Title

Inference of Gene Regulatory Networks Using Bayesian Nonparametric Regression and Topology Information.

Authors

Fan, Yue

Wang, Xiao

Peng, Qinke

Publication Date

2017

DOI

10.1155/2017/8307530

Peer reviewed

Hindawi Publishing Corporation
Computational and Mathematical Methods in Medicine
Volume 2017, Article ID 8307530, 8 pages
http://dx.doi.org/10.1155/2017/8307530

Research Article

Inference of Gene Regulatory Networks Using Bayesian Nonparametric Regression and Topology Information

Yue Fan, Xiao Wang, and Qinke Peng

Systems Engineering Institute, School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China

Correspondence should be addressed to Qinke Peng; qkpeng@xjtu.edu.cn

Received 18 August 2016; Accepted 24 November 2016; Published 4 January 2017

Academic Editor: Konstantin Blyuss

Gene regulatory networks (GRNs) play an important role in cellular systems and are important for understanding biological processes. Many algorithms have been developed to infer the GRNs. However, most algorithms only pay attention to the gene expression data but do not consider the topology information in their inference process, while incorporating this information can partially compensate for the lack of reliable expression data. Here we develop a Bayesian group lasso with spike and slab priors to perform gene selection and estimation for nonparametric models. B-spline basis functions are used to capture the nonlinear relationships flexibly and penalties are used to avoid overfitting. Further, we incorporate the topology information into the Bayesian method as a prior. We present the application of our method on DREAM3 and DREAM4 datasets and two real biological datasets. The results show that our method performs better than existing methods and the topology information prior can improve the result.

1. Introduction

Gene regulatory network plays an important role in diverse cellular functions. A reliable method to identify the structure and dynamics of such regulation is important for understanding complex biological processes and is helpful for treatment of diseases. With the development of high throughout technologies in recent years, gene expression data has provided a useful way to investigate the cellular system.

Generally, there are two types of gene expression data used to predict the structure of GRNs, which are steady-state data and time-series data. The steady-state data measures the steady-state levels in different samples, while time-series data measures the expression levels at several successive time points. Since the time-series data contains the dynamic information of the network while the steady-state data does not [1], we focus on the time-series data in this paper.

Over the last several years, a number of network inference methods have been developed to tackle this problem, including Bayesian network [2, 3], dynamic Bayesian network [4, 5], Boolean network [6, 7], ordinary differential equation [8, 9], and mutual information [10, 11]. A comprehensive review can be found in [12, 13]. Among these methods, dynamic Bayesian network has become the major focus for inferring gene regulatory network because it can infer causal interactions, model cyclic interactions, and has less computational complexity than ordinary differential equation.

Inferring a GRN from time-series data is known to be challenging partly due to the high number of genes relative to the number of data points. More importantly, the interactions between genes are typically nonlinear; thus linear model may be inefficient to recognize the nonlinear interactions. A flexible way to solve this problem is to use B-spline functions to describe the nonlinear interactions, and the B-spline functions have been used to infer GRNs in previous studies [14, 15]. A key problem in spline regression is the knot selection which greatly influences the curve fitting. Reference [14] suggested using penalized-splines to avoid overfitting and reduce the number of parameters to be estimated. Among many penalized methods, lasso [16] is the most popular method due to its ability to select and estimate simultaneously and can produce exact 0 estimates. Group lasso [17] was also developed to select grouped variables. Reference [18] proposed group lasso or Bayesian group lasso when spline regression was used because the predictors belong to a same gene forming a natural group. Reference [19]

also developed a Bayesian adaptive group lasso to perform simultaneous model selection and estimation for B-spline regression. However, Bayesian spline regression methods still predict a lot of false positive interactions because of the indirect effects existing in the GRNs.

Recently, [20] proposed a new method which uses network topology information to improve gene regulatory network inference; they used a prior that both prokaryotic and eukaryotic transcription networks exhibit an approximately scale-free out-degree distribution while the in-degree distribution is a more restricted exponential function; this structure property is described in [21]. Reference [20] also pointed out that 79% or more genes regulators are less than 3. This property means that most genes in a GRN are regulated by a few regulators and may be possible to be combined with the B-spline regression to improve the results of the GRN inference.

In this paper, we work with a dynamic Bayesian network and use spline regression to detect the nonlinear interactions between genes. A Bayesian group lasso is also used to avoid overfitting and reduce the number of parameters to be estimated. Comparing with group lasso, Bayesian group lasso is a better choice because there are 2 major advantages of Bayesian selection methods: (1) The tuning parameter can be set flexibly. (2) The topology information can be incorporated easily. Further, instead of taking a traditional Bayesian group lasso, we use a Bayesian group lasso model with spike and slab priors since this problem only requires the sparsity on the group level and spike and slab priors can exclude or include the entire group of B-spline basis functions. Finally, we incorporate the topology information as a prior in the Bayesian approach which controls the size of the selected model. This method is assessed by applying to DREAM3 and DREAM4 datasets and two real biological datasets.

2. Method

2.1. The Nonlinear Regression Model for GRN Inference. Consider an $G \times T$ matrix Y, where T is the number of the gene expression levels measured times and G is the number of genes. A DBN model represents probabilistic relationships between genes via a directed acyclic graph ϑ. In this graph, genes are represented by a set of nodes $V = \{V_1, \ldots, V_G\}$ and the interactions between genes are represented by a set of directed edges $E \subseteq \{(i, j) : i, j \in V\}$. A directed edge from node i to node j means gene i is a regulator of gene j. The probability distribution of genes Y_t given its parents can be expressed as

$$p\left(Y_t \mid Y_{t-1}\right) = \prod_{g=1}^{G} p\left(Y_{g,t} \mid Pa\left(Y_{g,t}\right)\right), \tag{1}$$

where $Y_{g,t}$ is the gene g expression level at time t and $Pa(Y_{g,t})$ is the set of all the parent nodes of gene g at time t. In the case of the regression-based DBN, the conditional distribution $p(Y_{g,t} \mid Pa(Y_{g,t}))$ can be written as

$$y_{g,t} = f^t\left(y_{-g,t-1}\right) + \varepsilon_g, \quad g = 1, \ldots, G, \; t = 1, \ldots, T, \tag{2}$$

where $y_{g,t}$ is the expression level of gene g and $y_{-g,t-1}$ is the vector without y_{t-1}:

$$y_{-g,t-1} = \left(y_{1,t-1}, \ldots, y_{g-1,t-1}, y_{g+1,t-1}, \ldots, y_{G,t-1}\right). \tag{3}$$

We assume that the GRN is a time-invariant network; thus $f^t(\cdot) = f(\cdot)$ and the error term $\varepsilon_g \sim N(0, \sigma^2)$. Although $f(\cdot)$ can be characterized by any nonlinear functional representation, [15] suggested using B-spline basis functions instead of using Fourier basis, wavelets, or other nonlinear basis functions because of the pattern of the relationship between genes is unknown. Therefore, we also use B-spline basis functions in this article and the regulatory relationships can be written as

$$\begin{aligned} y_{g,t} = \mu_g &+ f\left(y_{1,t-1}\right) + \cdots + f\left(y_{g-1,t-1}\right) \\ &+ f\left(y_{g+1,t-1}\right) + \cdots + f\left(y_{G,t-1}\right) + \varepsilon_g, \end{aligned} \tag{4}$$

where μ_g is the intercept and $f(y_{i,t}) = \sum_{k=1}^{M} \beta_{ik} B_{ik}(y_{i,t})$. $\{B_{ik}(y_{i,t})\}$ are M B-spline basis functions of degree l and β_{ik} is the parameters to estimate from data. Let $\{\kappa_i\}$ be the set of r equally spaced knots with $\min\{y_i\} = \kappa_{i1} < \kappa_{i2} < \cdots < \kappa_{ir} = \max\{y_i\}$, and $M = r + l$. We get rid of the subscript g for the variables for simplicity of notation. Then the regression equation can be written as

$$y = \mu + X\beta + \varepsilon, \tag{5}$$

where X is the bases matrix of size $T \times MG$ and β is the corresponding coefficients vector.

2.2. Incorporating the Topology Information and Bayesian Inference. We use the Bayesian group lasso method proposed in [22]; the hierarchical Bayesian model is

$$\begin{aligned} Y \mid X, \beta, \sigma^2 &\sim N_n\left(X\beta, \sigma^2 I_n\right), \\ \beta_g \mid \sigma^2, \tau_g^2 &\sim \gamma N_{m_g}\left(0, \sigma^2 \tau_g^2 I_{m_g}\right) + \left(1 - \gamma_g\right)\delta_0\left(\beta_g\right), \\ &\qquad g = 1, 2, \ldots, G, \\ \tau_g^2 &\sim \text{Gamma}\left(\frac{m_g + 1}{2}, \frac{\lambda^2}{2}\right), \\ &\qquad g = 1, 2, \ldots, G, \\ \sigma^2 &\sim \text{Inverse Gamma}\,(a, b), \\ \gamma_g &\sim \text{Bernoulli}\,(p), \\ p &\sim \text{Beta}\,(c, d), \end{aligned} \tag{6}$$

where $m_j = 1$ for $j = 1$ and $m_j = M$ otherwise. Here we use a spike and slab prior on β and get the ranking of the potential regulatory links from γ. Although we can place a positive and very small p as a prior when the in-degree of the target gene is small, there are still a lot of false positive interactions to be predicted. Inspired by the idea of maxP technique proposed

by [20], we use a prior proposed in [23], to place a restriction on γ, that only allow the model to be of small size.

$$\begin{aligned} &\gamma \sim \text{Bernoulli}(p) \quad |\gamma| \le k, \\ &\gamma = 0 \quad \text{otherwise}. \end{aligned} \tag{7}$$

Here the integer-valued hyperparameter k restricts the maximum number of parents for the target gene in each iteration. However, there are still some genes regulated by a large number of genes. Therefore, a fixed k will affect the accuracy of the prediction. Thus a uniform prior on $[1, m]$ is placed on k, where m is a predetermined integer. Then the model becomes

$$\begin{aligned} Y \mid X, \beta, \sigma^2 &\sim N_n\left(X\beta, \sigma^2 I_n\right), \\ \beta_g \mid \sigma^2, \tau_g^2 &\sim \gamma N_{m_g}\left(0, \sigma^2\tau_g^2 I_{m_g}\right) + \left(1-\gamma_g\right)\delta_0\left(\beta_g\right), \\ &\qquad g = 1, 2, \ldots, G, \\ \tau_g^2 &\sim \text{Gamma}\left(\frac{m_g+1}{2}, \frac{\lambda^2}{2}\right), \\ &\qquad g = 1, 2, \ldots, G, \\ \sigma^2 &\sim \text{Inverse Gamma}(a, b) \\ p(\gamma) &= \begin{cases} \gamma \sim \text{Bernoulli}(p), & |\gamma| \le k, \\ \gamma = 0, & \text{otherwise}, \end{cases} \\ p &\sim \text{Beta}(c, d), \\ k &\sim \text{uniform}(1, m). \end{aligned} \tag{8}$$

The likelihood is

$$\begin{aligned} &p\left(y \mid X, \beta, \sigma^2\right) \\ &\quad \propto \left(\sigma^2\right)^{-n/2} \exp\left(-\frac{(y-X\beta)^T(y-X\beta)}{2\sigma^2}\right). \end{aligned} \tag{9}$$

According to the prior and the likelihood above, the joint posterior distribution on data is

$$\begin{aligned} &p\left(\beta, \sigma^2, \tau^2, \gamma \mid Y, X\right) \propto \left(\sigma^2\right)^{-n/2} \\ &\cdot \exp\left(-\frac{(y-X\beta)^T(y-X\beta)}{2\sigma^2}\right) \\ &\cdot \prod_{g=1}^{G}\left[\gamma\left(\sigma^2\tau_g^2\right)^{-m_g/2}\exp\left(-\frac{\beta_g^T\beta_g}{2\sigma^2\tau_g^2}\right)\right. \\ &\left. + (1-\gamma)\,\delta_0\left(\beta_g\right)\right]\prod_{g=1}^{G}\left(\lambda^2\right)^{(m_g+1)/2}\left(\tau_g^2\right)^{(m_g+1)/2-1} \\ &\cdot \exp\left(-\frac{\lambda^2}{2}\tau_g^2\right)\left(\sigma^2\right)^{-a-1}\exp\left(-\frac{b}{\sigma^2}\right)p^{c-1}(1 \\ &- p)^{d-1}\,p(\gamma)\,p(k). \end{aligned} \tag{10}$$

The Gibbs sampling scheme is as follows: We use $\beta_{-g} = (\beta_1, \ldots, \beta_{g-1}, \beta_{g+1}, \ldots, \beta_G)$ to denote the coefficient vector β without the gth group and $X_{-g} = (X_1, \ldots, X_{g-1}, X_{g+1}, \ldots, X_G)$ to denote the covariate matrix corresponding to β_{-g}. The full conditions of $(\gamma_g = 1, \beta_g)$ and $(\gamma_g = 0, \beta_g)$ are

$$\begin{aligned} &p\left(\gamma_g = 1, \beta_g \mid \text{rest}\right) \propto \left(2\pi\tau_g^2\sigma^2\right)^{-m_g/2} \\ &\quad \cdot \exp\left(-\frac{\beta_g^T\Sigma_g\beta_g - 2u_g^T\beta_g}{2\sigma^2}\right)p\left(\gamma_g = 1\right), \\ &p\left(\gamma_g = 0, \beta_g \mid \text{rest}\right) \\ &\quad \propto \exp\left(-\frac{\beta_g^T X_g^T X_g\beta_g - 2u_g^T\beta_g}{2\sigma^2}\right)p\left(\gamma_g = 0\right) \\ &\quad \cdot \delta_0\left(\beta_g\right), \end{aligned} \tag{11}$$

where $\mu_g = \Sigma_g X_g^T(Y - X_{-g}\beta_{-g})$ and $\Sigma_g = (X_g^T X_g + (1/\tau_g^2)I_{m_g})^{-1}$.

Integrating out β_g, we have

$$\begin{aligned} &p\left(\gamma_g = 1, \beta_g \mid \text{rest}\right) \propto p\left(\gamma_g = 1\right)\left(\tau_g^2\right)^{-m_g/2}\left|\Sigma_g\right|^{1/2} \\ &\quad \cdot \exp\left\{\frac{\left(X_g^T\left(Y - X_{-g}\beta_{-g}\right)\right)^T\Sigma_g\left(X_g^T\left(Y - X_{-g}\beta_{-g}\right)\right)}{2\sigma^2}\right\}, \\ &p\left(\gamma_g = 0, \beta_g \mid \text{rest}\right) \propto p\left(\gamma_g = 0\right). \end{aligned} \tag{12}$$

From these equations, we can draw γ_g through

$$p\left(\gamma_g = 0 \mid \text{rest}\right) = \frac{p}{p + (1-p)\left(\tau_g^2\right)^{-m_g/2}\left|\Sigma_g\right|^{1/2}\exp\left(\left(X_g^T\left(y - X_{-g}\beta_{-g}\right)\right)^T\Sigma_g\left(X_g^T\left(y - X_{-g}\beta_{-g}\right)\right)/2\sigma^2\right)}. \tag{13}$$

Then the full conditional posterior distribution of β_g is

$$p\left(\beta_g \mid \gamma_g = 1, \text{rest}\right) \propto \exp\left(-\frac{\beta_g^2 - 2\mu_g}{2\sigma^2\Sigma_g}\right). \tag{14}$$

Thus, the full conditional distribution of β_g is a normal distribution:

$$\begin{aligned} \beta_g \mid \gamma_g = 1, \text{rest} &\sim N\left(u_g, \sigma^2\Sigma_g\right), \\ p\left(\beta_g = 0 \mid \gamma_g = 0, \text{rest}\right) &= 1. \end{aligned} \tag{15}$$

The full conditions of $(\tau_g^2, \gamma_g = 1)$ and $(\tau_g^2, \gamma_g = 0)$ are

$$\begin{aligned} p\left(\tau_g^2, \gamma_g = 1\right) &\propto \left(\tau_g^2\right)^{-1/2} \exp\left(-\frac{\beta_g^T\beta_g}{2\sigma^2\tau_g^2} - \frac{\lambda^2}{2}\tau_g^2\right), \\ p\left(\tau_g^2, \gamma_g = 0\right) &\propto \left(\tau_g^2\right)^{(m_g+1)/2-1} \exp\left(-\frac{\lambda^2}{2}\tau_g^2\right). \end{aligned} \tag{16}$$

Then the full conditional distribution of τ_g^2 is

$$\begin{aligned} \frac{1}{\tau_g^2}, \gamma_g = 1 \mid \text{rest} &\sim \text{Inverse Gaussian}\left(\frac{\lambda\sigma}{\left\|\beta_g\right\|_2}, \lambda^2\right), \\ \frac{1}{\tau_g^2}, \gamma_g = 0 \mid \text{rest} &\sim \text{Inverse Gamma}\left(\frac{m_j + 1}{2}, \frac{\lambda^2}{2}\right). \end{aligned} \tag{17}$$

The full conditional distribution of σ^2 is

$$\begin{aligned} p\left(\sigma^2 \mid \text{rest}\right) &\propto \left(\sigma^2\right)^{-n/2-(1/2)\sum_{g=1}^G m_g Z_g - a - 1} \\ &\cdot \exp\left(-\frac{\left(y - X\beta\right)^T\left(y - X\beta\right) + \beta^T D_\tau^{-1}\beta + b}{2\sigma^2}\right). \end{aligned} \tag{18}$$

Then the conditional posterior distribution of σ^2 is

$$\begin{aligned} \sigma^2 \mid \text{rest} \sim \text{Inverse Gamma}\Bigg(&\frac{n}{2} + \frac{1}{2}\sum_{g=1}^G m_g Z_g \\ &+ a, \frac{1}{2}\left[\left(Y - X\beta\right)^T\left(Y - X\beta\right) + \beta^T D_\tau^{-1}\beta\right] + b\Bigg), \end{aligned} \tag{19}$$

where $Z_g = \{1, \text{ if } \gamma_g = 0;\ 0, \text{ if } \gamma_g \neq 0\}$ and $D_\tau = \text{diag}\{\tau_1^2, \tau_2^2, \ldots, \tau_G^2\}$. And it can be verified that the conditional posterior distributions of other parameters are

$$\begin{aligned} p \mid \text{rest} &\sim \text{Beta}\left(c + \sum_{g=1}^G Z_g, b + \sum_{g=1}^G m_g - \sum_{g=1}^G Z_g\right), \\ k \mid \text{rest} &\sim \text{uniform}\left(\sum_{g=1}^G Z_g, m\right). \end{aligned} \tag{20}$$

And a Monte Carlo EM algorithm is used to estimate λ:

$$\lambda^{(k)} = \sqrt{\frac{p + G}{\sum_{g=1}^G E_{\lambda^{(k-1)}}\left[\tau_g^2 \mid Y\right]}}, \tag{21}$$

where p equal to $1 + m_j \times (G - 1)$ is the number of the total regressors and $E_{\lambda^{(k-1)}}[\tau_g^2 \mid Y]$ can be replaced by the sample average of τ_g^2 generated in the $k-1$th step of the Gibbs sampler. We choose the second half of the samples and the result is the average of the samples.

3. Results

To demonstrate the effectiveness of the topology information and the B-spline functions, our method is used to infer GRNs from in silico time-series data and real biological data; a linear model with topology information and a nonlinear model without topology information are also applied as competing methods. Here we use the time-series data in DREAM3 and DREAM4 challenges as the in silico data, and we use a cell cycle regulatory subnetwork in *Saccharomyces cerevisiae* and Human Hela cell network as the real biological datasets. We generate 10000 samples from the posterior distribution and choose the second half of the samples to derive the results. The posterior estimates of all the parameters are obtained through the posterior averages of the chains. For the B-spline functions, we adopt the setting as [14] and use a cubic B-spline with 10 interior knots. Here we choose $m = 5$ in our experiments.

3.1. Application to In Silico Networks. We first evaluate our method on DREAM4 challenges networks of sizes 10 and 100 [24–26]. The size 10 network data consists of 5 simulated networks, each of which consists of 21 time points and 5 replicates. The size 100 network data also consists of 5 simulated networks, each of which consists of 21 time points and 10 replicates. We also evaluate our method on DREAM3 challenges networks of sizes 10, which is also used in [27]. This data consists of 5 simulated networks, each of which consists of 21 time points. There are also steady-state data provided by the DREAM4 challenge. However, we only focus on time-series data in this article. Although the winning entry in DREAM4 competition used only the knock-out data [28] and combining the time-series and steady-state data can achieve much better results [27, 29], it is infeasible to do knock-out experiments for all genes in practice and generally the knock-out experiments only are done for a small part of genes [30].

Each of the five networks is inferred using all available time-series data, and the area under the receiver operating characteristic (AUROC) curve and the area under precision-recall (AUPR) curve are computed according to the gold standard network topology provided by DREAM3 and DREAM4 challenge. The prediction performances on the DREAM4 10-gene networks and 100-gene networks are summarized in Tables 1 and 2. Table 1 shows that the Bayesian lasso and Bayesian group lasso perform similarly on size 10 data while the BGL_prior has a better performance than the methods above in both average AUROC and AUPR. For net 2 and net 5, the BGL_prior outperforms other methods significantly. We also compared our method with another 2 dynamic Bayesian network methods [31]; the result of our method is also comparable to these methods. Table 2 shows that the nonlinear model performs poorly on this dataset;

Table 1: The prediction performances on the DREAM4 10-gene networks.

	Method	Net 1	Net 2	Net 3	Net 4	Net 5	Average
AUROC	BL	0.7956	0.6334	0.6356	0.8292	0.8034	0.7394
	BGL	0.7956	0.6537	0.6507	0.8422	0.8194	0.7523
	BGL_prior	0.8267	0.7188	0.6640	0.8472	0.8547	0.7823
	G1DBN	0.73	0.64	0.68	0.85	0.92	0.7640
	VBSSM	0.73	0.66	0.77	0.80	0.84	0.7600
AUPR	BL	0.4222	0.3711	0.3926	0.5235	0.4683	0.4355
	BGL	0.4613	0.3582	0.3781	0.5392	0.3368	0.4147
	BGL_prior	0.4750	0.4090	0.3062	0.6207	0.5022	0.4626
	G1DBN	0.37	0.34	0.45	0.69	0.77	0.5240
	VBSSM	0.38	0.41	0.49	0.46	0.64	0.4760

Table 2: The prediction performances on the DREAM4 100-gene networks.

	Method	Net 1	Net 2	Net 3	Net 4	Net 5	Average
AUROC	BL	0.7180	0.6040	0.6748	0.6551	0.7269	0.6758
	BGL	0.5768	0.5915	0.6148	0.5692	0.5829	0.5878
	BGL_prior	0.7117	0.6745	0.7062	0.6859	0.7327	0.7029
	G1DBN	0.68	0.64	0.68	0.66	0.72	0.6760
	VBSSM	0.59	0.56	0.59	0.67	0.71	0.6240
AUPR	BL	0.1177	0.0830	0.1154	0.1103	0.0776	0.1008
	BGL	0.0318	0.0984	0.0440	0.0685	0.0409	0.0567
	BGL_prior	0.508	0.0656	0.0967	0.1021	0.0730	0.0776
	G1DBN	0.11	0.10	0.13	0.10	0.11	0.1100
	VBSSM	0.08	0.05	0.11	0.10	0.09	0.0860

Table 3: The prediction performances on the DREAM3 10-gene networks.

	Method	Ecoli 1	Ecoli 2	Yeast 1	Yeast 2	Yeast 3	Average
AUROC	BL	0.4948	0.6880	0.6200	0.4412	0.4091	0.5306
	BGL	0.5339	0.7813	0.5525	0.5348	0.4646	0.5734
	BGL_prior	0.6237	0.7876	0.6363	0.5034	0.4987	0.6099
	Inferelator 1.0	0.49	0.52	0.56	0.45	0.48	0.5000
	Additive ODE	0.53	0.54	0.45	0.53	0.48	0.5060
AUPR	BL	0.2455	0.5325	0.2981	0.2979	0.2009	0.3150
	BGL	0.2076	0.6096	0.2749	0.2871	0.2199	0.3220
	BGL_prior	0.2427	0.6144	0.2795	0.2544	0.2347	0.3251
	Inferelator 1.0	0.15	0.21	0.22	0.33	0.28	0.2380
	Additive ODE	0.16	0.20	0.10	0.31	0.23	0.2000

while the topology information can remarkably improve the prediction performance of the nonlinear model, the Bayesian group lasso with topology information outperforms the Bayesian group lasso methods in both AUROC and AUPR, and these methods also have higher AUROC than linear model, although the AUPR is a little worse. Compared with the results of the other 2 DBN methods, the result of the Bayesian lasso is similar to them and our method still has the highest average AUROC. The prediction performances on the DREAM3 10 gene networks are summarized in Table 3. We also compared our method with another additive model based on ODE [27] and Inferelator 1.0 [32]. For Ecoli 1, Ecoli 2, Yeast 2, and Yeast 3, the 3 additive models perform better than the linear model; for Yeast 1, although BL performs much better than BGL, the BGL_prior still gets slightly better results. The average results show that BGL_prior outperforms the other methods in both AUROC and AUPR.

3.2. Application to IRMA Network. The IRMA network data is a subnetwork embedded in *Saccharomyces cerevisiae* consisting of 5 genes: CBF1, GAL4, SWI5, GAL80, and ASH1. Both of the two time-series gene expressions include switch-on data and switch-off data. The switch-on data is taken from 5 experiments and the switch-off data is taken

Table 4: The prediction performances on IRMA network.

Method	TP	FP	TN	FN	PR	RR	F
BL	5	9	3	3	0.3571	0.6250	0.4545
BGL	4	3	9	4	0.5714	0.5000	0.5333
BGL_prior	5	3	9	3	0.6250	0.6250	0.6250
Morrissey's method	1	1	13	5	0.5	0.1667	0.2500
TSNIF	5	2	10	3	0.7143	0.6250	0.6667
BANJO	5	6	6	3	0.4545	0.6250	0.5263

Table 5: The prediction performances on Hela network.

Method	TP	FP	TN	FN	PR	RR	F
BL	2	9	54	7	0.1818	0.2222	0.2000
BGL	5	18	45	4	0.2174	0.5556	0.3125
BGL_prior	5	14	49	4	0.2632	0.5556	0.3571
Morrissey's method	3	15	48	6	0.1667	0.6667	0.2222
TALasso	3	7	56	6	0.3	0.3333	0.3158
grpLasso	4	13	50	5	0.2353	0.4444	0.3076
CNET	4	7	56	5	0.3636	0.4444	0.4000

from 4 experiments with a total of 142 samples measured by [33] and also used in [20, 34]. The IRMA network is well studied and is a gold standard network. This network also has a fixed topology and the genes in the network are not regulated by other yeast genes. Here we use the precision rate (PR = TP/(TP + FP)), recall rate (RR = TP/(TP + FN)), and F-measure = $(2 \cdot \text{PR} \cdot \text{RR})/(\text{PR} + \text{RR})$ to evaluate the performance and select a best threshold as [35]. The signs of the interactions and self-regulations are not considered; thus the total number of the potential interactions is 20. Table 4 shows the inference performance for the IRMA network. The nonlinear model still performs better than linear model. The method with the prior has a higher TP than the Bayesian group lasso, which implies that the topology information improves the performance. Comparing with another B-spline based method [14] and the method used and compared in [36], although our method cannot achieve the best performance, it is still comparable to the TSNIF and performs much better than another B-spline based method.

3.3. Application to Hela Network. We then apply our method on the cell cycle genes in human cancer cell lines (HeLa) which were analyzed by Whitfield [37]. A subnet consisting of 9 Hela cell genes was extracted by Sambo et al. [38] and the topology of this gene regulatory network is determined in the BioGRID database. They also developed a method called CNET to analysis the Hela network. This network is also analyzed by Lozano et al. [39] and Shojaie and Michailidis [40]; they proposed 2 l_1 penalized method, grpLasso, and TAlasso to infer causal interactions.

Here we use the third experiment of Whitfield [37] as the previous studies, consisting of 47 samples. The results of CNET, grpLasso, and TAlasso are taken from [40]. Table 5 shows the inference performance for the Hela network. Comparing with the BGL, the BGL_prior has a higher precision. Comparing with other methods, the penalized method seems to perform better than another B-spline based method and has a similar performance to the other 2 l_1 penalized methods and all the true positives of Morrissey's method are also found by BGL and BGL_prior. On the other hand, the interactions from RFC4 to CDC2 and CDC2 to CCNE1 are found not only by BGL and BGL_prior, but also by 2 of other 3 comparable methods. It may be because these interactions exist in real regulatory network but are not included in the BioGRID dataset.

4. Conclusion

In this study, we propose a fully Bayesian method, based on B-spline, group lasso, and topology information to infer gene regulatory network from time-series data. We use B-spline functions to capture the nonlinear interactions between genes, l_1 norm penalty to prevent overfitting, and topology information, the knowledge of the exponential decrease in in-degree that most genes have only a small number of regulators as a prior. A spike and slab prior is used to facilitate variable selection by putting a multivariate point mass at $0_{m\times 1}$ for an m-dimensional coefficients group. The performance of the proposed method is demonstrated by applications to the DREAM4 in silico data of sizes 10 and 100 network challenges and the real biological data of IRMA and Hela cell network. The results show that the topology information indeed contributes to the gene regulatory network inference which can improve the AUROC remarkably of the DREAM4 in silico data and improve the results of the IRMA network and Hela cell data. B-spline regression model also performs better than linear model in real biological data. Therefore, our method is an effective way of inferencing gene regulatory network from the time-series data.

Competing Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

This work is supported in part by the National Science Foundation of China, under Grant 61173111.

References

[1] L.-Z. Liu, F.-X. Wu, and W.-J. Zhang, "Properties of sparse penalties on inferring gene regulatory networks from time-course gene expression data," *IET Systems Biology*, vol. 9, no. 1, pp. 16–24, 2015.

[2] A. V. Werhli and D. Husmeier, "Reconstructing gene regulatory networks with Bayesian networks by combining expression data with multiple sources of prior knowledge," *Statistical Applications in Genetics and Molecular Biology*, vol. 6, no. 1, 2007.

[3] Y. Watanabe, S. Seno, Y. Takenaka, and H. Matsuda, "An estimation method for inference of gene regulatory net-work using Bayesian network with uniting of partial problems," *BMC Genomics*, vol. 13, supplement 1, p. S12, 2012.

[4] M. Grzegorczyk and D. Husmeier, "Improvements in the reconstruction of time-varying gene regulatory networks: dynamic programming and regularization by information sharing among genes," *Bioinformatics*, vol. 27, no. 5, Article ID btq711, pp. 693–699, 2011.

[5] N. Xuan Vinh, M. Chetty, R. Coppel, and P. P. Wangikar, "Gene regulatory network modeling via global optimization of high-order dynamic Bayesian network," *BMC Bioinformatics*, vol. 13, article no. 131, 2012.

[6] T. Akutsu, S. Kuhara, O. Maruyama, and S. Miyano, "Identification of genetic networks by strategic gene disruptions and gene overexpressions under a Boolean model," *Theoretical Computer Science*, vol. 298, no. 1, pp. 235–251, 2003.

[7] M. I. Davidich and S. Bornholdt, "Boolean network model predicts cell cycle sequence of fission yeast," *PLoS ONE*, vol. 3, no. 2, Article ID e1672, 2008.

[8] K.-C. Chen, T.-Y. Wang, H.-H. Tseng, C.-Y. F. Huang, and C.-Y. Kao, "A stochastic differential equation model for quantifying transcriptional regulatory network in Saccharomyces cerevisiae," *Bioinformatics*, vol. 21, no. 12, pp. 2883–2890, 2005.

[9] A. Polynikis, S. J. Hogan, and M. di Bernardo, "Comparing different ODE modelling approaches for gene regulatory networks," *Journal of Theoretical Biology*, vol. 261, no. 4, pp. 511–530, 2009.

[10] A. A. Margolin, I. Nemenman, K. Basso et al., "ARACNE: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context," *BMC Bioinformatics*, vol. 7, no. 1, article no. S7, 2006.

[11] J. J. Faith, B. Hayete, J. T. Thaden et al., "Large-scale mapping and validation of Escherichia coli transcriptional regulation from a compendium of expression profiles," *PLoS biology*, vol. 5, no. 1, p. e8, 2007.

[12] G. Michailidis and F. d'Alché-Buc, "Autoregressive models for gene regulatory network inference: sparsity, stability and causality issues," *Mathematical Biosciences*, vol. 246, no. 2, pp. 326–334, 2013.

[13] L. E. Chai, S. K. Loh, S. T. Low, M. S. Mohamad, S. Deris, and Z. Zakaria, "A review on the computational approaches for gene regulatory network construction," *Computers in Biology and Medicine*, vol. 48, no. 1, pp. 55–65, 2014.

[14] E. R. Morrissey, M. A. Juárez, K. J. Denby, and N. J. Burroughs, "Inferring the time-invariant topology of a nonlinear sparse gene regulatory network using fully Bayesian spline autoregression," *Biostatistics*, vol. 12, no. 4, pp. 682–694, 2011.

[15] Y. Ni, F. C. Stingo, and V. Baladandayuthapani, "Bayesian nonlinear model selection for gene regulatory networks," *Biometrics*, vol. 71, no. 3, pp. 585–595, 2015.

[16] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society Series B: Methodological*, vol. 58, no. 1, pp. 267–288, 1996.

[17] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society, Series B: Statistical Methodology*, vol. 68, no. 1, pp. 49–67, 2006.

[18] S. McKay Curtis, S. Banerjee, and S. Ghosal, "Fast Bayesian model assessment for nonparametric additive regression," *Computational Statistics & Data Analysis*, vol. 71, pp. 347–358, 2014.

[19] X.-N. Feng, G.-C. Wang, Y.-F. Wang, and X.-Y. Song, "Structure detection of semiparametric structural equation models with Bayesian adaptive group lasso," *Statistics in Medicine*, vol. 34, no. 9, pp. 1527–1547, 2015.

[20] A. Nair, M. Chetty, and P. P. Wangikar, "Improving gene regulatory network inference using network topology information," *Molecular BioSystems*, vol. 11, no. 9, pp. 2449–2463, 2015.

[21] R. Albert, "Scale-free networks in cell biology," *Journal of Cell Science*, vol. 118, no. 21, pp. 4947–4957, 2005.

[22] X. Xu and M. Ghosh, "Bayesian variable selection and estimation for group lasso," *Bayesian Analysis*, vol. 10, no. 4, pp. 909–936, 2015.

[23] Z. Shang and P. Li, "High-dimensional Bayesian inference in nonparametric additive models," *Electronic Journal of Statistics*, vol. 8, no. 2, pp. 2804–2847, 2014.

[24] D. Marbach, R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky, "Revealing strengths and weaknesses of methods for gene network inference," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 14, pp. 6286–6291, 2010.

[25] D. Marbach, T. Schaffter, C. Mattiussi, and D. Floreano, "Generating realistic in silico gene networks for performance assessment of reverse engineering methods," *Journal of Computational Biology*, vol. 16, no. 2, pp. 229–239, 2009.

[26] R. J. Prill, D. Marbach, J. Saez-Rodriguez et al., "Towards a rigorous assessment of systems biology models: the DREAM3 challenges," *PLoS ONE*, vol. 5, no. 2, Article ID e9202, 2010.

[27] J. Henderson and G. Michailidis, "Network reconstruction using nonparametric additive ODE models," *PLoS ONE*, vol. 9, no. 4, Article ID A1455, 2014.

[28] A. Pinna, N. Soranzo, and A. de la Fuente, "From knockouts to networks: establishing direct cause-effect relationships through graph analysis," *PLoS ONE*, vol. 5, no. 10, Article ID e12912, 2010.

[29] A. Shojaie, A. Jauhiainen, M. Kallitsis, and G. Michailidis, "Inferring regulatory networks by combining perturbation screens and steady state gene expression profiles," *PLoS ONE*, vol. 9, no. 2, Article ID e82393, 2014.

[30] W. C. Young, A. E. Raftery, and K. Y. Yeung, "Fast Bayesian inference for gene regulatory networks using ScanBMA," *BMC Systems Biology*, vol. 8, article 47, 2014.

[31] M. Bansal, V. Belcastro, A. Ambesi-Impiombato, and D. Di Bernardo, "How to infer gene networks from expression profiles," *Molecular Systems Biology*, vol. 3, article no. 78, 2007.

[32] R. Bonneau, D. J. Reiss, P. Shannon et al., "The inferelator: an algorithm for learning parsimonious regulatory networks from systems-biology data sets de novo," *Genome Biology*, vol. 7, no. 5, article R36, 2006.

[33] I. Cantone, L. Marucci, F. Iorio et al., "A yeast synthetic network for in vivo assessment of reverse-engineering and modeling approaches," *Cell*, vol. 137, no. 1, pp. 172–181, 2009.

[34] A. Emad and O. Milenkovic, "CaSPIAN: a causal compressive sensing algorithm for discovering directed interactions in gene networks," *PLoS ONE*, vol. 9, no. 3, Article ID e90781, 2014.

[35] T. Hasegawa, R. Yamaguchi, M. Nagasaki, S. Miyano, and S. Imoto, "Inference of gene regulatory networks incorporating multi-source biological knowledge via a state space model with L1 regularization," *PLoS ONE*, vol. 9, no. 8, Article ID e105942, 2014.

[36] M. Ceccarelli, L. Cerulo, and A. Santone, "De novo reconstruction of gene regulatory networks from time series data, an approach based on formal methods," *Methods*, vol. 69, no. 3, pp. 298–305, 2014.

[37] M. L. Whitfield, "Identification of genes periodically expressed in the human cell cycle and their expression in tumors," *Molecular Biology of the Cell*, vol. 13, no. 6, pp. 1977–2000, 2002.

[38] F. Sambo, B. Di Camillo, and G. Toffolo, "CNET: an algorithm for reverse engineering of causal gene networks," in *Proceedings of the Network Tools and Applications in Biology Workshops (NETTAB '08)*, Varenna, Italy, 2008.

[39] A. C. Lozano, N. Abe, Y. Liu, and S. Rosset, "Grouped graphical Granger modeling for gene expression regulatory networks discovery," *Bioinformatics*, vol. 25, no. 12, pp. i110–i118, 2009.

[40] A. Shojaie and G. Michailidis, "Discovering graphical granger causality using the truncating lasso penalty," *Bioinformatics*, vol. 26, no. 18, pp. i517–i523, 2010.