ESTIMATION OF NONLINEAR GENE REGULATORY NETWORKS VIA L_1 REGULARIZED NVAR FROM TIME SERIES GENE EXPRESSION DATA

KANAME KOJIMA
kaname@ims.u-tokyo.ac.jp

ANDRÉ FUJITA
afujita@ims.u-tokyo.ac.jp

TEPPEI SHIMAMURA
shima@ims.u-tokyo.ac.jp

SEIYA IMOTO
imoto@ims.u-tokyo.ac.jp

SATORU MIYANO
miyano@ims.u-tokyo.ac.jp

*Human Genome Center, Institute of Medical Science, University of Tokyo, 4-6-1
Shirokanedai, Minato-ku, Tokyo 108-8639, Japan*

Recently, nonlinear vector autoregressive (NVAR) model based on Granger causality was proposed to infer nonlinear gene regulatory networks from time series gene expression data. Since NVAR requires a large number of parameters due to the basis expansion, the length of time series microarray data is insufficient for accurate parameter estimation and we need to limit the size of the gene set strongly. To address this limitation, we employ L_1 regularization technique to estimate NVAR. Under L_1 regularization, direct parents of each gene can be selected efficiently even when the number of parameters exceeds the number of data samples. We can thus estimate larger gene regulatory networks more accurately than those from existing methods. Through the simulation study, we verify the effectiveness of the proposed method by comparing its limitation in the number of genes to that of the existing NVAR. The proposed method is also applied to time series microarray data of Human hela cell cycle.

Keywords: time series gene expression data; gene regulatory networks; vector autoregression; B-spline; group LASSO

1. Introduction

Using time series microarray data, estimation of gene regulatory networks is one of the essential roles to elucidate transcriptional systems. Recently, various statistical approaches have been proposed to capture gene regulations using dynamic Bayesian network [13, 18], vector autoregressive model [7–9], and state space model [12, 25] based on statistical causality. In this study, we use vector autoregressive model and capture gene regulations based on Granger causality. Linear vector autoregressive models are well-established in statistics and in existing literature it has been applied to estimate gene regulatory networks. However, most of the regulations cannot be limited by linear [9], and we need to extend classical vector autoregressive models into nonlinear vector autoregressive models.

Fujita *et al.* [9] introduced nonparametric regression technique to vector autoregressive model for estimating nonlinear and nonmonotonic regulations in gene

regulatory networks. In nonparametric regression, since basis expansion technique was applied to build nonlinear mean function, the number of parameters increases rapidly. In addition, the number of genes that can be handled is highly limited by the fact that the length of time series microarray data is very short. Thus, we propose to use $L1$ regularization technique and address the estimation of nonlinear and nonmonotonic gene regulatory networks. $L1$ regularized nonparametric regression is reduced to group LASSO problem [16, 21]. For the solution of group LASSO, we show a new efficient method based on interior point method. Also, the estimates of group LASSO depend on the regularization parameter λ that determines which variables are chosen. Therefore, appropriate choice of λ is essential for statistical modeling based on group LASSO. We investigate this problem from a Bayesian point of view and derive an information criterion to choose the value of λ .

We apply the proposed method to the artificial network of ten genes and twenty edges [9]. From the comparison of true positive rates of our proposed method and the methods based on ordinary least square (OLS) and $L2$ penalization, i.e., ridge estimator, under false discovery rate control, the effectiveness of the proposed method is verified especially from the time series data of length less than 75. Our proposed method is also applied to time series gene expression data from Human hela cell cycle [24] and the obtained gene regulatory network is analyzed.

This manuscript is organized as follows: Section 2.1 gives the definition of group LASSO model and its efficient solution. $L1$ regularized spline additive model and its relationship to group LASSO are described in Section 2.2. In Section 2.3, an information criterion of group LASSO is derived and the selection of the regularized parameter λ is shown. Statistical test for Granger causality is illustrated in Section 2.4. In Section 3, our proposed method is applied to the time series data from the artificial network and real data. Finally, we discuss our work in Section 4.

2. $L1$ spline additive regression

2.1. Preliminary

2.1.1. Vector autoregressive model

Given gene expression profile vectors of p genes and T time points $\{\mathbf{x}_1, \dots, \mathbf{x}_T\}$, first order vector autoregressive (VAR(1)) model at time point t is given by:

$$\mathbf{x}_t = A\mathbf{x}_{t-1} + \boldsymbol{\varepsilon}_t, \quad (1)$$

where A is a $p \times p$ autoregressive coefficient matrix, and $\boldsymbol{\varepsilon}$ is a vector of normally distributed noise $\varepsilon_{i,t} \sim N(0, \sigma_i^2)$ for the expression of gene i at t time point. For simplicity of explanation, we use the following notations: $\mathbf{y}_i = (x_{i,2}, \dots, x_{i,T})'$, $X = (\mathbf{x}_1, \dots, \mathbf{x}_{T-1})'$, $\boldsymbol{\beta}_i = (a_{i,1}, \dots, a_{i,p})'$, and $\boldsymbol{\varepsilon}_i = (\varepsilon_{i,1}, \dots, \varepsilon_{i,T-1})'$. By using these notations, autoregressive model each gene i in Equation (1) can be given as:

$$\mathbf{y}_i = X\boldsymbol{\beta}_i + \boldsymbol{\varepsilon}_i. \quad (2)$$

Granger [11] defined a concept of Granger causality, in which a cause cannot come after the effect. Thus, if a gene x_i affects a gene x_j , the expression of gene x_i should help improving the prediction of the expression of gene x_j . To estimate x_i has significant Granger causality to x_j , we test whether the autoregressive coefficient $a_{j,i}$ is 0.

2.1.2. Linear autoregression with grouped covariates

We consider that p covariates are partitioned into disjoint G groups and rewrite the regression model in Equation (2) as:

$$\mathbf{y}_i = \sum_{g=1}^G X_g \boldsymbol{\beta}_{i,g} + \boldsymbol{\varepsilon}_i,$$

where $\boldsymbol{\beta}_{i,g}$ is a sub-vector of $\boldsymbol{\beta}_i$ corresponding to p_g covariates in the g th group, and $X_{i,g}$ is a $(T-1) \times p_g$ matrix of columns corresponding to covariates in the g th group. Like LASSO, group LASSO [26] can put the restriction that all coefficients in some $\boldsymbol{\beta}_{i,g}$'s are simultaneously and exactly zero. The estimates of group LASSO are obtained by solving the following minimization:

$$\arg \min_{\boldsymbol{\beta}_i} \left\{ (\mathbf{y}_i - \sum_g X_g \boldsymbol{\beta}_{i,g})' (\mathbf{y}_i - \sum_g X_g \boldsymbol{\beta}_{i,g}) + \lambda \sum_g \sqrt{\boldsymbol{\beta}_{i,g}' K_{i,g} \boldsymbol{\beta}_{i,g}} \right\}, \quad (3)$$

where $K_{i,g}$ is a $p_g \times p_g$ positive semi-definite matrix. Since Equation (3) is a convex optimization problem but not differentiable at $\boldsymbol{\beta}_{i,g} = \mathbf{0}$, Park and Hastie [21] proposed to use interior point method, introducing dummy variables.

2.1.3. Bayesian information criterion

Given data D , we may select a model M of maximum posterior probability $P(M|D)$ among the models of interest. If prior probability $P(M)$ for model is assumed to be uniform, due to the Bayes theorem, the posterior probability of M is proportional to the marginal likelihood $P(D|M)$. Suppose that a model M is characterized by a parametric model $f(D|\boldsymbol{\theta})$ and prior distribution $\pi(\boldsymbol{\theta})$ for parameter $\boldsymbol{\theta}$. Marginal likelihood of model M with respect to data D is given by:

$$P(D|M) = \int f(D|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}|M) d\boldsymbol{\theta}.$$

Bayesian information criterion [1, 22] was proposed as an approximation of the posterior probability of the model to select the optimal model based on the data:

$$\text{BIC} \approx -2 \log P(M|D) = -2 \log \int f(D|\boldsymbol{\theta}) \pi(\boldsymbol{\theta}|M) d\boldsymbol{\theta}.$$

2.2. L1 regularized spline additive model for gene regulatory network estimation

In nonparametric regression, spline function is often used for constructing regressors. Let $S_{i,j}(x_{j,t})$ be the spline function for the expression of gene j at time point t , $x_{j,t}$. In this study, third-order B-splines are used as base of spline function and spline function $S_{i,j}(x_{j,t})$ for variable $x_{j,t}$ is represented by $\sum \gamma_k b_{i,j,k}(x_{j,t})$. The smoothing spline additive model is obtained by minimizing the loss function:

$$\sum_{t=2}^T (x_{i,t} - \sum_{j=1}^p S_{i,j}(x_{j,t-1}))^2 + \sum_{j=1}^p \lambda \int \left\{ \frac{d^2}{dx^2} S_{i,j}(x) \right\}^2 dx.$$

Lin and Zhang [16], and Bach *et al.* [3] extended the above smoothing spline additive model to L1 regularized spline additive model in which L1 norms of first and second derivatives of spline functions are used as penalization. In L1 regularized spline additive model, the following loss function is optimized:

$$\sum_{t=2}^T \left\{ x_{i,t} - \sum_{j=1}^p S_{i,j}(x_{j,t-1}) \right\}^2 + \lambda \sum_{i=j}^p \left(\int \left\{ \frac{d}{dx} S_{i,j}(x) \right\}^2 dx + \int \left\{ \frac{d^2}{dx^2} S_{i,j}(x) \right\}^2 dx \right)^{1/2}, \quad (4)$$

In B-spline, $\int \left\{ \frac{d}{dx} S_{i,j}(x) \right\}^2 dx$ and $\int \left\{ \frac{d^2}{dx^2} S_{i,j}(x) \right\}^2 dx$ can be given as following forms [4]:

$$\begin{aligned} \int \left\{ \frac{d}{dx} S_{i,j}(x) \right\}^2 dx &= \gamma'_{i,j} D_{1,i,j} \gamma_{i,j}, \\ \int \left\{ \frac{d^2}{dx^2} S_{i,j}(x) \right\}^2 dx &= \gamma'_{i,j} D_{2,i,j} \gamma_{i,j}. \end{aligned}$$

Therefore, we can rewrite Equation (4) by:

$$(\mathbf{y}_i - \sum_{j=1}^p B_{i,j} \gamma_{i,j})' (\mathbf{y}_i - \sum_{j=1}^p B_{i,j} \gamma_{i,j}) + \lambda \sum_{j=1}^p \sqrt{\gamma'_{i,j} E_{i,j} \gamma_{i,j}}, \quad (5)$$

where $\mathbf{y}_i = (x_{i,1}, \dots, x_{i,T-1})'$, $m_{i,j}$ is the number of basis functions for variable x_j , $\gamma_{i,j} = (\gamma_{i,j,1}, \dots, \gamma_{i,j,m_i})'$, $E_{i,j} = D_{1,i,j} + D_{2,i,j}$, and $B_{i,j}$ is a $(T-1) \times m_{i,j}$ matrix:

$$B_{i,j} = \begin{bmatrix} b_{i,j,1}(x_{j,1}) & \cdots & b_{i,j,m_{i,j}}(x_{j,1}) \\ \vdots & \ddots & \vdots \\ b_{i,j,1}(x_{j,T-1}) & \cdots & b_{i,j,m_{i,j}}(x_{j,T-1}) \end{bmatrix}.$$

In the L1 regularized spline additive model, since we would like to evaluate whether all coefficients of some splines are simultaneously and exactly zero, we can thus use the procedure based on group LASSO.

However, use of dummy variables increases the number of variables to be concerned. In addition, unstable constraints caused by dummy variables induce the slow

convergence. Thus, we propose to convert the optimization problem in Equation (5) to:

$$\arg \min_{\gamma} (\gamma_i' B_i' B_i \gamma_i) \quad \text{subject to} \quad \lambda^2/4 \geq (\mathbf{y}_i - B_i \gamma_i)' B_{i,j} E_{i,j}^{-1} B_{i,j}' (\mathbf{y}_i - B_i \gamma_i), \quad (6)$$

where $B_i = (B_{i,1}, \dots, B_{i,p})$, and $\gamma_i = (\gamma_{i,1}', \dots, \gamma_{i,p}')'$. The optimization problem in Equation (6) can be solved by interior point method without using dummy variables. See Appendix A for details.

2.3. Bayesian information criterion for nonparametric group LASSO regression

Selection of regularization parameter λ is important for variable selection and coefficient estimation in group LASSO. We derive Bayesian information criterion for $L1$ spline additive model and λ minimizing the criterion. From the view point of Bayesian statistics, probabilistic model of $L1$ spline additive model can be characterized as likelihood function $f(\mathbf{y}_i | B_i, \gamma_i, \sigma_i^2)$ of linear regression with product of Laplacian prior $\pi_{i,j}(\gamma_{i,j} | \sigma_i^2, \lambda)$ for $\gamma_{i,j}$ given by:

$$f(\mathbf{y}_i | B_i, \gamma_i, \sigma_i^2) = \frac{1}{\sqrt{2\pi\sigma_i^2}^{T-1}} \exp \left\{ -\frac{1}{2\sigma_i^2} (\mathbf{y}_i - B_i \gamma_i)' (\mathbf{y}_i - B_i \gamma_i) \right\}, \quad (7)$$

$$\pi_{i,j}(\gamma_{i,j} | \sigma_i^2, \lambda) = L_{i,j} \exp \left(-\frac{\lambda}{2\sigma_i^2} \sqrt{\gamma_{i,j}' E_{i,j} \gamma_{i,j}} \right), \quad (8)$$

where $L_{i,j} = \frac{1}{2\pi^{p_{i,j}/2}} \left(\frac{\lambda}{2\sigma_i^2} \right)^{p_{i,j}} \det(E_{i,j})^{1/2} \frac{\Gamma(p_{i,j}/2)}{\Gamma(p_{i,j})}$. Using Equations (7) and (8) in Equation (4), the posterior probability of the model based on group LASSO can be given as:

$$P(D|M) = \int f(\mathbf{y}_i | B_i, \gamma_i, \sigma_i^2) \prod_j \pi_{i,j}(\gamma_{i,j} | \sigma_i^2, \lambda) d\gamma_i. \quad (9)$$

Note that the variance σ_i^2 is considered to be known. For unknown σ_i^2 , we use $\hat{\sigma}_i^2 = \sum_{t=2}^T (y_{i,t} - \hat{y}_{i,t})^2 / (T-1)$ as the estimator of σ_i^2 . Hereafter, we omit σ_i^2 and λ in $f(\mathbf{y}_i | B_i, \gamma_i, \sigma_i^2)$ and $\pi_{i,j}(\gamma_{i,j} | \sigma_i^2, \lambda)$ if no confusion occurs.

In the following, we explain how the integration in Equation (9) is solved. Let A_i be a set of group vector $\gamma_{i,j}$ estimated as non-zero in group LASSO. If $\gamma_{i,k}$ is not in A_i , i.e., estimated as exactly $\mathbf{0}$ in group LASSO, it implies Laplacian prior of $\gamma_{i,k}$ is much stronger than likelihood function. Thus, we approximately calculate

the integration these $\gamma_{i,k}$'s $\notin A_i$, ignoring $\gamma_{i,k}$ in the likelihood function:

$$\begin{aligned} & \int f(\mathbf{y}_i | B_i, \gamma_i) \pi_{i,j}(\gamma_{i,k}) d\gamma_{i,k} \\ &= \int \frac{1}{\sqrt{2\pi\sigma_i^2}^{T-1}} \exp \left\{ -\frac{1}{2\sigma_i^2} (\mathbf{y}_i - \sum_j B_{i,j} \gamma_{i,j})' (\mathbf{y}_i - \sum_j B_{i,j} \gamma_{i,j}) \right\} \pi_{i,k}(\gamma_{i,k}) d\gamma_{i,k} \\ &\approx \frac{1}{\sqrt{2\pi\sigma_i^2}^{T-1}} \exp \left\{ -\frac{1}{2\sigma_i^2} (\mathbf{y}_i - \sum_{j \neq k} B_{i,j} \gamma_{i,j})' (\mathbf{y}_i - \sum_{j \neq k} B_{i,j} \gamma_{i,j}) \right\}. \end{aligned}$$

After integrating all $\gamma_{i,j}, j \notin A_i$, we have:

$$\text{BIC}_{\text{GL}} \approx -2 \log \int f(\mathbf{y}_i | B_i, \gamma_{A_i}) \pi_{A_i}(\gamma_{A_i}) d\gamma_{A_i}, \quad (10)$$

where

$$\begin{aligned} f_{A_i}(\mathbf{y}_i | B_{A_i}, \gamma_{A_i}) &= \frac{1}{\sqrt{2\pi\sigma_i^2}^{T-1}} \exp \left\{ -\frac{1}{2\sigma_i^2} (\mathbf{y}_i - \sum_{j \in A_i} B_{i,j} \gamma_{i,j})' (\mathbf{y}_i - \sum_{j \in A_i} B_{i,j} \gamma_{i,j}) \right\}, \\ &= \frac{1}{\sqrt{2\pi\sigma_i^2}^{T-1}} \exp \left\{ -\frac{1}{2\sigma_i^2} (\mathbf{y}_i - B_{A_i} \gamma_{A_i})' (\mathbf{y}_i - B_{A_i} \gamma_{A_i}) \right\}, \\ \pi_{A_i}(\gamma_{A_i}) &= \prod_{j \in A_i} \pi_{i,j}(\gamma_{i,j}), \end{aligned}$$

Here, B_{A_i} is a sub-matrix B_i for covariates in A_i , and γ_{A_i} is a sub-vector of γ_i for covariates in A_i . For the integration with respect to $\gamma_{i,j}, j \in A_i$, Laplace approximation is used. By Laplace approximation, the integration is approximated as:

$$\int \exp \{q(\boldsymbol{\theta})\} d\boldsymbol{\theta} \approx \exp \{q(\hat{\boldsymbol{\theta}})\} (2\pi)^{p/2} / \left| -\frac{\partial^2 q(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}'} \right|,$$

where $\hat{\boldsymbol{\theta}} = \arg \max_{\boldsymbol{\theta}} q(\boldsymbol{\theta})$.

Applying Laplace approximation to Equation (10), we have:

$$\begin{aligned} & \log \int f_{A_i}(\mathbf{y}_i | B_{A_i}, \gamma_{A_i}) \pi_{A_i}(\gamma_{A_i}) d\gamma_{A_i} \\ & \approx \log f_{A_i}(\mathbf{y}_i | B_{A_i}, \hat{\gamma}_{A_i}) \pi_{A_i}(\hat{\gamma}_{A_i}) + \frac{|A_i|}{2} \log 2\pi - \log \det J(\hat{\gamma}_{A_i}), \end{aligned}$$

where $|A_i|$ is the length of γ_{A_i} , and $J(\hat{\gamma}_{A_i})$ is a $|A_i| \times |A_i|$ matrix given as:

$$\begin{aligned} J(\hat{\gamma}_{A_i}) &= -\frac{\partial^2}{\partial \gamma_{A_i} \partial \gamma_{A_i}'} \log f_{A_i}(\mathbf{y}_i | B_{A_i}, \gamma_{A_i}) \pi_{A_i}(\gamma_{A_i}) \Big|_{\gamma_{A_i} = \hat{\gamma}_{A_i}} \\ &= \frac{1}{\sigma^2} \left(B_{A_i}' B_{A_i} + \frac{\lambda}{2} \text{diag} \left[\frac{E_{i,j}}{\sqrt{\hat{\gamma}_{i,j}' E_{i,j} \hat{\gamma}_{i,j}}} - \frac{E_{i,j} \hat{\gamma}_{i,j} \{E_{i,j} \hat{\gamma}_{i,j}\}'}{\sqrt{\hat{\gamma}_{i,j}' E_{i,j} \hat{\gamma}_{i,j}}^3} \right] \right) \end{aligned} \quad (11)$$

Thus, BIC_{GL} is given as:

$$\begin{aligned} \text{BIC}_{\text{GL}} &= -2 \log f_{A_i}(\mathbf{y}_i | B_{A_i}, \hat{\gamma}_{A_i}) \pi_{A_i}(\hat{\gamma}_{A_i}) - |A_i| \log 2\pi + 2 \log \det J(\hat{\gamma}_{A_i}) \\ &= -(T - 1 + 2|A_i|) \log 2 - (T - 1) \log \pi - (T - 1 + |A_i|) \log \sigma^2 + 2|A_i| \log \lambda \\ &\quad + 2 \sum_{j \in A_i} \left(\log \Gamma(p_{i,j}/2) - \log \Gamma(p_{i,j}) + \frac{1}{2} \log |E_{i,j}| \right) \\ &\quad - \frac{1}{\sigma^2} \left\{ (\mathbf{y}_i - B_{A_i} \hat{\gamma}_{A_i})' (\mathbf{y}_i - B_{A_i} \hat{\gamma}_{A_i}) + \lambda \sum_{j \in A_i} \sqrt{\hat{\gamma}_{i,j}' E_{i,j} \hat{\gamma}_{i,j}} \right\} - \log |J(\hat{\gamma}_{A_i})|. \end{aligned} \quad (12)$$

2.4. Wald test for Granger causality

The variables selected by group LASSO are considered as the candidate variables having Granger causality to the response. In order to control false discovery rate of those candidates, we test the coefficients of basis functions $\gamma_{i,j}$ corresponding to each selected variable x_j . Usually, to test whether all the coefficients of grouped variables in linear regression are simultaneously zero, i.e., $\gamma_{i,j} = \mathbf{0}$, we may use Wald test. However, Wald test is based on asymptotic normality of maximum likelihood estimators. Since group LASSO is not a maximum likelihood method due to the existence of Laplacian prior, it is impossible to use Wald test directly for the estimators of group LASSO.

Konishi and Kitagawa [15] considered that a parameter θ is represented as a functional $T(G)$ for the true distribution $G(x)$ and the estimator $\hat{\theta}$ for θ is given by $T(\hat{G})$, where $\hat{G}$ is the empirical distribution of G . The asymptotic normality of $T(\hat{G})$ was shown:

$$\sqrt{n}(T(\hat{G}) - T(G)) \rightarrow N(0, \int T^{(1)}(G) \{T^{(1)}(G)\}' dG(x)) \quad \text{in law,}$$

where $T^{(1)}(x, G)$ is influence function for $T^{(1)}$ given by:

$$T^{(1)}(x, G) = \lim_{\epsilon \rightarrow 0} \frac{T((1 - \epsilon)G + \epsilon\delta_x) - T(G)}{\epsilon}.$$

Here, δ_x is a distribution function having a probability of 1 at point x . Since various estimators including maximum likelihood estimator and maximum penalized likelihood estimator can be represented as $T(\hat{G})$, we exploit this property for Wald test.

Let $T_{A_i}(G)$ be a functional for γ_{A_i} , group LASSO coefficients in A_i . Due to the KKT conditions for group LASSO estimators [3, 21], functional $T_{A_i}(G)$ for γ_{A_i} satisfies

$$\int \Psi_{A_i}(y, T_{A_i}(G)) dG(y) = 0,$$

where

$$\Psi_{A_i}(y, T_{A_i}(G)) = \frac{\partial}{\partial \gamma_{A_i}} \log f_{A_i}(y | \mathbf{b}_{A_i}, \gamma_{A_i}) \pi_{A_i}(\gamma_{A_i}).$$

In addition, it is natural to assume that set of groups selected A_i by group LASSO is invariant for small perturbation $\epsilon\delta_x$ to the distribution [23]. Thus, for small ϵ , we have:

$$\int \Psi_{A_i}(y, \mathbf{T}_{A_i}((1-\epsilon)G + \epsilon\delta_x)) d((1-\epsilon)G(y) + \epsilon\delta_x(y)) = 0.$$

By following the derivation of the influence function for M -estimator in [15], influence function $\mathbf{T}_{A_i}^{(1)}(x, G)$ for $\mathbf{T}_{A_i}(G)$ is given as:

$$\mathbf{T}_{A_i}^{(1)}(x, G) = \left\{ \int \frac{\partial^2}{\partial \gamma_{A_i}} \Psi_{A_i}(y, \gamma_{A_i}) \Big|_{\gamma_{A_i} = \mathbf{T}_{A_i}(G)} dG(y) \right\}^{-1} \Psi_{A_i}(x, \mathbf{T}_{A_i}(G)).$$

By using empirical distribution $\hat{G}$ for G , we have the covariance matrix of group LASSO estimator:

$$\begin{aligned} \Sigma(\hat{\gamma}_{A_i}) &= \int \mathbf{T}_{A_i}^{(1)}(y, \hat{G}) \left\{ \mathbf{T}_{A_i}^{(1)}(y, \hat{G}) \right\}' d\hat{G}(y) \\ &= \frac{1}{n} J(\hat{\gamma}_{A_i})^{-1} I(\hat{\gamma}_{A_i}) J(\hat{\gamma}_{A_i})^{-1}, \end{aligned}$$

where

$$I(\hat{\gamma}_{A_i}) = \frac{1}{n\sigma^4} \left(B_{A_i}' \Lambda^2 B_{A_i} - \frac{\lambda}{2} \omega_{A_i} \mathbf{1}' \Lambda B_{A_i} - \frac{\lambda}{2} B_{A_i}' \Lambda \mathbf{1} \omega_{A_i}' + \frac{\lambda^2}{4n} \omega_{A_i} \omega_{A_i}' \right),$$

and $J(\hat{\gamma}_{A_i})$ is given by Equation (11). Here, $\Lambda = \text{diag}[x_{i,t} - \hat{\gamma}_{A_i}' \mathbf{b}_{A_i,t}]$, and ω_{A_i} is a vector comprised of $\frac{E_{i,j} \hat{\gamma}_{i,j}}{\sqrt{\hat{\gamma}_{i,j}' E_{i,j} \hat{\gamma}_{i,j}}}$, $j \in A_i$.

For the null hypothesis, $H_0: R\gamma_{A_i} = \mathbf{r}$, we can derive Wald statistics W_{GL} for group LASSO coefficients as follows:

$$W_{GL} = (R\hat{\gamma}_{A_i} - \mathbf{r})' \{R\Sigma(\hat{\gamma}_{A_i})R'\}^{-1} (R\hat{\gamma}_{A_i} - \mathbf{r}) \rightarrow \chi_{\text{rank}(R)}^2 \quad \text{in law.}$$

For example, suppose that $\gamma_{A_i} = (\gamma'_{i,1}, \gamma'_{i,2}, \gamma'_{i,3}, \gamma'_{i,4})'$ and we would like to evaluate the null hypothesis $H_0: \gamma_2 = \mathbf{0}$, we set $R = (O_{m_{i,2}, m_{i,1}}, I_{m_{i,2}}, O_{m_{i,2}, m_{i,3}}, O_{m_{i,2}, m_{i,4}})$ and $\mathbf{r} = \mathbf{0}$, where $O_{m,n}$ is an $m \times n$ matrix whose elements are zero and I_m is the identity matrix of size m .

3. Numerical examples

3.1. Simulation data examples

We use an artificial network of ten genes having twenty linear and nonlinear relationships and show the performance of our proposed method, $L1$ NVAR. For the competitors of $L1$ NVAR, OLS based nonlinear vector autoregressive based (NVAR) model [9] and nonlinear vector autoregressive model with $L2$ penalization ($L2$ NVAR) are employed. In $L2$ NVAR, $L1$ penalization in Equation (5) is replaced by $L2$ penalization, and regularization parameter is selected by Bayesian information criterion. A Wald test derived in a similar manner for $L1$ NVAR is used to

capture significant Granger causalities of $L2$ NVAR. Twenty edges in the artificial network are set as follows:

$$\begin{cases} x_{1,t} = 0.5x_{1,t-1} + \varepsilon_{1,t} \\ x_{2,t} = 0.6x_{2,t-1} + \varepsilon_{2,t} \\ x_{3,t} = 0.7x_{3,t-1} + \varepsilon_{3,t} \\ x_{4,t} = 0.8x_{4,t-1} + \varepsilon_{4,t} \\ x_{5,t} = 0.9x_{5,t-1} + \varepsilon_{5,t} \\ x_{6,t} = \sin(x_{1,t-1}) + 0.5x_{2,t-1} - 0.5x_{9,t-1} + 2 + \varepsilon_{6,t} \\ x_{7,t} = 2\cos(x_{2,t-1}) - 2\sin(x_{3,t-1}) + 0.6x_{10,t-1} + \varepsilon_{7,t} \\ x_{8,t} = 0.8\cos(x_{3,t-1}) + 0.6x_{4,t-1} + \cos(x_{6,t-1}) + 1 + \varepsilon_{8,t} \\ x_{9,t} = \sin(x_{4,t-1}) + \cos(x_{5,t-1}) - 0.8x_{7,t-1} + \varepsilon_{9,t} \\ x_{10,t} = \sin(x_{1,t-1}) - 0.8x_{5,t-1} + \cos(x_{8,t-1})\varepsilon_{10,t} \end{cases}$$

Graphical representation of the artificial network drawn with Cell Illustrator [19, 27] is shown in Figure 1. From the artificial network, we generate time series data of various length $\{10, 20, 30, 40, 50, 75, 100\}$ and apply NVAR, $L1$ NVAR and $L2$ NVAR to them. Since time series length is not sufficient for the estimation, the number of B-splines is set to four. We repeat the experiment 100 times for each time series length. Granger causalities are estimated under false discovery rate 5%.

Fig. 1. An artificial network of 10 genes and 20 edges used in a simulation study. Some of edges represent nonlinear causality.

First, in order to verify the false discovery rate control, we calculate the true false discovery rates by comparing edges in the artificial network and significantly estimated Granger causalities in NVAR, $L1$ NVAR, and $L2$ NVAR for each time series length. Those true false discovery rates are summarized in Table 1. When the length of time series is short, i.e., data is not enough, false discovery rate is not controlled within 5% in $L1$ NVAR. In $L2$ NVAR, false discovery rate is exploded for all the time series length. This problem may be related to the convergence of covariance matrix for coefficients.

In Wald tests for *L1* NVAR and *L2* NVAR, asymptotic normality is used for the derivation of covariance matrix. On the other hand, covariance matrix in NVAR coincides with its unbiased estimator, and thus asymptotic normality of maximum likelihood estimator is actually not used. This hypothesis is supported by the fact that false discovery rate is controlled for all the time series length in NVAR. For relatively long time series data, e.g., times series data of length 50, false discovery rate is correctly controlled in *L1* NVAR, while it is out of control in *L2* NVAR. In *L1* NVAR, some variables are dropped in estimation, and thus convergence of covariance matrix is faster than the case considering all the variables. However, in *L2* NVAR, no variable is dropped for the estimation. Thus, false discovery rate is converging to 5 % as time series length increase, but it is still not converged in time series data of sufficient length.

Table 1. True false discovery rates obtained by comparing the artificial network and • estimated Granger causalities in *L1* NVAR and NVAR under false discovery rate controlled 5 % (mean $\pm$ standard deviation, in %).

Time series length	NVAR	<i>L1</i> NVAR	<i>L2</i> NVAR
10	-	40.54 $\pm$ 35.71	67.13 $\pm$ 20.25
20	-	18.72 $\pm$ 16.23	65.11 $\pm$ 13.55
30	-	11.64 $\pm$ 11.04	61.23 $\pm$ 8.20
40	-	7.40 $\pm$ 8.95	55.59 $\pm$ 8.27
50	2.23 $\pm$ 9.02	4.13 $\pm$ 5.88	49.84 $\pm$ 8.49
75	3.83 $\pm$ 6.02	3.48 $\pm$ 4.42	39.67 $\pm$ 10.01
100	4.37 $\pm$ 5.21	2.18 $\pm$ 4.03	31.91 $\pm$ 8.88

True positive rates of estimated Granger causalities in NVAR, *L1* NVAR, and *L2* NVAR under false discovery rate 5 % are compared in Figure 2(a). Since false discovery rate in *L2* NVAR is completely out of control, we also calculate the true positive rates obtained by controlling true false discovery rate within 5 % in Figure 2(b). According to the results in Figure 2(a), *L1* NVAR overwhelms NVAR for time series data of length less than 75. On the other hand NVAR gives slightly better performance than *L1* NVAR for long time series data. However, our interest and design of *L1* NVAR is the estimation of Granger causality using insufficient time series data, and this point do not have to be concerned. *L2* NVAR seems to give the best performance among the three methods in Figure 2(a), but under the control of true false discovery rate within 5 %, *L1* NVAR gives the best performance among them. Therefore we conclude that *L1* NVAR is the best of option among them to estimate the nonlinear and nonmonotonic gene regulatory network from the short time series data.

3.2. Application of expression data of human hela cell

We apply the proposed method to the time series gene expression data of Human hela cell [24]. 48 times points for 94 genes selected by [8] are used in our study. The

number of B-splines is set to four, that is, the number of parameters is approximately eight times as many as the length of data. Figure 3 shows the estimated gene regulatory network under false discovery rate 5%.

In the following, significantly estimated Granger causalities and biologically reported facts are compared:

- Transcription factor NF- κ B is known to work as the central mediator of the human immune response [20]. ICAM-1, Cyclin D1, A20, IAP are reported to be target genes of NF- κ B [20]. In our estimated network, IAP is estimated as the Granger causal of NF- κ B, Cyclin D1, A20, and ICAM-1. PKR is known to activate NF- κ B. In the estimated network, PKR and NF- κ B have connection with the Granger causal of IAP.
- Bcl-2 is known to inhibit PERP-induced cell death [2]. This regulation coincides the Granger causality, Bcl-2 $\rightarrow$ PERP in the estimated network.
- E2F1 is a transcription factor known to regulate the transcription of Cyclin E1 [10]. The estimated Granger causality Cyclin E1 $\rightarrow$ E2F1 is opposite to the biologically known fact.
- Puma is known to bind Mcl-1 to maintain the expression level of Mcl-1 [6]. This coincides the estimated Granger causality Puma $\rightarrow$ Mcl-1.
- In colon cancer cell, over expression of E2F-1 is reported to down-regulate Mcl-1, up-regulate c-myc, and induce apoptosis [5]. In hela cell, this mechanism may be different, but interestingly, in the estimated network there is a completely opposite Granger causality path, c-myc $\rightarrow$ Mcl-1 $\rightarrow$ PUMA $\rightarrow$ E2F-1.
- Fas is a well known target gene of P53. On the other hand, Fas is Granger causal of P53 in the estimated network.
- P21 is estimated to have self loop. This self loop is also detected by NVAR based on OLS and verified [9].

Fig. 2. True positive rates for estimated Granger causalities in NVAR, L1 NVAR, and L2 NVAR. (a) True positive rates under false discovery control 5 %. (b) True positive rates obtained by controlling true false discovery rate within 5 %.

Fig. 3. The resulting gene regulatory network from *L1* NVAR under false discovery rate 5%.

4. Discussion

In this study, we proposed *L1* regularized nonlinear vector autoregressive (*L1* NVAR) models to estimate gene regulatory networks from short time series microarray data. The difficulty in short time series microarray data is that the number of parameters in the model is greater than the number of samples and this leads overfitting and decreases the predictive power of the model. To overcome the difficulty, we apply group LASSO technique to build nonparametric regressions for gene regulation models. For the edge selection in gene regulatory networks, we derived an information criterion from a Bayesian point of view. For the post-treatment of the obtained model to find Granger causality, we established Wald test for regression parameters estimated by group LASSO procedure.

The simulation study indicated that the proposed method outperforms nonlinear vector autoregressive model estimated by ordinary least squares in terms of the accuracy of the estimated networks. We also applied the proposed method to human hela cell time series microarray data. Some of the estimated edges are supported by the literature, but we observed that some edges have opposite direction. We need to investigate this point in future paper, the relationship between biological regulation

and Granger causality.

References

- [1] Akaike, H., Likelihood and the Bayes procedure, *Bayesian Statistic* (eds. J.M. Bernardo, M.H. DeGroot, D.V. Lindley and A.F.M. Smith), Univ. Press, Valencia, Spain, 1980.
- [2] Attardi, L.D., Reczek, E.E., Cosmas, C., Demicco, E.G., McCurrach, M.E., Lowe, S.W., and Jacks, T., Perp, an apoptosis-associated target of p53, is a novel member of the PMP-22/gas3 family, *Genes & Development*, 14(6): 704–718, 2000.
- [3] Bach, F.R., Thibaux, R., and Jordan, M.I., Computing regularization paths for learning multiple kernels, *In Advanced in Neural Information Processing System 17*, 2004.
- [4] Eilers, P., and Marx, B., Flexible smoothing with *B*-splines and penalties (with discussion), *Statistical Science*, 11: 89–121, 1996.
- [5] Elliott, M.J., Dong, Y.B., Yang, H., and McMasters, K.M., E2F-1 up-regulates c-myc and p14arf and induces apoptosis in colon cancer cells, *Clinical Cancer Research*, 7: 3590–3597, 2001.
- [6] Ewings, K.E., Hadfield-Moorhouse, K., Wiggins, C.M., Wickenden, J.A., Balmanno, K., Gilley, R., Degenhardt, K., White, E., and Cook, S.J., Erk1/2-dependent phosphorylation of bim promotes its rapid dissociation from mcl-1 and bcl-xl, *EMBO Journal*, 26(12): 2856–2867, 2007.
- [7] Fujita, A., Sato, J.R., Garay-Malpartida, H.M., Morettin, P.A., Sogayar, M.C., Ferreira, C.E., Time-varying modeling of gene expression regulatory networks using the wavelet dynamic vector autoregressive method, *Bioinformatics*, 23(13):1623–1630, 2007.
- [8] Fujita, A., Sato, J.R., Garay-Malpartida, H.M., Morettin, P.A., Yamaguchi, R., Miyano, S., Sogayar, M.C., Ferreira, C.E., Modeling gene expression regulatory networks with the sparse vector autoregressive model, *BMC Systems Biology*, 1:39, 2007.
- [9] Fujita, A., Sato, J.R., Garay-Malpartida, H.M., Sogayar, M.C., Ferreira, C.E., and Miyano, S., Modeling nonlinear gene regulatory networks from time series gene expression data, *Journal of Bioinformatics and Computational Biology*, in press, 2008.
- [10] Geng, Y., Eaton, E.N., Picón, M., Roberts, J.M., Lundberg, A.S., Gifford, A., Sardet, C., Weinberg, R.A., Regulation of cyclin E transcription by E2Fs and retinoblastoma protein, *Oncogene*, 12(6): 1173–1180, 1996.
- [11] Granger, C.W.J., Investigating causal relations by econometric models and cross-spectral methods, *Econometrica* 37: 424–438, 1969.
- [12] Hirose, O., Yoshida, R., Imoto, S., Yamaguchi, R., Higuchi, T., Charnock-Jones, S.D., Print, C., and Miyano, S., Statistical inference of transcriptional module-based gene networks from time course gene expression profiles by using state space models, module finder on gene expression profiles, *Bioinformatics*, 24(7): 932–942, 2008.
- [13] Kim, S., Imoto, S., and Miyano, S., Dynamic Bayesian network and nonparametric regression for nonlinear modeling of gene networks from time series gene expression data, *Biosystems*, 75(1–3): 57–65, 2004.
- [14] Kim, S.J., Koh, K., and Lustig, M., An interior-point method for large-scale l_1 -regularized least squares, *IEEE Journal on Selected Topics in Signal Processing*, 4(1): 606–617, 2007.
- [15] Konishi, S., and Kitagawa, G., Generalized information criteria in model selection, *Biometrika*, 83(4): 875–890, 1996.
- [16] Lin, Y., and Zhang, H.H., Component selection and smoothing in multivariate nonparametric regression, *Annals of Statistics*, 34(5): 2272–2297, 2006.

- [17] Lobo, M.S., Vandenberghe, L., Boyd, S., Applications of Second-order Cone Programming, *Linear Algebra and its Applications*, 284: 193–228, 1998.
- [18] Nachman, I., Regev, A., and Friedman, N., Inferring quantitative models of regulatory networks from expression data, *Bioinformatics*, 4(20): i248–i256, 2004.
- [19] Nagasaki, M., Doi, A., Matsuno, H., Miyano, S., Genomic Object Net: I. A platform for modeling and simulating biopathways, *Applied Bioinformatics*, 2: 181–184, 2003.
- [20] Pahl, H.L., Activators and target genes of rel/nf- κ B transcription factors, *Oncogene*, 18: 6853–6866, 1999.
- [21] Park, M., and Hastie, T., Regularization path algorithm for detecting gene interactions, preprint, 2006.
- [22] Schwarz, G., Estimating the dimension of a model, *Annals of Statistics*, 6: 461–464, 1978.
- [23] Shimamura, T., Model selection with empirical Bayes criteria L_1 -regularization, Ph D dissertation, Hokkaido University, 2007.
- [24] Whitfield, M.L., Sherlock, G., Saldanha, A.J., Murray, J.I., Ball C.A., Alexander, K.E., Matese, J.C., Perou, C.M., Hurt, M.M., Brown, P.O., Botstein, D., Identification of genes periodically expressed in the human cell cycle and their expression in tumors, *Molecular Biology of the Cell*, 13: 1977–2000, 2002.
- [25] Yamaguchi, R., Yoshida, R., Imoto, S., Higuchi, T., Miyano, S., Finding module-based gene networks in time-course gene expression data with state space models, *IEEE Signal Processing Magazine*, 24: 37–46, 2007.
- [26] Yuan, M., and Lin, Y., Model selection and estimation in regression with grouped variables, *Journal of The Royal Statistical Society Series B*, 68(1): 49–67, 2006.
- [27] <http://www.cellillustrator.com/>

Appendix A.

We define optimization problem for L_1 regularized spline additive model in Equation (5) as primal problem $\mathcal{P}$:

$$\mathcal{P}(\gamma_i) = \min_{\gamma_i} (\mathbf{y}_i - B_i \gamma_i)' (\mathbf{y}_i - B_i \gamma_i) + \lambda \sum_j \sqrt{\gamma'_{i,j} E_{i,j} \gamma_{i,j}}, \quad (\text{A.1})$$

and derive its dual. Letting $\mathbf{z} = B_i \gamma_i - \mathbf{y}_i$, we can rewrite Equation (A.1) as:

$$\min_{\gamma} \left\{ \mathbf{z}' \mathbf{z} + \lambda \sum_j \sqrt{\gamma_{i,j} E_{i,j} \gamma_{i,j}} \right\} \quad \text{subject to} \quad \mathbf{z} = B_i \gamma_i - \mathbf{y}_i.$$

Thus, dual problem can be given as:

$$\begin{aligned} \mathcal{D}(\alpha) &= \max_{\alpha} \inf_{\mathbf{z}, \gamma_i} \mathbf{z}' \mathbf{z} + \lambda \sum_j \sqrt{\gamma'_{i,j} E_{i,j} \gamma_{i,j}} + \alpha' (B_i \gamma_i - \mathbf{y}_i - \mathbf{z}) \\ &= \max_{\alpha} \inf_{\gamma_i} -\frac{\alpha' \alpha}{4} - \alpha' \mathbf{y}_i + \lambda \sum_j \sqrt{\gamma'_{i,j} E_{i,j} \gamma_{i,j}} + \alpha' B_i \gamma_i \\ &= \begin{cases} \max_{\alpha} \left\{ -\frac{\alpha' \alpha}{4} - \alpha' \mathbf{y}_i \right\} & \text{if } \lambda^2 \geq \alpha' B_{i,j} E_{i,j}^{-1} B'_{i,j} \alpha \\ \infty & \text{otherwise} \end{cases} \end{aligned}$$

where $\alpha = 2z = 2(B_i\gamma_i - \mathbf{y}_i)$. The dual problem can be converted as:

$$\begin{aligned} \tilde{D}(\gamma_i) &= \arg \min_{\gamma_i} (\gamma_i' B_i' B_i \gamma_i) \\ \text{subject to } \lambda^2/4 &\geq (\mathbf{y}_i - B_i\gamma_i)' B_{i,j} E_{i,j}^{-1} B_{i,j}' (\mathbf{y}_i - B_i\gamma_i). \end{aligned}$$

By using barrier function $\phi_j(\gamma_i)$:

$$\phi_j(\gamma_i) = \log \{ \lambda^2/4 - (\mathbf{y}_i - B_i\gamma_i)' B_{i,j} E_{i,j}^{-1} B_{i,j}' (\mathbf{y}_i - B_i\gamma_i) \},$$

we have the following potential function to be minimized in interior point method:

$$\psi(\gamma_i) = \gamma_i' B_i' B_i \gamma_i + \mu \sum_j \phi_j(\gamma_i).$$

The dual gap η of the interior point method can be given as:

$$2(\gamma_i' B_i' B_i \gamma_i - \mathbf{y}_i' B_i \gamma_i + \mathbf{y}_i' \mathbf{y}_i) + \lambda \sum_j \sqrt{\gamma_i' B_{i,j} E_{i,j} \gamma_i}.$$

Here, we show a brief algorithm of interior point method:

- step 1** Set barrier strength μ , acceleration rate ν , and stopping criterion τ to some positive values, for instance, $\mu = 0.2$, $\nu = 6.5$, and $\tau = 0.00001$.
- step 2** Calculate Newton direction and select step size by the Armijo rule.
- step 3** Update γ_i according to the Newton direction and step size obtained at step 2.
- step 3** If dual gap η is less than τ , finish the algorithm and output γ_i
- step 4** Update μ as $\mu \leftarrow \mu/\nu$ and go to step 2.

For the details of interior point method, see [14, 17].