RESEARCH ARTICLE

Using single-index ODEs to study dynamic gene regulatory network

Qi Zhang[1], **Yao Yu**[2], **Jun Zhang**[3], **Hua Liang**[4]*

1 Department of Statistics, Qingdao University, Qingdao, China, **2** Department of Biostatistics and Computational Biology, University of Rochester School of Medicine and Dentistry, Rochester, New York, United States of America, **3** Institute of Statistical Sciences at Shenzhen University, Shenzhen University, Shenzhen, China, **4** Department of Statistics, George Washington University, Washington, D.C., United States of America

* hliang@gwu.edu

OPEN ACCESS

Citation: Zhang Q, Yu Y, Zhang J, Liang H (2018) Using single-index ODEs to study dynamic gene regulatory network. PLoS ONE 13(2): e0192833. https://doi.org/10.1371/journal.pone.0192833

Editor: Alan M. Moses, University of Toronto, CANADA

Received: February 28, 2017

Accepted: January 31, 2018

Published: February 23, 2018

Data Availability Statement: All relevant data are available in the Supporting Information files.

Funding: Liang's research was partially supported by NSF grants DMS-1418042 and DMS-1620898, and by Award Number 11529101, made by National Natural Science Foundation of China. This study was supported by all these external funding, and there was no additional external funding received for this study.

Competing interests: The authors have declared that no competing interests exist.

Abstract

With the development of biotechnology, high-throughput studies on protein-protein, protein-gene, and gene-gene interactions become possible and attract remarkable attention. To explore the interactions in dynamic gene regulatory networks, we propose a single-index ordinary differential equation (ODE) model and develop a variable selection procedure. We employ the smoothly clipped absolute deviation penalty (SCAD) penalized function for variable selection. We analyze a yeast cell cycle gene expression data set to illustrate the usefulness of the single-index ODE model. In real data analysis, we group genes into functional modules using the smoothing spline clustering approach. We estimate state functions and their first derivatives for functional modules using penalized spline-based nonparametric mixed-effects models and the spline method. We substitute the estimates into the single-index ODE models, and then use the penalized profile least-squares procedure to identify network structures among the models. The results indicate that our model fits the data better than linear ODE models and our variable selection procedure identifies the interactions that may be missed by linear ODE models but confirmed in biological studies. In addition, Monte Carlo simulation studies are used to evaluate and compare the methods.

Introduction

Gene regulatory networks (GRN) are complex and dynamic systems in nature. They are composed of genes that interact with each other and with other substances inside cells, such as RNAs and proteins. Over the past few decades, a variety of methods have been proposed to model GRN. Commonly used models include information theory models, Boolean networks, ordinary differential equation (ODE) models, and Bayesian networks [1]. Information theory models [2–4] construct network architecture on correlation coefficients. Such models are simple and have a low computation cost, but cannot take into account the dynamic processes and situations when multiple genes participate in regulations. Boolean networks [5–7] are discrete dynamic networks and easy to understand, but have limitations because their networks' nodes are binary states: “off” or “on”. Due to these simplifying assumptions, the study of kinetic gene regulation is still challenging because of the complexity of the biological process [8].

The Bayesian networks [9–12] integrate biological knowledge and measurements to infer network structures. But the estimated results obtained from Bayesian networks depend on the quality and completeness of prior knowledge. As pointed out by [13], the existing ODE models and associated methods used to study GRN are flexible but are limited to small scale gene expression levels. ODE models describe the dynamic behaviors of GRN in a quantitative manner and represent gene expression level changes by functions of gene expression levels:

$$\frac{dX_k(t)}{dt} = F(t, \mathbf{X}(t), \boldsymbol{\theta}), \;\; k = 1, \ldots, p, \tag{1}$$

where $\mathbf{X}(t) = (X_1(t), \cdots, X_p(t))^T$ represents gene expression levels at the time t of the p genes; $F(\cdot, \cdot, \cdot)$ is a function which can be linear or nonlinear; and $\boldsymbol{\theta}$ is an unknown parameter vector which quantifies the regulations or interactions among the genes in GRN.

Once we can determine $\mathbf{X}(t)$, the gene expression levels which should be included in the ODE model (1), we can infer the interactions within a dynamic GRN. This motivates us to use appropriate models and to develop associated techniques in order to construct dynamic GRN for time course gene expression data. Within a dynamic GRN, the majority of the genes are not significantly relevant to each other. The precision of parameter estimation, model interpretability, and the accuracy of forecasting will be reduced when irrelevant genes are included in models [14]. Thus, those irrelevant genes should be excluded from the final model. However, variable selection for ODE models using traditional statistical methods is important but challenging, especially when it comes to dynamic GRN. The difficulties arise from two aspects: one is the collinearity among genes, i.e., genes sharing same “pathway” are highly correlated in expressions; the other is the high-dimensional feature of GRN, i.e., a large-scale GRN involves hundreds or even thousands of genes. When the number (n) of measurements for individual genes is much smaller than the number (p) of genes, traditional statistical methods face significant challenges in developing statistical procedures and deriving theory [15].

Pioneering research has investigated gene regulatory networks using variable selection techniques. For example, [13] proposed linear ODE models: $dX_k(t)/dt = \gamma^T\mathbf{X}(t)$ and developed a variable selection procedure based on SCAD penalty. [13] further employed their method to construct a module-based dynamic network. However a linear ODE model has many limitations and is unable to capture certain patterns. In reality, the first derivatives of the gene expression profiles (the time-related changes of a gene expression) can be quantified as a function of gene expression levels of all related genes. The link functions that quantify the regulatory effects of genes on the first derivatives may be nonlinear. In other words, systems of cellular regulations may be nonlinear [1, 16]. Due to the limitations of linear ODE models, developing a flexible modeling approach to explore the interactions among genes has become necessary. When the linear assumption cannot be satisfied, it is natural to consider a single-index model, $E(Y|X) = \eta(X^T\beta)$ with η being an *unknown* differentiable function and β an unknown parameter to be estimated. Single-index models have many advantages, such as being able to model the curvature of a smooth curve and circumventing the so-called “curse of dimensionality”. More discussions about the usefulness of single-index models are provided in [17]. A nonlinear ODE model (given the function η) may suffer from misspecification and “the curse of dimensionality”, whereas single-index ODE models can avoid these two problems and are more flexible, and the index parameter (β) can be estimated with the root—n convergence rate though the link function is unknown. More importantly, single index ODE models allow the predictors to have interactions, which is common in characterizing gene-gene regulation.

Various methods have been proposed to estimate regression coefficients for single-index models. See [18–23] for parameter estimators. In addition, much research has been done on variable selection for single-index models. For example, [24] developed a variable selection method based on sliced inverse regression. [25] proposed a leave-m-out cross-validation method to select variables in a single-index model. [26] proposed semiparametrically efficient profile least-squares estimators for parameter estimation, and employed the SCAD approach to simultaneously select variables and estimate regression coefficients. [27] studied estimation and variable selection coupling with dimension reduction procedures.

Although parameter estimation and variable selection for single-index models have gained fruitful results, to the best of our knowledge, no method that couples single-index models with ODE to study dynamic GRN is available. In this paper, we propose a single-index ODE model to study dynamic GRN with the aim of overcoming the inadequacy of linear ODE models. This model can be written as

$$\frac{dX_k(t)}{dt} = \eta_k\left(\mathbf{X}(t)^{\mathrm{T}}\boldsymbol{\beta}_0^{[k]}\right) + \varepsilon, k = 1, \ldots, p, \quad (2)$$

where $\eta_k(\cdot)$ is an unknown differentiable function; $\boldsymbol{\beta}_0^{[k]}$ is a parameter vector with $\left\|\boldsymbol{\beta}_0^{[k]}\right\| = 1$, and the first element of $\boldsymbol{\beta}_0^{[k]}$ is positive (for identifiability), where $\| \cdot \|$ denotes the Euclidean norm. $\mathbf{X}(t) = (x_1(t), \cdots, x_{\mathrm{p}}(t))^{\mathrm{T}}$ are state functions. Here $\mathbf{X}(t)$ can be gene-expressing levels of genes or population mean curves for functional modules. To study the interactions within dynamic GRN, one needs to identify the relevant $\mathbf{X}(t)$ for ODE models, that is $\boldsymbol{\beta}_0^{[k]} \neq 0$. We therefore apply the penalized least-squares approach for this aspect and for estimating dynamic parameters $\boldsymbol{\beta}_0^{[k]}$.

We will apply the mixed-effects nonparametric model with a mixture distribution framework to cluster the genes into functional modules in the first step. This clustering approach allows us to build the module-based dynamic network and identify the interesting functional modules. These interesting modules may play important roles in 'dynamic' regulations. Although these interesting modules may contain many genes with heterogeneous functions, it can allow scientists to focus on the genes in each module for further investigations. As shown in Fig 1 (below), most gene expression levels can be grouped in several clusters. In each cluster, these expression levels share a similar pattern. The genes in a cluster (represented by a node) may play a common function in biological procession. Such a network can single out regulator-regulator interactions which are helpful to avoid tedious experiments and to speed biological studies.

In Section of Methods, we briefly describe the procedure for GRN construction with details for penalized profile least-squares (PPrLS) estimation and variable selection. In Section of Numerical Results, we construct a module-based GRN structure by using PPrLS estimator for the yeast cell cycle gene expression data with additional results (S1 and S2 Tables), and conduct Monte Carlo simulation studies to evaluate the performance of the proposed procedure. The simulation settings were designed to mimic the gene expression patterns from the real data example. In Section of Discussions, we conclude the article with a brief discussion. All theory and associated technical details are given in the supporting materials (S1–S7 Files).

Methods

Time-course gene expressions are synchronized to many ongoing biological processes such as tissue repair, cell differentiation, or cell cycles [28, 29]. Through understanding the genes underlying the cell cycles, we can study the mechanisms of many diseases at a molecular level and in turn provide potential drug targets for treating those diseases. From the end of the last

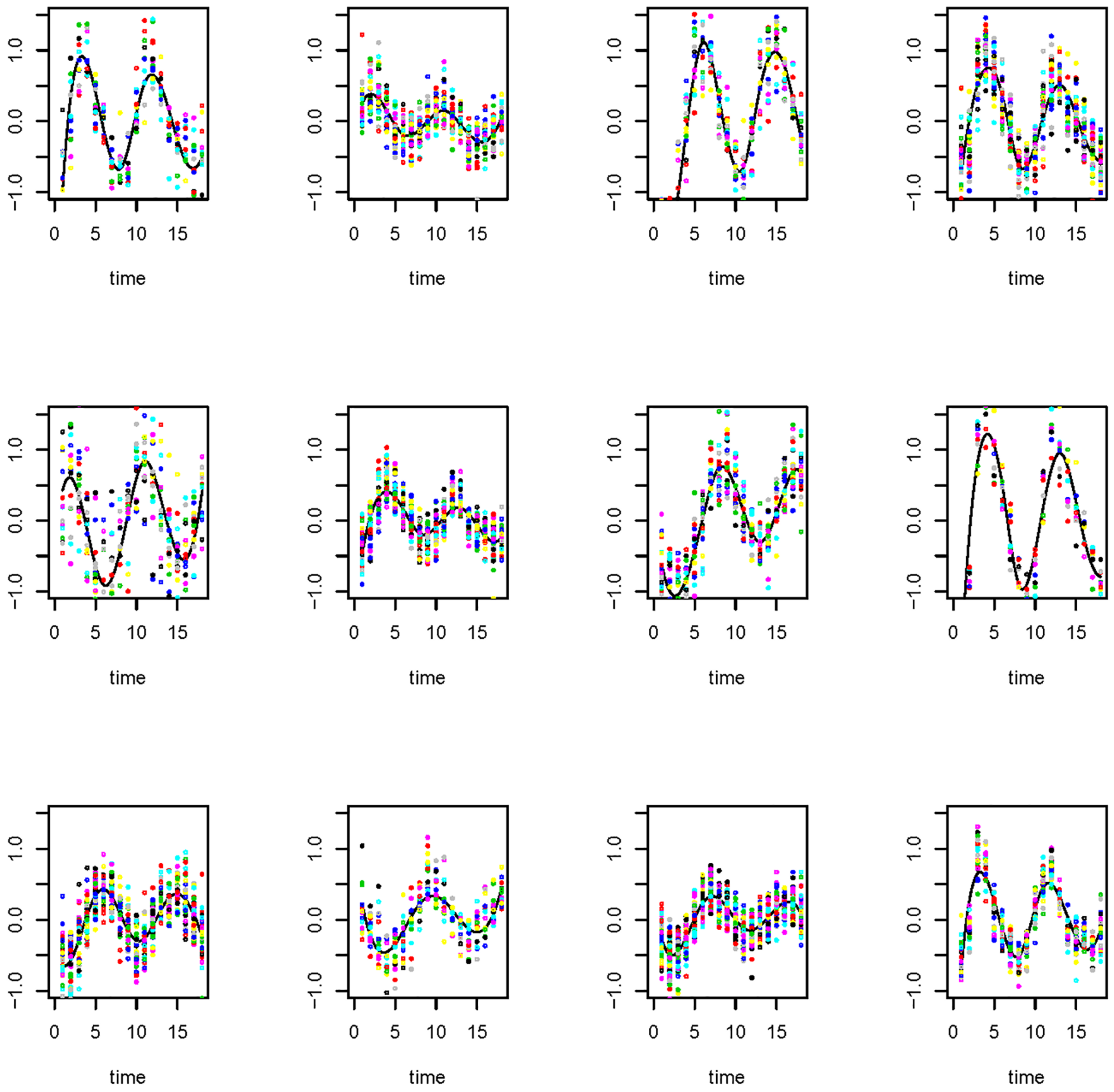

Fig 1. The scatterplot of gene expressions against time (the same color for each individual gene in a module) and the population mean curve (solid line) of 12 modules for the time course yeast cell data set.

https://doi.org/10.1371/journal.pone.0192833.g001

century, the identification of cell cycles associated genes has attracted considerable attention in biological study. For example, [28, 30] performed genome-wide transcriptional analysis of the cell cycle process of yeast using microarrays and identified about 800 cell-cycle-regulated genes. GRN include genes, the products of genes, and the interactions among them, which together affect many cellular processes. To understand the dynamic mechanism of cellular processes, modeling and analysis of dynamic gene regulatory networks using time-course gene expression data has attracted much attention. We model dynamic network for functional modules based on observed time course gene expression levels in three steps.

Step 1. Group genes into functional modules using the smoothing spline clustering (SSC) approach [31–33]. Final number of clusters was selected by the Bayesian information criterion (BIC), and penalty parameters were determined by the leave-one-out cross-validation procedure (GCV);

Step 2. Estimate state function $X(t)$ and first derivative $X'(t)$ for each functional module using nonparametric mixed-effect models (NPME) and the spline method respectively;

Step 3. Select modules and estimate dynamic parameters using the PPrLS procedure given below, for which tuning parameter was selected by the BIC, and bandwidths were determined by the GCV.

We now describe the details for these three steps.

Step 1—Clustering process

We assume the time-course gene expression levels for gene *i* can be represented by a smooth function of time and follow a mixture Gaussian distribution:

$$g_i(t) \sim p_1 N(\mu_1, \Sigma_1) + p_2 N(\mu_2, \Sigma_2) + \cdots + p_p N(\mu_p, \Sigma_p), \tag{3}$$

where p_k, $k = 1, \cdots, p$ are the probability that gene *i* belongs to cluster *k*; μ_k, $k = 1, \cdots, p$ and Σ_k, $k = 1, \cdots, p$ are the vector representations of the mean curve and variances components for each cluster respectively. The gene expression levels for individual genes are assumed to follow an overall mean curve (fixed-effect) while having a gene-specific shift (random-effect). Therefore nonparametric mixed-effect model can be constructed by fitting the time-course gene expressions for each gene to a function over time by using the smoothing spline method. Through maximizing the penalized log-likelihood, the SSC procedure estimates the probabilities p_k. The means μ_k and variances component Σ_k can be estimated as by-products also. More details about the SSC procedure are available in [32] and [33].

Step 2—Applications of nonparametric mixed-effect models

After grouping genes into functional modules, we apply NPME models to estimate the state function $X(t)$ and its first derivative $X'(t)$ for each functional module. For notation simplicity, we consider the estimation of the state function and its first derivative for module 1 ($k = 1$) and denote them by $X(t)$ and $X'(t)$ respectively. Suppose the number of genes in module 1 is m, and the number of measurements collected from each gene is m_i. The NPME model can be described as

$$g_i(t = X(t) + v_i(t) + \varepsilon_i(t), \; i = 1, \cdots, m, \tag{4}$$

where $g_i(t)$ is the observed gene expression level for the i^{th} gene; $X(t)$ presents the fixed-effect or population curve which reflects an overall time-related trend of the gene expression level for module 1; $v_i(t)$ describe individual curve variations; $\varepsilon_i(t)$ are measurement errors; and $v_i(t)$ and $\varepsilon_i(t)$ are assumed to be independent.

We can combine the penalized spline [34–36] with the linear mixed-effects (LME) modeling framework [37] to approximate $X(t)$. For presentation completeness, we briefly summarize the estimation procedure. We first approximate X(t) and $v_i(t)$ by $\widetilde{X}(t)$ and $\widetilde{v}_i(t)$, respectively, which are expressed as:

$$\widetilde{X}(t) = \sum_{r=0}^{l} \alpha_r t^r + \sum_{r=1}^{R} u_r (t - \zeta_r)_+^l, \text{ and } \widetilde{v}_i(t) = \sum_{r=0}^{l} b_{ir} t^r + \sum_{r=1}^{R} w_{ir} (t - \zeta_r)_+^l, \tag{5}$$

where $l \geq 1$ is an integer, $\zeta_1 < \cdots < \zeta_R$ are fixed knots, $u_r(t - \zeta_r)_+ = \max(0, t - \zeta_r)$, $\boldsymbol{\alpha} = (\alpha_1, \cdots, \alpha_l)$, $\mathbf{u} = (u_1, \cdots, u_R)$, $\mathbf{b}_i = (b_{i0}, \cdots, b_{il})$, and $\mathbf{w}_i = (w_{i0}, \cdots, w_{iR})$. Let

$$S_i = \begin{pmatrix} 1 & t_{i1} & \cdots & t_{i1}^l \\ 1 & t_{i2} & \cdots & t_{i2}^l \\ \vdots & \vdots & \ddots & \vdots \\ 1 & t_{im_i} & \cdots & t_{im_i}^l \end{pmatrix}, \text{ and } Z_i = \begin{pmatrix} (t_{i1} - \zeta_1)_+^l & \cdots & (t_{i1} - \zeta_R)_+^l \\ (t_{i2} - \zeta_1)_+^l & \cdots & (t_{i1} - \zeta_R)_+^l \\ \vdots & \ddots & \vdots \\ (t_{im_i} - \zeta_1)_+^l & \cdots & (t_{im_i} - \zeta_R)_+^l \end{pmatrix}.$$

The approximation of model (4) can be expressed as

$$\mathbf{g} = \mathbf{S}\boldsymbol{\alpha} + \Lambda\mathbf{b} + \mathbf{Z}\mathbf{u} + \boldsymbol{\Gamma}\mathbf{w} + \boldsymbol{\varepsilon}, \tag{6}$$

where $\mathbf{S} = (S_1^{\mathrm{T}}, \cdots, S_m^{\mathrm{T}})^{\mathrm{T}}$, $\mathbf{g} = (g_1^{\mathrm{T}}, \cdots, g_m^{\mathrm{T}})^{\mathrm{T}}$, $\Lambda = \mathrm{diag}(S_1^{\mathrm{T}}, \cdots, S_m^{\mathrm{T}})$, $\mathbf{Z} = (Z_1^{\mathrm{T}}, \cdots, Z_m^{\mathrm{T}})^{\mathrm{T}}$, $\boldsymbol{\Gamma} = \mathrm{diag}(Z_1^{\mathrm{T}}, \cdots, Z_m^{\mathrm{T}})$, $\mathbf{b} = (b_1^{\mathrm{T}}, \cdots, b_m^{\mathrm{T}})^{\mathrm{T}}$, and $\mathbf{w} = (w_1^{\mathrm{T}}, \cdots, w_m^{\mathrm{T}})^{\mathrm{T}}$. Model (6) is a standard LME model. As a result, $\boldsymbol{\alpha}$, $\mathbf{b}$, $\mathbf{u}$ and $\mathbf{w}$ can be estimated by using the function `lme` (available in the R package `nlme`). Substituting the estimated $\widehat{\boldsymbol{\alpha}}$ and $\widehat{\mathbf{u}}$ in Eq (5), we estimate the $X(t)$ for module 1. After estimating $\mathbf{X}(t)$, we apply the spline method (available in the R package `splines`) to estimate the first derivative of $\widehat{X}(t)$. The detailed estimation procedure is referred to [38] and [39].

Step 3—Estimation procedure based on the penalized profile least-squares approach

Suppose a genome-wide time course gene expression levels were clustered into p modules; $X_j(t)$, $j = 1, \cdots, p$ are the population mean curves estimated by NPME models; and $\widehat{X}'_k(t)$ are the estimates of the first derivative $dX_k(t)/dt$ for the k-th module. Substituting $X_j(t)$, $j = 1, \cdots, p$ and the first derivative $\widehat{X}'_k(t)$ for the k-th module in model (2), we obtain a single-index ODE model for the k-th functional module which can be written as

$$Y_k(t) = \eta_k(\mathbf{X}(t)^{\mathrm{T}}\boldsymbol{\beta}_0^{[k]}) + \boldsymbol{\varepsilon}, \ k = 1, \cdots, p, \tag{7}$$

where η_k is an unknown differentiable function, $Y_k(t) = \widehat{X}'_k(t)$, $\mathbf{X}(t) = (X_1(t), \ldots, X_\mathrm{p}(t))^{\mathrm{T}}$, $\boldsymbol{\beta}_0^{[k]} = (\beta_{01}^{[k]}, \cdots, \beta_{0p}^{[k]})^{\mathrm{T}}$, and ε is the sum of numerical errors due to integration and estimation. This complexity $\boldsymbol{\varepsilon}$ makes it challenging to study the properties of the proposed estimator for $\boldsymbol{\beta}'s$, For simplicity, we adopt an additive error model used in the literature [13, 40, 41]. Once the $\mathbf{X}(t)$ can be identified, we construct a module-based network. Here we develop the variable (population mean of functional modules) selection and estimation procedure for model (7) based on the penalized profile least-squares approach as follows.

Selecting variables by penalized least squares has been widely studied in literature. See, for example, the least absolute shrinkage and selection operator (LASSO) [42], the smoothly clipped absolute deviation (SCAD) approach [43], the adaptive lasso estimator [44], the elastic-net estimator [45] and the adaptive elastic-net estimator [46]. However, the variable selection problem for single-index ODE models has not been addressed in the literature. In this paper, we extend the approach proposed by [26] to the single-index ODE model (7).

Let p be the number of all modules; $X_i = (X_1(t_i), \ldots, X_p(t_i))^{\mathrm{T}}$, $i = 1, \ldots, N$, $Y_i = \widehat{X}'_k(t_i)$ be the vector representations of the mean curves of p function modules and the estimates of the first derivative $dX_k(t)/dt$ for the k-th module. Assume the functional data of the kth module follows the single-index ODE model

$$Y_i = \eta_k(X_i^{\mathrm{T}}\boldsymbol{\beta}^{[k]}) + \varepsilon_i,\ \ k = 1, \cdots, p, \tag{8}$$

Let $\Lambda_i = X_i^{\mathrm{T}}\boldsymbol{\beta}^{[k]}$. $\eta_k(u)$ can be estimated utilizing the local linear regression method [47], i.e., minimizing

$$\sum_{i=1}^{N}\{a_k + b_k(\Lambda_i - u) - Y_i\}^2 K_h(\Lambda_i - u), \tag{9}$$

with respect to a_k and b_k, where $K_h(\cdot) = K(\cdot/h)/h$, $K(\cdot)$ is a kernel function and h is a bandwidth. We can then obtain

$$\hat{\eta}_k(u, \boldsymbol{\beta}) = \hat{a}_k = \frac{K_{20}(u,\boldsymbol{\beta})K_{01}(u,\boldsymbol{\beta}) - K_{10}(u,\boldsymbol{\beta})K_{11}(u,\boldsymbol{\beta})}{K_{00}(u,\boldsymbol{\beta})K_{20}(u,\boldsymbol{\beta}) - K_{10}^2(u,\boldsymbol{\beta})}, \tag{10}$$

where $K_{jl}(u, \boldsymbol{\beta}) = \sum_{i=1}^{N} K_h(X_i^{\mathrm{T}}\boldsymbol{\beta}^{[k]} - u)(X_i^{\mathrm{T}}\boldsymbol{\beta}^{[k]} - u)^j Y_i^l$, for $j = 0, 1, 2$ and $l = 0, 1, 2$. Consequently, the profile least squares function can be proposed as a function of $\beta^{[k]}$

$$Q(\boldsymbol{\beta}^{[k]}) = \sum_{i=1}^{N}\{Y_i - \hat{\eta}_k(X_i^{\mathrm{T}}\boldsymbol{\beta}^{[k]})\}^2. \tag{11}$$

The above estimation procedure can be used when the true model is known a priori. Because we wish to identify GRN structure and enhance the predictive power of a proposed model, we apply the penalized least-squares approach to simultaneously select modules and estimate parameters. Define a penalized profile least-squares (PPrLS) function

$$\mathcal{L}_P\left(\boldsymbol{\beta}^{[k]}\right) = \frac{1}{2}Q\left(\boldsymbol{\beta}^{[k]}\right) + N\sum_{j=1}^{p} p_{\lambda^{[k]}}\left(|\beta_j^{[k]}|\right), \tag{12}$$

where $p_{\lambda^{[k]}}(\cdot)$ is a penalty function with a regularization parameter $\lambda^{[k]}$. The PPrLS estimator of $\boldsymbol{\beta}^{[k]}$ is the minimizer of Eq (12); i.e.,

$$\widehat{\boldsymbol{\beta}}^{[k]} = \operatorname{argmin}\mathcal{L}_P(\boldsymbol{\beta}^{[k]}). \tag{13}$$

For a given tuning parameter $\lambda^{[k]}$, we can estimate $\boldsymbol{\beta}^{[k]}$ by minimizing $\mathcal{L}_P(\boldsymbol{\beta}^{[k]})$ with respect to $\boldsymbol{\beta}^{[k]}$. By determining non-zero $\boldsymbol{\beta}^{[k]}$, we identify the modules having impacts on the kth module and therefore construct GRN.

There are various penalty functions in the literature of variable selection for semiparametric models. Considering the SCAD method has many good theoretical properties, we adopt the SCAD penalty function [43], and adopt BIC selector proposed by [48] to choose the regularization parameters $\lambda^{[k]}$ by minimizing the following objective function:

$$\mathrm{BIC}(\lambda^{[k]}) = \log\{\mathrm{MSE}(\lambda^{[k]})\} + \{\log(N)/N\}\mathrm{DF}_{\lambda^{[k]}}, \tag{14}$$

where $\mathrm{MSE}(\lambda^{[k]}) = N^{-1}\sum_{i=1}^{N}\left\{Y_i - \hat{\eta}_k\left(X_i^{\mathrm{T}}\widehat{\boldsymbol{\beta}}_{\lambda^{[k]}}^{[k]}\right)\right\}^2$ and $\mathrm{DF}_{\lambda^{[k]}}$ is the number of nonzero coefficients of $\widehat{\boldsymbol{\beta}}_{\lambda^{[k]}}^{[k]}$, the PPrLS obtained from (12) for each $\lambda^{[k]}$.

Remark. Although the proposed method needs three steps to implement and its computational cost is high, compared to the existing methods, its gain in computational efficiency is significant. Most of dynamic network models such as dynamic Bayesian networks and random graph models require extensive computations for posterior inference. As a result, Bayesian based methods allow one to deal with only small networks. The proposed method can avoid numerically solving the differential equations directly, and does not need the initial or boundary conditions of the state variables. The method also incorporate the high-dimensional ODEs to allow us to perform variable selection and parameter estimation for one equation. These good features gain computational efficiency.

Numerical results

Real data analysis

We used the procedure introduced in Section of Methods to analyze a time-course yeast cell cycle gene expression data set. These 297 genes were identified as expressions across 18 time points during approximate two cell cycles; i.e., each gene has 18 time-related observations [49].

We implemented Step 1 using the MFDA function (available in the R package `MFDA`), and identified 12 functional modules. The population mean curves for the functional modules are given in Fig 1. We can see that for each functional module, the genes included share a similar pattern. These time-related patterns show two cell cycles (Fig 1). The number of genes included in each functional module ranges from 9 to 53.

In order to construct a functional landscape of the genome-wide regulatory network through identifying interactions among modules, we used the Database for Annotation, Visualization and Integrated Discovery [50, 51] to identify enriched functional annotations in Gene ontology and Kyoto Encyclopedia of Genes and Genomes pathways for each functional module. A modified Fisher exact test was used to test the null hypothesis that a certain function is not over-represented in the module compared to the background population. Due to space limitation, we displayed part of the selected functional annotations in Table 1. All enriched functional annotations were provided in S1 Table.

As shown in Table 1, the function annotation analysis suggested that genes in the identified functional modules participate in broad biological process such as cell cycle, DNA replication or packaging, meiosis, regulation of transcription etc. For example, module 3 was highly enriched in DNA packaging; module 7 was enriched in cell-division cycle and mitosis; and DNA metabolic process was related to module 12. Although each functional module has multiple enriched annotations, but most annotations can be grouped into one or two clusters.

After grouping genes into functional modules, we applied step 2 to all functional modules, and obtained $X_i(t)$ and $\widehat{X}_i'(t),\ \ i = 1, \cdots, 12$. Following the data augmentation strategy used in [13, 52, 53] and [54], we selected 300 time points from $X_i(t)$ and the first derivative $\widehat{X}_1'(t)$ for the module. Therefore, the sample size is $N = 300$. After substituting the estimates into single-index models, we built the full model for module 1, for instance, as follows.

$$y_1 = \eta_1(\mathbf{X}(t)^{\mathrm{T}}\boldsymbol{\beta}_0^{[1]}) + \varepsilon, \tag{15}$$

where the response variable $y_1 = \widehat{X}_1'(t)$, the estimated first derivatives; $\mathbf{X}(t) = (X_1(t), \ldots, X_{12}(t))^{\mathrm{T}}$ are the population mean estimates of 12 functional modules; and $\boldsymbol{\beta}_0^{[1]} = (\beta_{01}^{[1]}, \ldots, \beta_{012}^{[1]})^{\mathrm{T}}$. Applying the PPrLS procedure given in step 3 to model (15), we detected significant variables $\mathbf{X}(t)$ and obtained nonzero $\widehat{\boldsymbol{\beta}}^{[1]}$. As a result, we identified the modules related to the gene-expression changes of the module 1. For a comparison, we also fitted y to $\mathbf{X}(t)$ by using a linear ODE model [13]

$$y_1 = \mathbf{X}(t)^{\mathrm{T}}\boldsymbol{\beta}_{L0}^{[1]} + \varepsilon. \tag{16}$$

Table 1. The inward and outward regulations in the module-based regulatory network and RSS based on the linear ODE (L-ODE) and the single-index ODE (Si-ODE).

Module	Selected function annotation and associated p-values (in parentheses)	Outward influence modules		Inward influence modules		RSS	
		L-ODE	Si-ODE	L-ODE	Si-ODE	L-ODE	Si-ODE
module1 (12)	DNA replication (0.013), regulation of RNA metabolic process (0.014), meiosis (0.021)	3, 5, 7, 8	1, 3, 7, 8, 9, 12	2, 6, 9, 12	1, 2, 3, 4, 5, 8, 9, 10, 11	3.07E-03	1.41E-04
module2 (30)	cellular carbohydrate biosynthetic process (0.006),	1, 7, 9	1, 2, 7, 8, 9, 12	7	2, 4, 5, 8, 9, 10, 12	5.20E-04	3.38E-05
module3 (15)	protein—DNA complex assembly (< 0.001), DNA packaging (< 0.001)	3, 7, 8, 10	1, 7, 8, 12	1, 3, 4, 5, 6, 7, 8, 12	1, 4, 5, 6, 7, 8, 10, 11, 12	3.37E-03	3.58E-03
module4 (32)	DNA metabolic process (< 0.001), DNA replication (< 0.001), DNA repair (< 0.001), cell-division cycle (0.025)	3, 5, 7, 8	1, 2, 3, 7, 8, 9, 11, 12	NA	6, 8, 9, 11, 12	4.52E-03	1.05E-03
module5 (16)	interphase of mitotic cell cycle (< 0.001), DNA replication initiation (< 0.001)	3, 5, 7, 8	1, 2, 3, 5, 8, 9, 11, 12	1, 4, 5, 6, 7, 8, 9, 10, 11, 12	5, 7, 10, 11	5.28E-03	3.45E-04
module6 (38)	lipoprotein biosynthetic process and metabolic process (0.004), regulation of DNA metabolic process (0.005), chromosome organization (< 0.001)	1, 3, 5, 7	3, 4, 6, 7, 8, 9, 12	NA	6, 9	1.10E-03	2.88E-05
module7 (20)	nuclear division (< 0.001), cell-division (< 0.001), mitosis (< 0.001)	2, 3, 5, 8	3, 5, 7, 8, 9, 10, 12	1, 2, 3, 4, 5, 6, 8, 9, 10, 11	1, 2, 3, 4, 6, 7, 8, 10	2.10E-03	2.67E-04
module8 (9)	cell cycle (0.007), regulation of cell cycle (0.025)	3, 5, 7, 8	1, 2, 3, 4, 7, 8, 9, 11, 12	1, 3, 4, 5, 7, 8, 9, 10, 11	1, 2, 3, 4, 5, 6, 7, 8, 9, 11	1.01E-02	7.90E-04
module9 (35)	Glycosylation (< 0.001), mitotic cell cycle (< 0.001), nuclear division (< 0.001)	1, 5, 7, 8	1, 2, 4, 6, 8, 11	2	1, 2, 4, 5, 6, 7, 8, 11	5.20E-04	1.34E-05
module10 (14)	regulation of cell cycle (< 0.001), regulation of cell cycle process (0.001)	5, 7, 8, 12	1, 2, 3, 5, 7, 11, 12	3, 11	7, 11	6.55E-03	6.82 E-06
module11 (53)	cell cycle (0.043), nuclear migration along microtubule (0.012)	5, 7, 8, 10	1, 3, 4, 5, 8, 9, 10	NA	4, 5, 8, 9, 10	8.03E-05	8.70E-06
module12 (23)	mitotic recombination (< 0.001), DNA metabolic process (< 0.001)	1, 3, 5	2, 3, 4	10	1, 2, 3, 4, 5, 6, 7, 8, 10	2.27E-03	1.41E-04

https://doi.org/10.1371/journal.pone.0192833.t001

We also selected $\mathbf{X}(t)$ and estimated $\boldsymbol{\beta}_{L0}^{[1]} = (\beta_{L01}^{[1]}, \ldots, \beta_{L012}^{[1]})^{\mathrm{T}}$ by applying the SCAD method to the linear ODE model (16).

Applying the procedure to all functional modules, we constructed a regulatory network among modules (Figs 2 and 3) and estimated their corresponding dynamic coefficients by both single-index and linear ODE models.

To compare the results provided by the single-index ODE and linear ODE models, we summarized the inward (significantly impact on) and outward (impacted by) regulatory relationships between modules in Table 1. The number of genes in each module was displayed in the parentheses. One can see that the residuals of sum squares (RSS) of single-index ODE models were smaller than those of the linear ODE models. We can also observe that the single-index ODE models selected more modules than the linear ODE models did. For example, the single-index ODE model indicated that module 2 was impacted by modules 1, 2, 7, 8, 9 and 12 of which only modules 1, 7 and 9 were selected by the linear ODE model. Both linear ODE and single-index ODE indicated that modules 3, 7 and 8 were important because they regulated more than 50% modules. We also noted that module 8 only included 9 genes. Further experiments are needed to explore these new discoveries in biological progression.

A simulation study

In this part we conducted Monte Carlo simulation studies to validate the proposed procedure for the single-index ODE models. Due to the intensive computational cost, we designed a system with 7 ODEs, which include following linear and nonlinear forms. The simulation settings

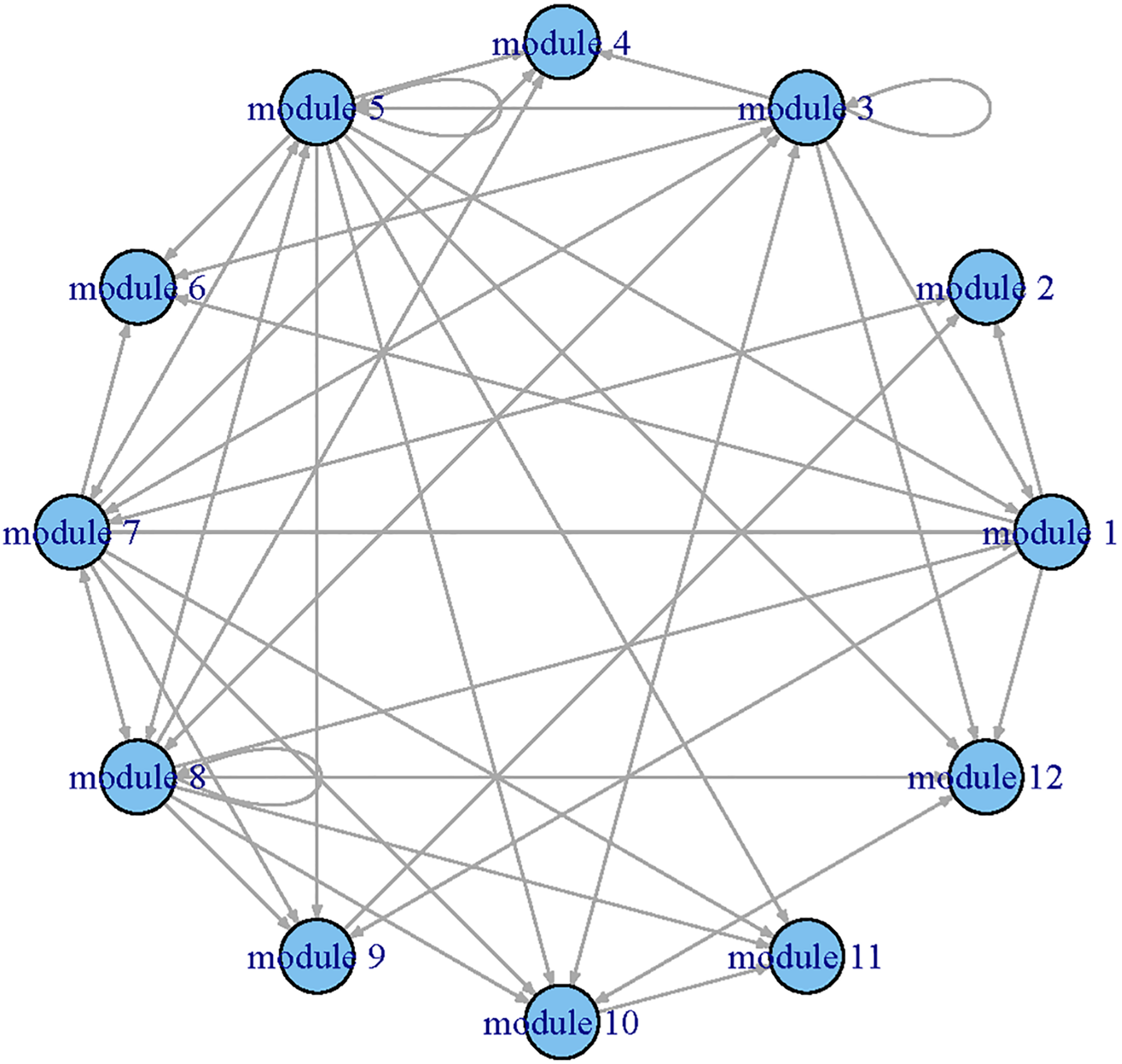

Fig 2. The GRN identified by the linear ODE models for the time course yeast cell data set. Each node represents a module and the arrows presents the direction of influence.

https://doi.org/10.1371/journal.pone.0192833.g002

are data-driven because the gene expression pattern show sine and cosines patterns (Fig 1)

$$\begin{aligned}\frac{dX_1(t)}{dt} &= 0.05*(\beta_{01}*X_1+\beta_{02}*X_2),\ \frac{dX_2(t)}{dt}=\cos(\beta_{03}*X_2+\beta_{04}*X_3),\\ \frac{dX_3(t)}{dt} &= \sin(\beta_{05}*X_2+\beta_{06}*X_3),\ \frac{dX_4(t)}{dt}=0.1*(\beta_{07}*X_2+\beta_{08}*X_4),\\ \frac{dX_5(t)}{dt} &= \sin(\beta_{09}*X_2+\beta_{010}*X_4),\ \frac{dX_6(t)}{dt}=0.05*\exp(\beta_{011}*X_3+\beta_{012}*X_6),\\ \frac{dX_7(t)}{dt} &= 0.2*(\beta_{013}*X_2+\beta_{014}*X_3),\ X_p(t_0)=X_{p0},\ p=1,\cdots,7,\end{aligned}\tag{17}$$

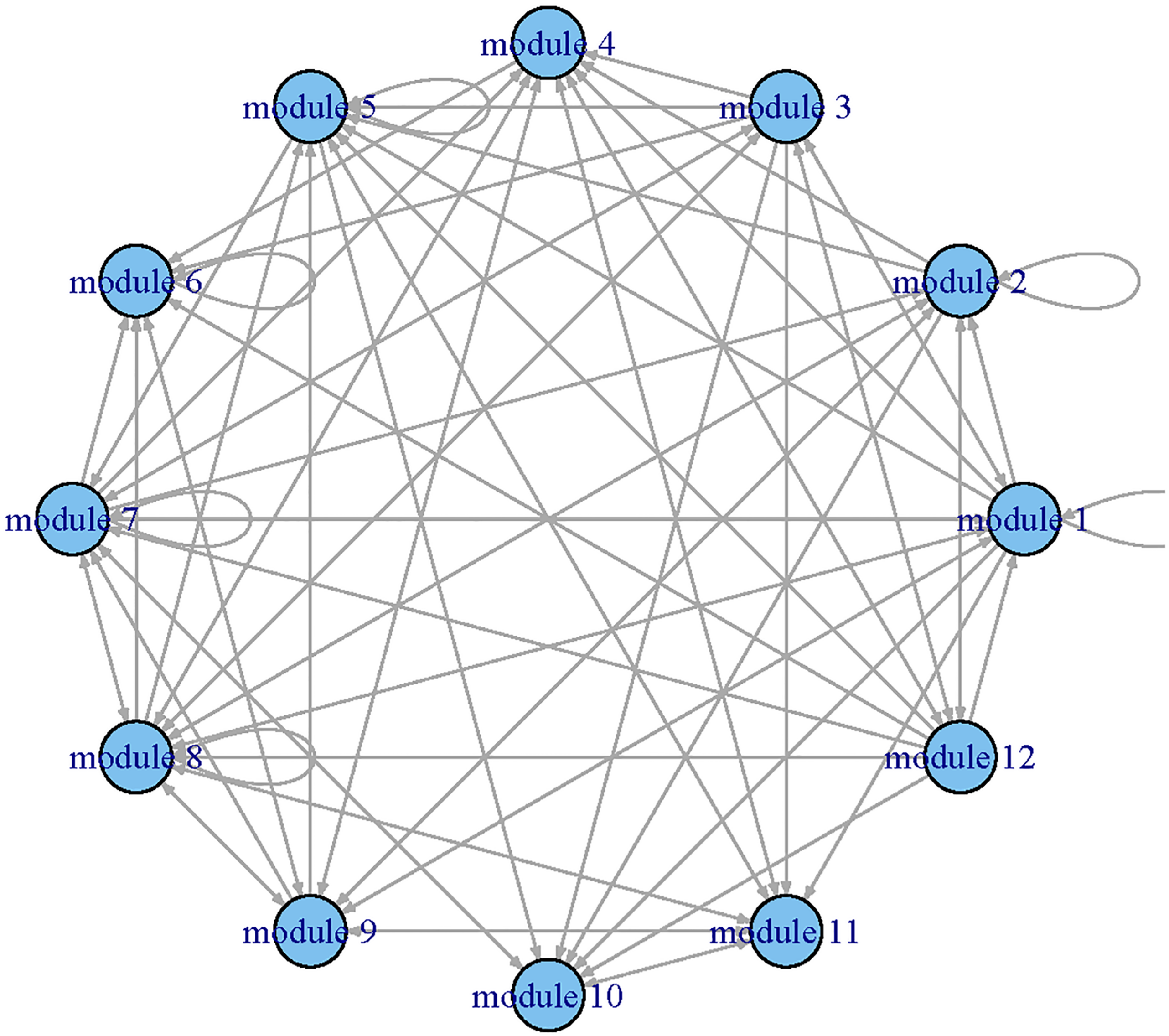

Fig 3. The GRN identified by the single-index ODE models for the time course yeast cell data set. Each node represents a module and the arrows presents the direction of influence.

https://doi.org/10.1371/journal.pone.0192833.g003

where $\boldsymbol{\beta}_0^{[1]} = (\beta_{01}, \beta_{02}, 0, 0, 0, 0, 0)^{\mathrm{T}} = (0.707, 0.707, 0, 0, 0, 0, 0)^{\mathrm{T}}$,
$\boldsymbol{\beta}_0^{[2]} = (0, \beta_{03}, \beta_{04}, 0, 0, 0, 0)^{\mathrm{T}} = (0, 0.555, -0.832, 0, 0, 0, 0)^{\mathrm{T}}$,
$\boldsymbol{\beta}_0^{[3]} = (0, \beta_{05}, \beta_{06}, 0, 0, 0, 0)^{\mathrm{T}} = (0, 0.832, -0.555, 0, 0, 0, 0)^{\mathrm{T}}$,
$\boldsymbol{\beta}_0^{[4]} = (0, \beta_{07}, 0, \beta_{08}, 0, 0, 0)^{\mathrm{T}} = (0, 0.600, 0, -0.800, 0, 0, 0)^{\mathrm{T}}$,
$\boldsymbol{\beta}_0^{[5]} = (0, \beta_{09}, 0, \beta_{010}, 0, 0, 0)^{\mathrm{T}} = (0, 0.894, 0, 0.447, 0, 0, 0)^{\mathrm{T}}$,
$\boldsymbol{\beta}_0^{[6]} = (0, 0, \beta_{011}, 0, 0, \beta_{012}, 0)^{\mathrm{T}} = (0, 0, 0.894, 0, 0, -0.447, 0)^{\mathrm{T}}$, and
$\boldsymbol{\beta}_0^{[7]} = (0, \beta_{013}, \beta_{014}, 0, 0, 0, 0)^{\mathrm{T}} = (0, 0.894, -0.447, 0, 0, 0, 0)^{\mathrm{T}}$.

Given initial values X_{p0}, $p = 1, \ldots, 7$, we can numerically solve the above ODE system and obtain the numerical solution $X_p(t)$, $p = 1, \ldots, 7$. In this simulation study, we first generated initial values $X_p(0)$, $p = 1, \ldots, 7$ by using

$$X_{p0} = X_0 + 0.5 * e_p,\ \ p = 1, \ldots, 7,$$

where e_p follows $N(0, 1)$ and $X_0 = (0.7628, 0.6789, 1.2351, 0.6170, 2.7800, 0.2906, 0.4441)$. We then numerically solved the ODE system (17) and output $X_p(t)$, $p = 1, \ldots, 7$, using three different schedules: equally spaced time points on the ranges of [0, 18], but three different intervals between time points. As a result, we simulated seven population means $X_p(t_i)$, $p = 1, \ldots, 7$, $i = 1, \ldots, N$ with sample sizes $N = 180, 288, 360$. After generating the population mean curves, we used the spline method to estimate the first derivatives, which are denoted by $\widehat{X}'_p(t)$, $p = 1, \ldots, 7$. For notation simplicity, we gave the model structure and estimation procedure for the first ODE ($k = 1$). The same procedure can be applied to the rest of the ODEs. Substituting the generated $X_p(t)$, $p = 1, \ldots, 7$ and estimated $\widehat{X}'_1(t)$ into the single-index ODE models, we obtained the largest model for the first ODE as follows:

$$\widehat{X}'_1(t_i) = \eta_1(\mathbf{X}(t_i)^{\mathrm{T}}\boldsymbol{\beta}_0^{[1]}) + \varepsilon_i,\ \ i = 1, \ldots, N,$$

where $\boldsymbol{\beta}_0^{[1]} = (\beta_{01}^{[1]}, \ldots, \beta_{07}^{[1]})^{\mathrm{T}}$ and $\mathbf{X}(t_i) = (X_1(t_i), \cdots, X_7(t_i))^{\mathrm{T}}$, $i = 1, \cdots, N$. Applying the procedure given in Step 3, we selected $\mathbf{X}(t)$ and estimated $\boldsymbol{\beta}_0^{[1]}$ for the first ODE. Applying the same procedure to the other six ODEs, we estimated $\boldsymbol{\beta}_0^{[k]}$, $k = 2, \ldots, 7$. As a result, we constructed GRN for the simulated functional modules. We repeated the same procedure 100 time and summarized the $\mathrm{MSE}_q = \sum_{j=1}^{100}(\widehat{\beta}_{qj} - \beta_{0q})^2/100$ and $\mathrm{ARE}_q = \sum_{j=1}^{100}\dfrac{|\widehat{\beta}_{qj} - \beta_{0q}|}{|\beta_{0q}|}$, $q = 1, \ldots, 14$, where $\widehat{\beta}_{qj}$ is the estimated β_q for j_{th} iteration. In Table 2, "overfitted (O)" represents extra variables; "underfitted (U)" represents incorrectly deleting necessary variables. We can see that the PPrLS method can correctly select the variables for most cases in terms of the number of correctly fitted model. Larger sample sizes lead to better performance. For ODEs with a linear form, namely ODE1, ODE4, and ODE7, both variable selection and parameter estimation procedures have good performance when the sample size is 180. For the nonlinear case, with the increase of the sample size, both variable selection and parameter estimation tend to work better. In addition, we reported the 10% trimmed MSE and ARE (discarding 5% of the lowest and the highest values). Meanwhile, we constructed networks among simulated functional modules for each iteration (see Figs 4, 5 and 6). The thick lines represents true connection, and the numbers present the times which were found by our method in 100 iterations. From Figs 4, 5 and 6, we can see that the constructed GRN match the true network in most cases.

Conclusions and discussions

In this paper, we have proposed single-index ODE models and developed a procedure to select variables and estimate parameters. The procedure has further been used to analyze a time-course data set with the aim of exploring the module based and regulator-regulator interactions. We found the interactions identified by using single-index ODE were more accurate, i.e., the linear ODE models overlooked some confirmed regulator-regulator interactions [55]. We took module 12 as an example. MBP1 is a DNA-binding protein that forms MBF complex; a protein complex that binds to the Mlu1 cell cycle box promoter element. [56, 57] showed that MBP1 is topologically related to transcription factors, including SWI4 in *Saccharomyces cerevisiae*. In addition, there is physical and genetic evidence that MBP1 interacts with SKN7, a transcription factor [58]. These two interactions are identified as potential interactions in module 12 by single-index ODE models, but are overlooked by the linear ODE models.

Table 2. The simulation results for the SCAD method for scenarios with different sample sizes based on 100 replications. The simulation results for the SCAD method for scenarios with different sample sizes based on 100 replications. Correctly fitted (C); underfitted (U); overfitted(O).

ODE	$\widehat{\beta}$'s	C	U	O	MSE	MSE$_{trim}$	ARE(%)	ARE$_{trim}$(%)
	$N = 180$							
1	$\widehat{\beta}_1$	100	0	0	< 0.001	< 0.001	0.005	0.004
	$\widehat{\beta}_2$				< 0.001	< 0.001	0.005	0.004
2	$\widehat{\beta}_3$	96	0	3	0.012	< 0.001	4.012	0.013
	$\widehat{\beta}_4$				0.028	< 0.001	4.005	0.006
3	$\widehat{\beta}_5$	97	1	2	0.014	< 0.001	2.271	0.006
	$\widehat{\beta}_6$				0.007	< 0.001	2.366	0.013
4	$\widehat{\beta}_7$	99	0	1	0.004	< 0.001	1.025	0.024
	$\widehat{\beta}_8$				0.006	< 0.001	1.014	0.013
5	$\widehat{\beta}_9$	65	6	24	0.295	0.255	35.181	32.363
	$\widehat{\beta}_{10}$				0.064	0.057	32.669	30.219
6	$\widehat{\beta}_{11}$	93	2	3	0.053	0.005	6.443	1.105
	$\widehat{\beta}_{12}$				0.014	0.003	6.811	1.538
7	$\widehat{\beta}_{13}$	100	0	0	< 0.001	< 0.001	0.004	0.003
	$\widehat{\beta}_{14}$				< 0.001	< 0.001	0.016	0.014
	$N = 288$							
1	$\widehat{\beta}_1$	100	0	0	< 0.001	< 0.001	0.002	0.001
	$\widehat{\beta}_2$				< 0.001	< 0.001	0.002	0.001
2	$\widehat{\beta}_3$	96	0	4	0.012	< 0.001	4.003	0.003
	$\widehat{\beta}_4$				0.028	< 0.001	4.001	0.001
3	$\widehat{\beta}_5$	97	1	2	0.014	< 0.001	2.264	0.002
	$\widehat{\beta}_6$				0.007	< 0.001	2.49	0.003
4	$\widehat{\beta}_7$	100	0	0	< 0.001	< 0.001	0.007	0.007
	$\widehat{\beta}_8$				< 0.001	< 0.001	0.004	0.004
5	$\widehat{\beta}_9$	77	5	13	0.166	0.14	21.242	18.046
	$\widehat{\beta}_{10}$				0.055	0.035	23.601	18.503
6	$\widehat{\beta}_{11}$	96	1	2	0.024	< 0.001	3.167	0.001
	$\widehat{\beta}_{12}$				0.006	< 0.001	3.19	0.003
7	$\widehat{\beta}_{13}$	100	0	0	< 0.001	< 0.001	0.001	0.001
	$\widehat{\beta}_{14}$				< 0.001	< 0.001	0.004	0.004
	$N = 360$							
1	$\widehat{\beta}_1$	100	0	0	< 0.001	< 0.001	0.001	0.001
	$\widehat{\beta}_2$				< 0.001	< 0.001	0.001	0.001
2	$\widehat{\beta}_3$	96	0	4	0.012	< 0.001	4.002	0.002
	$\widehat{\beta}_4$				0.028	< 0.001	4.001	0.001
3	$\widehat{\beta}_5$	97	1	2	0.014	< 0.001	2.278	0.001
	$\widehat{\beta}_6$				0.007	< 0.001	2.402	0.002
4	$\widehat{\beta}_7$	100	0	0	< 0.001	< 0.001	0.005	0.005
	$\widehat{\beta}_8$				< 0.001	< 0.001	0.003	0.003

(Continued)

Table 2. (Continued)

ODE	$\widehat{\beta}$'s	C	U	O	MSE	MSE$_{trim}$	ARE(%)	ARE$_{trim}$(%)
5	$\widehat{\beta}_9$	78	5	10	0.15	0.122	19.267	15.852
	$\widehat{\beta}_{10}$				0.039	0.032	20.493	17.214
6	$\widehat{\beta}_{11}$	96	0	4	0.032	< 0.001	4	< 0.001
	$\widehat{\beta}_{12}$				0.008	< 0.001	4.002	0.002
7	$\widehat{\beta}_{13}$	100	0	0	< 0.001	< 0.001	0.001	< 0.001
	$\widehat{\beta}_{14}$				< 0.001	< 0.001	0.002	0.002

https://doi.org/10.1371/journal.pone.0192833.t002

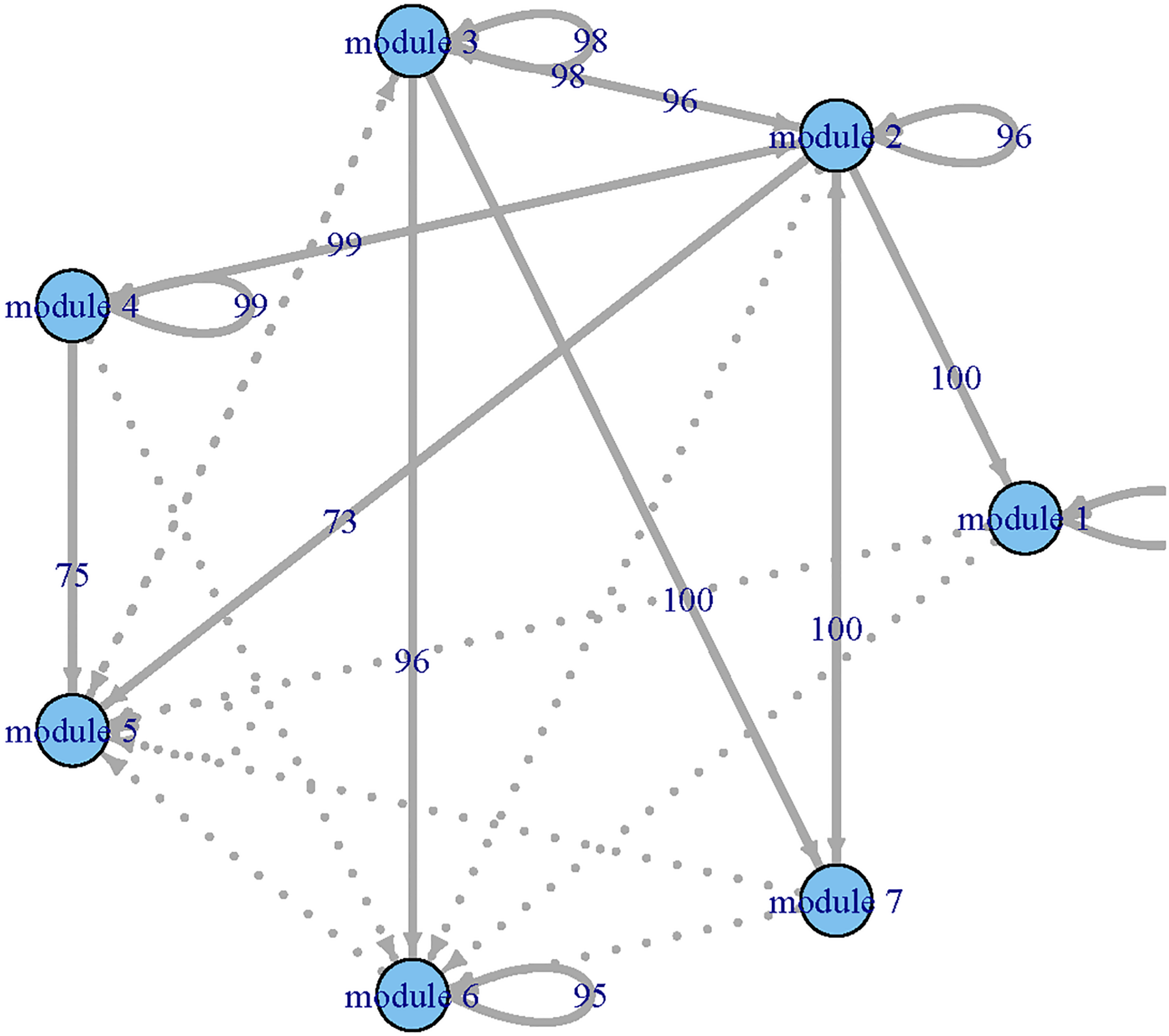

Fig 4. The constructed gene regulatory networks for simulation studies with $N = 180$ and 100 iterations. Solid lines: the true connections, numbers present: the times correctly identified using our procedure in 100 iteration, dots line: incorrectly identified connections.

https://doi.org/10.1371/journal.pone.0192833.g004

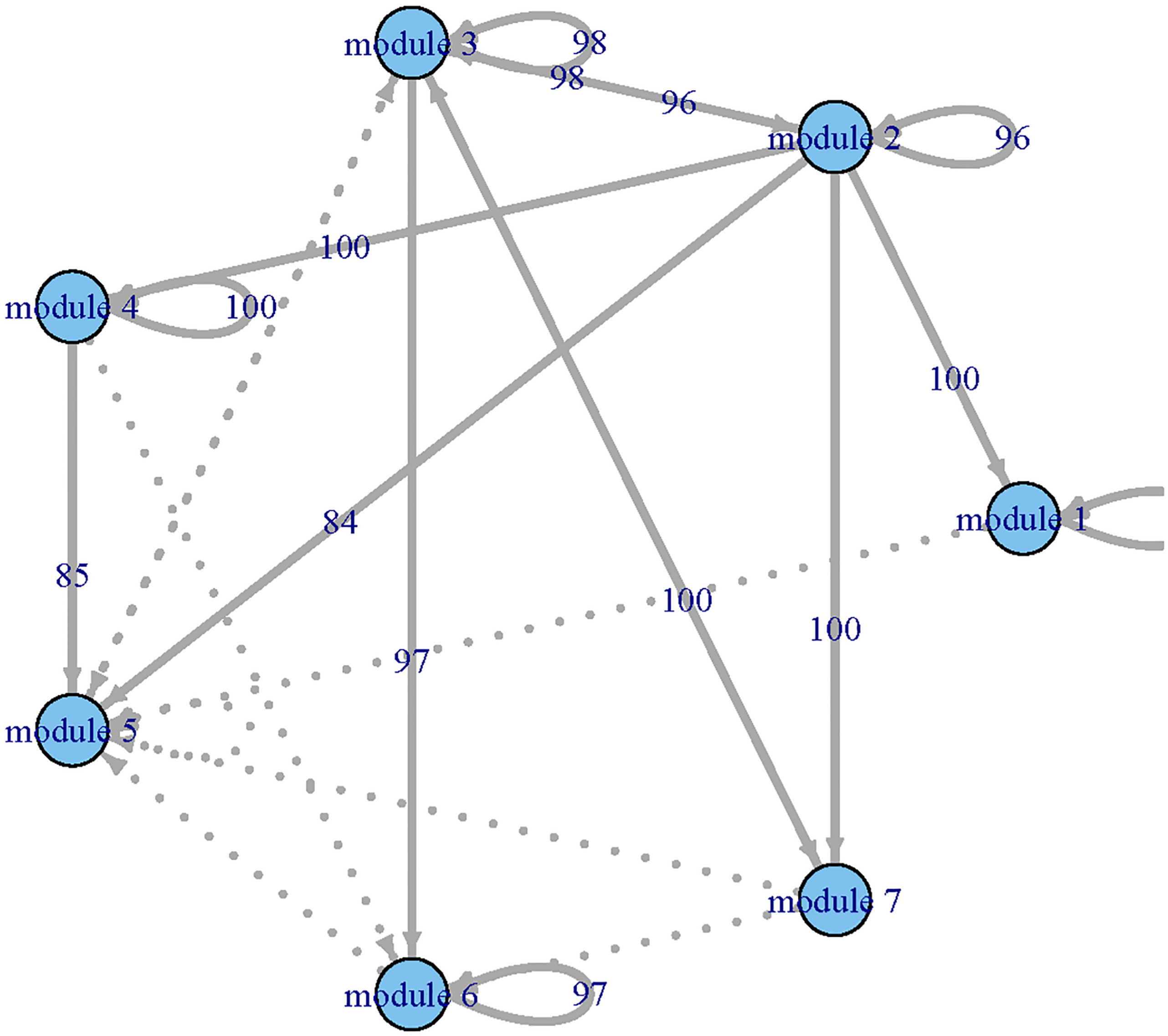

Fig 5. The constructed gene regulatory networks for simulation studies with $N = 288$ and 100 iterations. The legend is the same as in Fig 4.

https://doi.org/10.1371/journal.pone.0192833.g005

The advantages of our method include (i) single-index ODE models can fit the data better than linear ODE models; (ii) the interactions found by single-index ODE models can cover most of the interactions identified by linear ODE models for some of the modules; and (iii) our method is computationally efficient because we can select significant modules and estimate index coefficients simultaneously.

Similar to the linear ODE model, our method needs estimated population means and their corresponding first derivatives, which may be treated as the limitation of the proposed procedure. The PPrLS estimator has good performance in identifying significant modules. But new stable techniques are still needed to group genes to reduce the gene cluster uncertainty because cluster assignment still plays an important role in enhancing the usefulness of this research. It is worth noting that the PPrLS estimates may not be most efficient in terms of estimation accuracy because PPrLS estimation is a nonparametric method and inherits error if the data

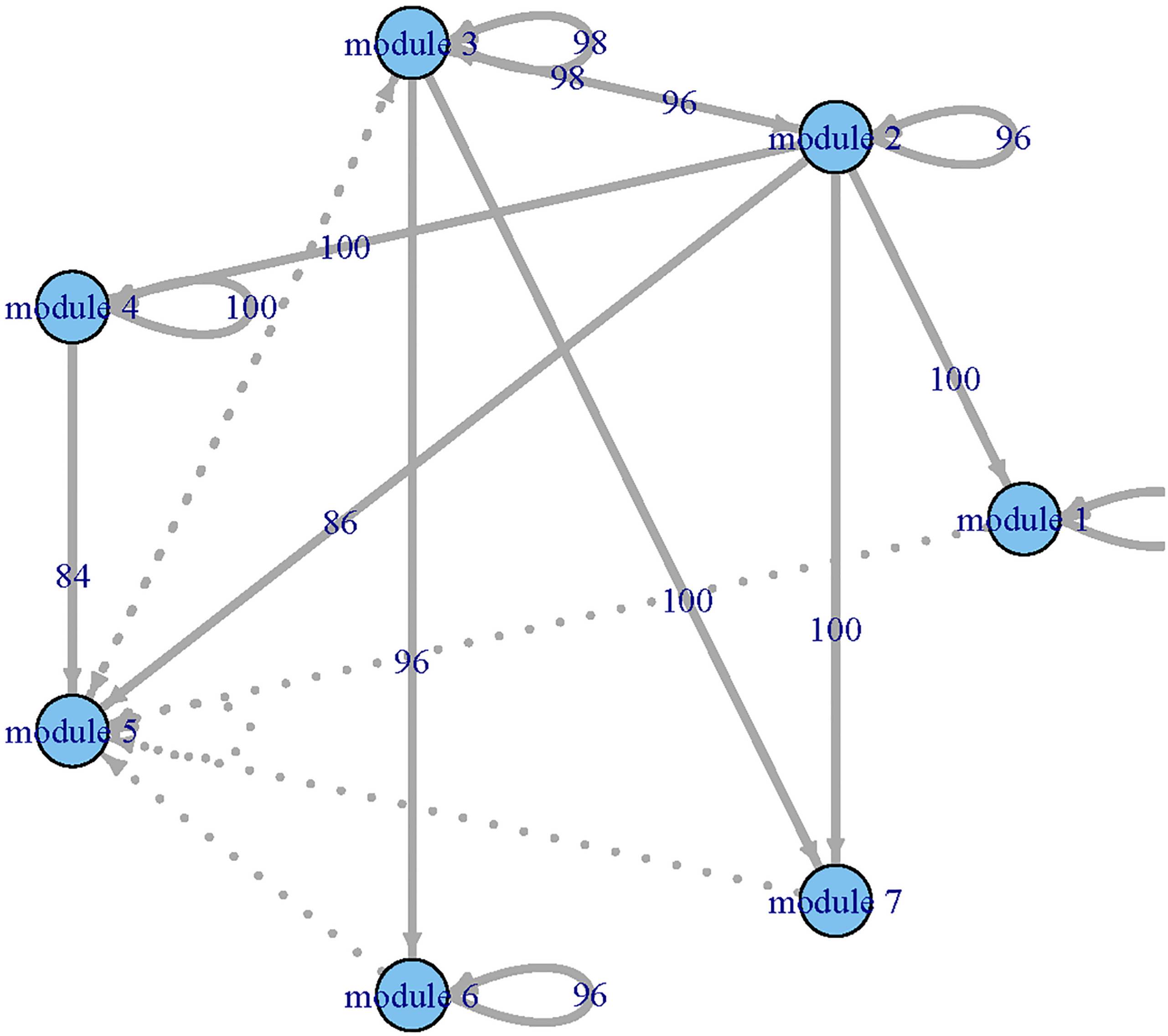

Fig 6. The constructed gene regulatory networks for simulation studies with $N = 360$ and 100 iterations. The legend is the same as in Fig 4.

https://doi.org/10.1371/journal.pone.0192833.g006

contain a large noise. Regulator-regulator interaction exploration depends on the knowledge of gene-regulator relationship, which we study. So the proposed method may provide valuable insights into complicated biological processes with understanding gene-gene and gene-regulator relationships. Overall, our procedure is useful to single out high level (module based) and potential regulator-regulator interactions which are helpful to provide guidance for tedious and costly experiments.

Supporting information

S1 Table. All enriched functional annotations.
(PDF)

S2 Table. The estimated regression coefficients for every functional modules using single-index and linear ODE models.
(PDF)

S1 File. Large-sample properties of the PPrLS procedure and discussion of computation cost.
(PDF)

S2 File. Clustered data.
(TXT)

S3 File. Estimated coefficients using the proposed model and methods.
(TXT)

S4 File. Estimated coefficients using the linear ODE model.
(TXT)

S5 File. Functional annotations.
(XLS)

S6 File. R code for clustering.
(R)

S7 File. R code for drawing Fig 3.
(R)

Acknowledgments

The authors thank the editors and two referees for their constructive comments that have remarkably improved an earlier version of this paper.

Author Contributions

Formal analysis: Qi Zhang, Yao Yu.

Methodology: Qi Zhang, Yao Yu, Jun Zhang, Hua Liang.

Software: Yao Yu.

Supervision: Hua Liang.

Validation: Yao Yu, Hua Liang.

Writing – original draft: Yao Yu, Jun Zhang, Hua Liang.

Writing – review & editing: Qi Zhang, Hua Liang.

References

1. Hecker M, Lambeck S, Toepfer S, van Someren E, Guthke R. Gene regulatory network inference: Data integration in dynamic models– a review. Biosystems. 2009; 96:86–103. https://doi.org/10.1016/j.biosystems.2008.12.004 PMID: 19150482
2. Steuer R, Kurths J, Daub CO, Weise J, Selbig J. The mutual information: Detecting and evaluating dependencies between variables. Bioinformatics. 2002; 18:S231–S240. https://doi.org/10.1093/bioinformatics/18.suppl_2.S231 PMID: 12386007
3. Stuart JM, Segal E, Koller D, Kim SK. A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules. Science. 2003; 302:249–255. https://doi.org/10.1126/science.1087447 PMID: 12934013

4. Rao A, Hero AO, States DJ, Engel JD. Using directed information to build biologically relevant influence networks. Computational systems bioinformatics / Life Sciences Society Computational Systems Bioinformatics Conference. 2007; 6:145–156.

5. Kauffman SA. Metabolic stability and epigenesis in randomly constructed genetic nets. Journal of Theoretical Biology. 1969; 22:437–467. https://doi.org/10.1016/0022-5193(69)90015-0 PMID: 5803332

6. Thomas R. Boolean formalization of genetic control circuits. Journal of Theoretical Biology. 1973; 42:563–585. https://doi.org/10.1016/0022-5193(73)90247-6 PMID: 4588055

7. Bornholdt S. Boolean network models of cellular regulation: prospects and limitations. Journal of the Royal Society, Interface / the Royal Society. 2008; 5:S85–S94. https://doi.org/10.1098/rsif.2008.0132.focus

8. Marbach D, Prill RJ, Schaffter T, Mattiussi C, Floreano D, Stolovitzky G. Revealing strengths and weaknesses of methods for gene network inference. Proceedings of the National Academy of Sciences. 2010; 107:6286–6291. https://doi.org/10.1073/pnas.0913357107

9. Friedman N, Linial M, Nachman I, Pe'er D. Using Bayesian Networks to Analyze Expression Data. Journal of Computational Biology. 2000; 7:601–620. https://doi.org/10.1089/106652700750050961 PMID: 11108481

10. Perrin BEE, Ralaivola L, Mazurie A, Bottani S, Mallet J, d'Alché Buc F. Gene networks inference using dynamic Bayesian networks. Bioinformatics. 2003; 19 Suppl 2:ii138–ii148. https://doi.org/10.1093/bioinformatics/btg1071 PMID: 14534183

11. van Berlo RJP, van Someren EP, Reinders MJT. Studying the Conditions for Learning Dynamic Bayesian Networks to Discover Genetic Regulatory Networks. Simulation. 2003; 79:689–702.

12. Needham CJ, Bradford JR, Bulpitt AJ, Westhead DR. A primer on learning in Bayesian networks for computational biology. PLoS Computational Biology. 2007; 3:e129+. https://doi.org/10.1371/journal.pcbi.0030129 PMID: 17784779

13. Lu T, Liang H, Li H, Wu H. High-dimensional ODEs coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. Journal of the American Statistical Association. 2011; 106:1242–1258. https://doi.org/10.1198/jasa.2011.ap10194 PMID: 23204614

14. Altham PME. Improving the precision of estimation by fitting a model. Journal of the Royal Statistical Society, Series B. 1984; 46:118–119.

15. Fan J, Lv J. A Selective Overview of Variable Selection in High Dimensional Feature Space (Invited Review Article). Statistica Sinica. 2009; 20:101–148.

16. Heinrich R, Schuster S. The Regulation Of Cellular Systems. Springer; 1996.

17. Horowitz JL. Semiparametric and Nonparametric Methods in Econometrics. Springer Series in Statistics. New York: Springer; 2009.

18. Horowitz JL, Härdle W. Direct semiparametric estimation of single-index models with discrete covariates. Journal of the American Statistical Association. 1996; 91(436):1632–1640. https://doi.org/10.1080/01621459.1996.10476732

19. Härdle W, Hall P, Ichimura H. Optimal smoothing in single-index models. The Annals of Statistics. 1993; 21:157–178. https://doi.org/10.1214/aos/1176349020

20. Liang H, Wang NS. Partially linear single-index measurement error models. Statistica Sinica. 2005; 15:99–116.

21. Ichimura H. Semiparametric least squares (SLS) and weighted SLS estimation of single-index models. Journal of Econometrics. 1993; 58:71–120. https://doi.org/10.1016/0304-4076(93)90114-K

22. Duan N, Li KC. Slicing regression: A link-free regression method. The Annals of Statistics. 1991; 19:505–530. https://doi.org/10.1214/aos/1176348109

23. Powell JL, Stock JH, Stoker TM. Semiparametric Estimation of index coefficients. Econometrica. 1989; 57(6):1403–1430. https://doi.org/10.2307/1913713

24. Naik PA, Tsai CL. Single-index model selections. Biometrika. 2001; 88:821–832. https://doi.org/10.1093/biomet/88.3.821

25. Kong E, Xia Y. Variable selection for the single-index model. Biometrika. 2007; 94:217–229. https://doi.org/10.1093/biomet/asm008

26. Liang H, Liu X, Li R, Tsai CL. Estimation and testing for partially linear single-index models. The Annals of Statistics. 2010; 38:3811–3836. https://doi.org/10.1214/10-AOS835 PMID: 21625330

27. Zhang J, Wang T, Zhu L, Liang H. A dimension reduction based approach for estimation and variable selection in partially linear single-index models with high-dimensional covariates. Electronic Journal of Statistics. 2012; 6:2235–2273. https://doi.org/10.1214/12-EJS744

28. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, et al. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization.

Molecular Biology of the Cell. 1998; 9:3273–3297. https://doi.org/10.1091/mbc.9.12.3273 PMID: 9843569

29. Luan Y, Li H. Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data. Bioinformatics. 2004; 20:332–339. https://doi.org/10.1093/bioinformatics/btg413 PMID: 14960459

30. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, et al. A genome-wide transcriptional analysis of the mitotic cell cycle. Molecular cell. 1998; 2:65–73. https://doi.org/10.1016/S1097-2765(00)80114-8 PMID: 9702192

31. Gu C. Smoothing Spline ANOVA Models. New York: Springer-Verlag; 2002.

32. Ma P, Castillo-Davis CI, Zhong W, Liu JS. A data-driven clustering method for time course gene expression data. Nucleic Acids Research. 2006; 34:1261–1269. https://doi.org/10.1093/nar/gkl013 PMID: 16510852

33. Ma P, Zhong W. Penalized clustering of large-scale functional data with multiple covariates. Journal of the American Statistical Association. 2008; 103:625–636. https://doi.org/10.1198/016214508000000247

34. Eilers PHC, Marx BD. Flexible Smoothing with B-splines and Penalties. Statistical Science. 1996; 11:89–102. https://doi.org/10.1214/ss/1038425655

35. Carroll RJ. Spatially-adaptive penalties for spline fitting. Australian and New Zealand Journal of Statistics. 2000; 42:205–223. https://doi.org/10.1111/1467-842X.00119

36. Brumback BA, Ruppert D, Wand MP. Variable Selection and Function Estimation in Additive Nonparametric Regression Using a Data-Based Prior: Comment. Journal of the American Statistical Association. 1999; 94:794–797.

37. Laird NM, Ware JH. Random-effects models for longitudinal data. Biometrics. 1982; 38:963–974. https://doi.org/10.2307/2529876 PMID: 7168798

38. Liang H. Modeling antitumor activity in xenograft tumor treatment. Biometrical Journal. 2005; 47:358–368. https://doi.org/10.1002/bimj.200310113 PMID: 16053259

39. Ruppert D, Wand M, Carroll R. Semiparametric Regression. New York: Cambridge University Press; 2003.

40. Davidian M, Giltinan DM. Nonlinear Models for Repeated Measurement Data. New York: Chapman and Hall; 1995.

41. Liang H, Wu H. Parameter estimation for differential equation models using a framework of measurement error in regression models. Journal of the American Statistical Association. 2008; 103. https://doi.org/10.1198/016214508000000797 PMID: 19956350

42. Tibshirani R. Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistical Society, Series B. 1996; 58:267–288.

43. Fan J, Li R. Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties. Journal of the American Statistical Association. 2001; 96:1348–1360. https://doi.org/10.1198/016214501753382273

44. Zou H. The Adaptive LASSO and Its Oracle Properties. Journal of the American Statistical Association. 2006; 101:1418–1429. https://doi.org/10.1198/016214506000000735

45. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society Series B. 2005; 67:301–320. https://doi.org/10.1111/j.1467-9868.2005.00503.x

46. Zou H, Zhang HH. On the adaptive elastic-net with a diverging number of parameters. The Annals of Statistics. 2009; 37:1733–1751. https://doi.org/10.1214/08-AOS625 PMID: 20445770

47. Fan J, Gijbels I. Local Polynomial Modelling and Its Applications. vol. 66. London: Chapman & Hall; 1996.

48. Wang H, Li R, Tsai CL. Tuning parameter selectors for the smoothly clipped absolute deviation method. Biometrika. 2007; 94:553–568. https://doi.org/10.1093/biomet/asm053 PMID: 19343105

49. Wang L, Chen G, Li H. Group SCAD regression analysis for microarray time course gene expression data. Bioinformatics. 2007; 23:1486–1494. https://doi.org/10.1093/bioinformatics/btm125 PMID: 17463025

50. Huang DW, Sherman BT, Lempicki RA. Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nature protocols. 2008; 4:44–57. https://doi.org/10.1038/nprot.2008.211

51. Huang DW, Sherman BT, Lempicki RA. Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. Nucleic acids research. 2009; 37:1–13. https://doi.org/10.1093/nar/gkn923

52. D'Haeseleer P, Wen X, Fuhrman S, Somogyi R. Linear Modeling Of mRNA Expression Levels During CNS Development And Injury. Pacific Symposium on Biocomputing. 1999; 4:41–52.

53. Bansal M, Della Gatta G, di Bernardo D. Inference of gene regulatory networks and compound mode of action from time course gene expression profiles. Bioinformatics. 2006; 22:815–822. https://doi.org/10.1093/bioinformatics/btl003 PMID: 16418235

54. Wessels LFA, van Someren EP, Reinders MJT. A comparison of genetic network models. Pacific Symposium on Biocomputing (PSB). 2001; 6:508–519.

55. Lee TI, Rinaldi NJ, Robert F, Odom DT, Bar-Joseph Z, Gerber GK, et al. Transcriptional Regulatory Networks in Saccharomyces cerevisiae. Science. 2002; 298:799–804. https://doi.org/10.1126/science.1075090 PMID: 12399584

56. Taylor I, McIntosh P, Pala P, Treiber M, Howell S, Lane A, et al. Characterization of the DNA-binding domains from the yeast cell-cycle transcription factors Mbp1 and Swi4. Biochemistry. 2000; 39:3943–3954. https://doi.org/10.1021/bi992212i PMID: 10747782

57. Nair M, McIntosh P, Frenkiel T, Kelly G, Taylor I, Smerdon S, et al. NMR structure of the DNA-binding domain of the cell cycle protein Mbp1 from Saccharomyces cerevisiae. Biochemistry. 2003; 42:1266–1273. https://doi.org/10.1021/bi0205247 PMID: 12564929

58. Bouquin N, Johnson AL, Morgan BA, Johnston LH. Association of the Cell Cycle Transcription Factor Mbp1 with the Skn7 Response Regulator in Budding Yeast. Molecular Biology of the Cell. 1999; 10:3389–3400. https://doi.org/10.1091/mbc.10.10.3389 PMID: 10512874