene regulates a small portion of a genome, not all the genes in a genome. This indicates that gene regulation network is expected to be a sparse network rather than a complete graph. Therefore, two covariance matrices indicating sparse networks (Σ_2, Σ_3) were prepared in addition to a covariance matrix derived from empirical data (Σ_1). Generally, although hundreds of genes contribute



(a) Σ_1, lasso

(b) Σ_1, PCR

(c) Σ_2, lasso

(d) Σ_2, PCR

(e) Σ_3, lasso

(f) Σ_3, PCR

Figure 4: Box and whisker plot of RMSE. The variance-covariance matrix used in the simulations is Σ_1 in (a)–(b), Σ_2 in (c)–(d), and Σ_3 in (e)–(f). The regression model is estimated by the lasso (Section 2.1) in (a), (c), and (e) and by PCR (Section 2.2) in (b), (d), and (f).

to defining a trait, their contributions are not equal. It is frequently observed that genes regulating a trait include a few large-effect genes and several small-effect genes. This property was reflected in the distribution of $\boldsymbol{\beta}$ (Fig. 3). When a limited number of regression coefficients had a large contribution to the definition of a trait, and the gene regulation network was random (Σ_2), the simulation indicated that models with high estimation accuracy could be developed from a small number of observations (Fig. 4). However, a real gene regulation network is likely to exhibit scale-free structure. In such cases, the simulation indicated that the prediction of traits from a transcriptome requires a relatively large number of observations to produce good performance (Σ_1, Σ_3, Fig. 4). In conclusion, it is necessary to secure sufficiently large sample sizes when performing regression analysis of data with scale-free network.

Conventional theory on the relationship between RMSE and sample size has been developed under the assumption that the sample size exceeds the number of exploratory variables (e.g., Fahrmeir *et al.*, 2007). However, omics data, which is rapidly being accumulated, results in high dimensional data with strong correlations. Thus, our simulation study considered more complicated settings than the traditional ones. Our simulation, or its extension, may be used in the future to find clues about theoretical aspects that may ultimately lead to the development of a sample size determination technique for omics data. Another important future research topic is the development of methods that have better estimation accuracy than the lasso in the case of small sample sizes.

Acknowledgements

The authors would like to thank Mr. Kanta Miura for the valuable discussions.

Funding

This work was partially supported by the Japan Society for the Promotion of Science KAKENHI 19K11862 (KH) and JST CREST Grant Number JPMJCR15O2 (AJN).

References

Albert, R. (2005). Scale-free networks in cell biology. *Journal of cell science*, **118**(21), 4947–4957.

Arita, M. (2005). Scale-freeness and biological networks. *Journal of biochemistry*, **138**(1), 1–4.

Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *Science*, **286**(5439), 509–512.

Barabasi, A.-L. and Oltvai, Z. N. (2004). Network biology: understanding the cell's functional organization. *Nature reviews genetics*, **5**(2), 101–113.

Bøvelstad, H. M. *et al.* (2007). Predicting survival from microarray data - A comparative study. *Bioinformatics*, **23**(16), 2080–2087.

Boyd, S. (2011). Alternating direction method of multipliers. In *Talk at NIPS Workshop on Optimization and Machine Learning*.

Broido, A. D. and Clauset, A. (2019). Scale-free networks are rare. *Nature communications*, **10**(1), 1017.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Chan, A. W. *et al.* (2016). 1H-NMR urinary metabolomic profiling for diagnosis of gastric cancer. *British journal of cancer*, **114**(1), 59–62.

Chen, J. and Chen, Z. (2008). Extended bayesian information criteria for model selection with large model spaces. *Biometrika*, **95**(3), 759–771.

Dermauw, W. *et al.* (2013). A link between host plant adaptation and pesticide resistance in the polyphagous spider mite tetranychus urticae. *Proceedings of the National Academy of Sciences*, **110**(2), E113–E122.

Dobra, A. *et al.* (2004). Sparse graphical models for exploring gene expression data. *Journal of Multivariate Analysis*, **90**(1), 196–212.

Fahrmeir, L. *et al.* (2007). *Regression*. Springer.

Foygel, R. and Drton, M. (2010). Extended bayesian information criteria for gaussian graphical models. In *Advances in neural information processing systems*, pages 604–612.

Friedman, J. *et al.* (2008). Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, **9**(3), 432–441.

Gehlenborg, N. *et al.* (2010). Visualization of omics data for systems biology. *Nature methods*, **7**(3s), S56.

Hasin, Y. *et al.* (2017). Multi-omics approaches to disease. *Genome biology*, **18**(1), 83.

Hunter, D. R. and Lange, K. (2004). A tutorial on mm algorithms. *The American Statistician*, **58**(1), 30–37.

Johnstone, I. M. *et al.* (2001). On the distribution of the largest eigenvalue in principal components analysis. *The Annals of statistics*, **29**(2), 295–327.

Jolliffe, I. T. (1986). Principal components in regression analysis. In *Principal component analysis*, pages 129–155. Springer.

Kremling, K. A. *et al.* (2018). Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature*, **555**(7697), 520–523.

Li, Z. and Sillanpää, M. J. (2012). Overview of lasso-related penalized regression methods for quantitative trait mapping and genomic selection. *Theoretical and applied genetics*, **125**(3), 419–435.

Liu, Q. and Ihler, A. T. (2011). Learning scale free networks by reweighted l1 regularization. In *International Conference on Artificial Intelligence and Statistics*, pages 40–48.

Milo, R. *et al.* (2002). Network motifs: simple building blocks of complex networks. *Science*, **298**(5594), 824–827.

Mochida, K. and Shinozaki, K. (2011). Advances in omics and bioinformatics tools for systems analyses of plant functions. *Plant and Cell Physiology*, **52**(12), 2017–2038.

Nagano, A. *et al.* (2012). Deciphering and prediction of transcriptome dynamics under fluctuating field conditions. *Cell*, **151**(6), 1358–1369.

Nagano, A. J. *et al.* (2019). Annual transcriptome dynamics in natural environments reveals plant seasonal adaptation. *Nature plants*, **5**(1), 74–83.

Nandagopal, V. *et al.* (2019). Feasible analysis of gene expression–a computational based classification for breast cancer. *Measurement*, **140**, 120–125.

Schwarz, G. (1978). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics*, **9**, 1135–1151.

Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *Journal of the Royal Statistical Society Series B-Methodological*, **58**(1), 267–288.

van de Geer, S. A. and Bühlmann, P. (2009). On the conditions used to prove oracle results for the lasso. *Electron. J. Statist.*, **3**, 1360–1392.

van 't Veer, L. J. *et al.* (2002). Gene expression profiling predicts clinical outcome of breast cancer. *nature*, **415**(6871), 530–536.

Wei, Z. and Li, H. (2007). A markov random field model for network-based analysis of genomic data. *Bioinformatics*, **23**(12), 1537–1544.

Wille, A. and Bühlmann, P. (2006). Low-order conditional independence graphs for inferring genetic networks. *Statistical applications in genetics and molecular biology*, **5**(1).

Witten, D. M. *et al.* (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, **20**(4), 892–900.

Yu, D. *et al.* (2013). Review of biological network data and its applications. *Genomics & informatics*, **11**(4), 200.

Yuan, M. and Lin, Y. (2007). Model selection and estimation in the gaussian graphical model. *Biometrika*, **94**(1), 19–35.