Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: https://www.tandfonline.com/loi/uasa20

Sparse Additive Ordinary Differential Equations for Dynamic Gene Regulatory Network Modeling

Hulin Wu, Tao Lu, Hongqi Xue & Hua Liang

To cite this article: Hulin Wu, Tao Lu, Hongqi Xue & Hua Liang (2014) Sparse Additive Ordinary Differential Equations for Dynamic Gene Regulatory Network Modeling, Journal of the American Statistical Association, 109:506, 700-716, DOI: 10.1080/01621459.2013.859617

To link to this article: https://doi.org/10.1080/01621459.2013.859617

Accepted author version posted online: 03 Dec 2013.
Published online: 03 Dec 2013.

Sparse Additive Ordinary Differential Equations for Dynamic Gene Regulatory Network Modeling

Hulin WU, Tao LU, Hongqi XUE, and Hua LIANG

The gene regulation network (GRN) is a high-dimensional complex system, which can be represented by various mathematical or statistical models. The ordinary differential equation (ODE) model is one of the popular dynamic GRN models. High-dimensional linear ODE models have been proposed to identify GRNs, but with a limitation of the linear regulation effect assumption. In this article, we propose a sparse additive ODE (SA-ODE) model, coupled with ODE estimation methods and adaptive group least absolute shrinkage and selection operator (LASSO) techniques, to model dynamic GRNs that could flexibly deal with nonlinear regulation effects. The asymptotic properties of the proposed method are established and simulation studies are performed to validate the proposed approach. An application example for identifying the nonlinear dynamic GRN of T-cell activation is used to illustrate the usefulness of the proposed method.

KEY WORDS: Adaptive group LASSO; Dynamic systems; High-dimensional data; Nonparametric additive models; Time course microarray data; Variable selection.

1. INTRODUCTION

The gene regulatory network (GRN) is a complex system associated with biological activities at the cellular level, such as cell growth, division, development, and response to environmental stimulus (Carthew and Sontheimer 2009), which should be modeled in a dynamic way (Hecker et al. 2009). The new high-throughput technologies such as DNA microarray and next generation RNA-Seq enable us to observe the dynamic features of gene expression profiles in a genome scale. In particular, the time course gene expression data collected from these new technologies allow investigators to study gene regulatory networks from a dynamic point of view in more details. Currently, several models have been proposed for GRN construction, such as information theory models (Steuer et al. 2002; Stuart et al. 2003); Boolean networks (Kauffman 1969; Thomas 1973; Bornholdt 2008); Bayesian networks (Heckerman 1996; Imoto et al. 2003; Needham et al. 2007; Werhli and Husmeier 2007); latent variable models (Shojaie and Michailidis 2009); and other regression models (Kim et al. 2009).

In particular, many popular models have been proposed for inferring gene regulatory networks using time course gene expression data. For example, dynamic Boolean networks and probabilistic Boolean networks (Liang, Fuhrman, and Somogyi 1998; Akutsu, Miyano, and Kuhara 2000; Shmulevich et al. 2002; Martin et al. 2007), dynamic Bayesian networks (Murphy and Mian 1999; Friedman et al. 2000; Hartemink et al. 2001; Zou and Conzen 2005; Song, Kolar, and Xing 2009), vector autoregressive and state space models (SSMs; Hirose et al. 2008; Kojima et al. 2009; Shimamura et al. 2009); and differential equation models (Voit 2000; Holter et al. 2001; DeJong 2002; Yeung, Tegner, and Collins 2002). Especially Xing and his associates have recently introduced temporal exponential random graph models and time-varying networks to capture dynamics of networks (Hanneke and Xing 2006; Guo et al. 2007; Kolar et al. 2010). Most of these dynamic network models such as dynamic Bayesian networks and random graph models require extensive computations for posterior inference, which only allow us to deal with small networks. Song, Kolar, and Xing (2009) also introduced a time-varying dynamic Bayesian network to model structurally varying directed graphs. Gupta, Qu, and Ibrahim (2007) proposed a hierarchical hidden Markov regression model for determining gene regulatory networks from gene expression microarray data, which also allows for covariate effects varying between states and gene clusters varying over time. Furthermore, Gupta and Ibrahim (2007) introduced a hierarchical regression mixture model to combine gene clustering and motif discovery in a unified framework, in which a Monte Carlo method was used for simultaneous variable selection (for motifs) and clustering (for time course gene expression data). Shojaie, Basu, and Michailidis (2012) recently proposed an adaptive thresholding estimate under the framework of graphical Granger causality for reconstructing regulatory networks from time course gene expression data.

In this article, we focus on ordinary differential equation (ODE) models for dynamic GRN construction. The ODE approach models the dynamic change of a gene expression (the derivative of the expression) as a function of expression levels of all related genes. So the dynamic feature of the GRN is automatically and naturally quantified. Both positive and negative as well as the feedback effects of gene regulations can be appropriately captured by the ODE model in a systematic way. A

Hulin Wu is Professor (E-mail: *Hulin_Wu@urmc.rochester.edu*) and Hongqi Xue is Research Assistant Professor (E-mail: *Hongqi_Xue@urmc.rochester.edu*), Department of Biostatistics and Computational Biology, School of Medicine and Dentistry, University of Rochester, Rochester, NY 14642. Tao Lu is Assistant Professor, Department of Epidemiology and Biostatistics, State University of New York, Albany, NY 12144 (E-mail: *stat.lu11@gmail.com*). Hua Liang is Professor, Department of Statistics, George Washington University, Washington, DC 20052 (E-mail: *hliang@gwu.edu*). The authors thank the editor, an associate editor, and two referees for their constructive comments and suggestions. This research was partially supported by the NIAID/NIH grants HHSN272201000055C and AI087135 as well as two University of Rochester CTSI pilot awards (UL1RR024160) from the National Center For Research Resources. Liang's research was partially supported by NSF grants DMS-1007167 and DMS-1207444, and by Award Number 11228103, made by National Natural Science Foundation of China.

Color versions of one or more of the figures in the article can be found online at *www.tandfonline.com/r/jasa*.

DOI: 10.1080/01621459.2013.859617

general ODE model for GRNs can be written as

$$\boldsymbol{X}'(t) = \boldsymbol{F}(t, \boldsymbol{X}(t), \boldsymbol{\theta}), \tag{1.1}$$

where $t \in [t_0, T]$ $(0 \le t_0 < T < \infty)$ is time, $\boldsymbol{X}(t) = (X_1(t), \ldots, X_p(t))^T$ is a vector representing the gene expression level of gene $1, \ldots, p$ at time t, and $\boldsymbol{X}'(t)$ is the first-order derivative of $\boldsymbol{X}(t)$. $\boldsymbol{F}$ serves as the link function that quantifies the regulatory effects of regulator genes on the expression change of a target gene, which depends on a vector of parameters $\boldsymbol{\theta}$. In general, $\boldsymbol{F}$ can take any linear or nonlinear functional forms. Many GRN models are based on linear ODEs due to its simplicity. Lu et al. (2011) proposed using the following linear ODEs for dynamic GRN identification and applied the smoothly clipped absolute deviation (SCAD) approach for variable selection,

$$X_k'(t) = \sum_{j=1}^{p} \theta_{kj} X_j(t), \quad k = 1, 2, \ldots, p, \tag{1.2}$$

where parameters $\boldsymbol{\theta} = \{\theta_{kj}\}_{k,j=1,\ldots,p}$ quantify the regulations and interactions among the genes in the network. In practice, however, there is little a priori justification for assuming that the effects of regulatory genes take a linear form. Thus, the linear ODE models may be very restrictive for practical applications. In fact, nonlinear parametric ODE models have been proposed for gene regulatory networks (Weaver, Workman, and Stormo 1999; Sakamoto and Iba 2001; Spieth, Hassis, and Streichert 2006), but the variable selection (network edge identification) problem for high-dimensional nonlinear ODEs has not been addressed. In this article, we intend to extend the high-dimensional linear ODE models in Lu et al. (2011) to a more general additive nonparametric ODE model for modeling high-dimensional nonlinear GRNs:

$$X_k'(t) = \mu_k + \sum_{j=1}^{p} f_{kj}(X_j(t)), \quad k = 1, 2, \ldots, p, \tag{1.3}$$

where μ_k is an intercept term and $f_{kj}(\cdot)$ is a smooth function to quantify the nonlinear relationship among related genes in the GRN. Based on the sparseness principle of gene regulatory networks and other biological systems, we usually assume that the number of significant nonlinear effects, $f_{kj}(\cdot)$, is small for each of the p variables (genes), X_k, although the total number of variables (genes), p, in the network may be large. Thus, we refer to model (1.3) as the sparse additive ODE (SA-ODE) model. Also we assume that the measurements of gene expression for the kth gene are obtained at multiple time points, $t_i, i = 1, \ldots, n$, that is,

$$Y_k(t_i) = X_k(t_i) + \varepsilon_k(t_i), \quad k = 1, 2, \ldots, p, \tag{1.4}$$

where the measurement errors $\varepsilon_k(t_i)$ $(i = 1, \ldots, n)$ are assumed to be iid with mean zero and variance σ_k^2 $(0 < \sigma_k^2 < \infty)$. The challenging question is how to perform model selection for the nonparametric SA-ODE model (1.3) under the assumption of sparsity constraints on the index set $\{j : f_{kj}(\cdot) \ne 0\}$ of functions $f_{kj}(\cdot)$ that are not identically zero.

There exist several classes of parameter estimation methods for ODE models, which include the nonlinear least-square (NLS) method (Hemker 1972; Bard 1974; Li, Osborne, and Pravan 2005; Xue, Miao, and Wu 2010), the two-stage smoothing-based estimation method (Varah 1982; Brunel 2008; Chen and Wu 2008a,b; Liang and Wu 2008; Wu, Xue, and Kumar 2012), the principal differential analysis (PDA) and its extensions (Ramsay 1996; Heckman and Ramsay 2000; Ramsay and Silverman 2005; Poyton et al. 2006; Ramsay et al. 2007; Varziri et al. 2008; Qi and Zhao 2010), and the Bayesian approaches (Putter et al. 2002; Huang, Liu, and Wu 2006; Donnet and Samson 2007). Among these methods, we are more interested in the two-stage smoothing-based estimation method, where in the first stage, a nonparametric smoothing approach is used to obtain the estimates of both the state variables and their derivatives from the observed data, and then in the second stage, these estimated functions are plugged into the ODEs to estimate the unknown parameters using a formulated pseudo-regression model. In particular, the two-stage smoothing-based estimation method avoids numerically solving the differential equations directly and does not need the initial or boundary conditions of the state variables. This method also decouples the high-dimensional ODEs to allow us to perform variable selection and parameter estimation for one equation at a time (Voit and Almeida 2004; Jia, Stephanopoulos, and Gunawan 2011; Lu et al. 2011). These good features of the two-stage smoothing-based estimation method in addition to its computational efficiency greatly outweigh its disadvantage in a small loss of estimation accuracy in dealing with high-dimensional nonlinear ODE models (Lu et al. 2011).

In the past two decades, there has been much work on penalization methods for variable selection and parameter estimation for high-dimensional data, including the bridge estimator (Frank and Friedman 1993); the least absolute shrinkage and selection operator (LASSO) (Tibshirani 1996) and its extensions, such as the adaptive LASSO (Zou 2006), the group LASSO (Yuan and Lin 2006), and the adaptive group LASSO (Wang and Leng 2008); the SCAD penalty (Fan and Li 2001); and the elastic net (Zou and Hastie 2006) among others. Recently, the LASSO approach has been applied to high-dimensional nonparametric sparse additive (regression) models to perform variable selection and parameter estimation simultaneously (Meier, van de Geer, and Buhlmann 2009; Ravikumar et al. 2009; Huang, Horowitz, and Wei 2010; Cantoni, Flemming, and Ronchetti 2011). In this article, we propose to couple the ideas of the two-stage smoothing-based estimation method and the high-dimensional variable selection techniques to perform variable selection and nonparametric function estimation for the proposed SA-ODE model (1.3). However, this is not just to trivially combine the two ideas, instead we have to tackle several critical challenges that include: (i) the two-stage smoothing-based estimation method allows us to convert our nonparametric SA-ODE model into a pseudo-sparse additive regression model where the covariates and the response variables are derived from nonparametric smoothing estimates of the ODE state variables and their derivatives (instead of direct observed data); (ii) the resulting errors in the pseudo-sparse additive (PSA) regression model are not iid, but dependent; and (iii) the number of covariates, p, in the SA-ODE models is usually large (maybe greatly larger than the sample size). Thus, it is not trivial to establish the theoretical properties for the proposed method. To the best of our knowledge, this is the first attempt to propose a variable selection method for a high-dimensional nonparametric ODE model, and it is also the first time to establish the theoretical results for the

penalized estimators for ODE models under the "large p, small n" setting.

The remainder of this article is organized as follows. In Section 2, we propose a five-step variable selection procedure for the SA-ODE model (1.3). In Section 3, the theoretical properties of the proposed method are established in the "large p, small n" setting. In Section 4, we present some simulation results to validate the proposed method. To illustrate the usefulness of the proposed methods, we apply the proposed method to identify a dynamic gene regulatory network based on time course gene expression data from a T-cell activation study in Section 5. We conclude the article with some remarks in Section 6. The detailed technical proofs are given in the Appendix.

2. PSEUDO-SPARSE ADDITIVE MODEL AND VARIABLE SELECTION

In this section, we propose a five-step variable selection procedure for the SA-ODE model (1.3). In Step I, we use one of the nonparametric smoothing approaches to estimate both the ODE state variables and their derivatives based on the measurement model (1.4). In Step II, we use spline functions to approximate each of the nonparametric additive components in the SA-ODE model (1.3), and then substitute the estimated state variables and their derivatives from Step I into the SA-ODE model (1.3) to form a "pseudo" sparse additive model. In Step III, we apply the group LASSO approach to obtain an initial estimator and reduce the dimension of the problem. In Step IV, we apply the adaptive group LASSO approach (Wang and Leng 2008; Huang, Horowitz, and Wei 2010) for component selection, which combines ideas from the adaptive LASSO (Zou 2006) and the group LASSO (Yuan and Lin 2006). In Step II, we usually use a larger number of basis functions to approximate the nonparametric functions and some of these basis functions may not be necessary. In Step V, we propose to use the regular LASSO again to the selected model from Step IV to further shrink some of the coefficients of B-spline basis to zero so that we can obtain a more parsimonious model at the end.

2.1 Step I. Nonparametric Smoothing

First, we apply one of nonparametric smoothing approaches such as smoothing splines, regression splines, penalized splines or local polynomial to estimate the state variables and their derivatives, $X_k(t)$ and $X_k'(t)$ based on model (1.4). In this article, we adopt the penalized splines (Ruppert, Wand, and Carroll 2003; Li and Ruppert 2008; Claeskens, Krivobokova, and Opsomer 2009; Wu, Xue, and Kumar 2012) to obtain the estimates, $\hat{X}_k(t)$ and $\hat{X}_k'(t)$, $k = 1, \ldots, p$. That is, approximate $X_k(t)$ one by one for $1 \le k \le p$ by $X_k(t) \approx \sum_{j=-\nu}^{K_k} \delta_{k,j} N_{k,j,\nu+1}(t) = \mathbf{N}_{k,\nu+1}^T(t)\boldsymbol{\delta}_k$, where $\boldsymbol{\delta}_k = (\delta_{k,-\nu}, \ldots, \delta_{k,K_k})^T$ is the unknown coefficient vector to be estimated from the data, and $\mathbf{N}_{k,\nu+1}(t) = \{N_{k,-\nu,\nu+1}(t), \ldots, N_{k,K_k,\nu+1}(t)\}^T$ is the B-spline basis function vector of degree ν and dimension $K_k + \nu + 1$ at a sequence of knots $t_0 = \tau_{k,-\nu} = \tau_{k,-\nu+1} = \cdots = \tau_{k,-1} = \tau_{k,0} < \tau_{k,1} < \cdots < \tau_{k,K_k} < \tau_{k,K_k+1} = \tau_{k,K_k+2} = \cdots = \tau_{k,K_k+\nu+1} = T$ on $[t_0, T]$ (Schumaker 1981). Define $n \times (K_k + \nu + 1)$ matrix $\mathbf{N}_k = \{\mathbf{N}_{k,\nu+1}(t_1), \ldots, \mathbf{N}_{k,\nu+1}(t_n)\}^T$, $\mathbf{Y}_k = (Y_k(t_1), \ldots, Y_k(t_n))^T$ and let $\mathbf{V}_k = \int_{t_0}^T [\mathbf{N}_{k,\nu+1}''(t)][\mathbf{N}_{k,\nu+1}''(t)]^T dt$. The penalized spline (P-spline) objective function contains a penalized sum of squared differences with a penalized term by the integrated squared second-order derivative of the spline function as

$$L_k(\boldsymbol{\delta}_k; \lambda_k) = (\mathbf{Y}_k - \mathbf{N}_k\boldsymbol{\delta}_k)^T(\mathbf{Y}_k - \mathbf{N}_k\boldsymbol{\delta}_k) + \lambda_k\boldsymbol{\delta}_k^T\mathbf{V}_k\boldsymbol{\delta}_k. \quad (2.1)$$

The minimizer of (2.1) takes the form $\hat{\boldsymbol{\delta}}_k = (\mathbf{N}_k^T\mathbf{N}_k + \lambda_k\mathbf{V}_k)^{-1}\mathbf{N}_k^T\mathbf{Y}_k$. Then we have

$$\hat{X}_k(t) = \mathbf{N}_{k,\nu+1}^T(t)\hat{\boldsymbol{\delta}}_k, \quad \hat{X}_k'(t) = [\mathbf{N}_{k,\nu+1}'(t)]^T\hat{\boldsymbol{\delta}}_k. \quad (2.2)$$

From de Boor (2001), the derivatives of spline functions can be simply expressed in terms of lower order spline functions, then we can obtain the explicit expressions of $\mathbf{N}_{k,\nu+1}'(t)$ and $\mathbf{N}_{k,\nu+1}''(t)$. To determine λ_k, we use the standard generalized cross-validation (GCV) method (Craven and Wahba 1979). Note that, if the longitudinally replicate data are available (see our application example in Section 5), the nonparametric mixed-effects smoothing methods can be used in this step to obtain better smoothing results (Wu and Zhang 2006; Lu et al. 2011).

2.2 Step II. Pseudo-Sparse Additive Models

In this step, we propose a method to identify significant functions in model (1.3) using a high-dimensional variable selection technique, the group LASSO. First, following the idea similar to Varah (1982), Brunel (2008), Chen and Wu (2008a), Chen and Wu (2008b), Liang and Wu (2008), and Wu, Xue, and Kumar (2012), we substitute the estimated state variables $\hat{X}_k(t)$ and their derivatives $\hat{X}_k'(t)$ into ODE model (1.3) to form the following PSA model:

$$H_{ki} = \mu_k + \sum_{j=1}^{p} f_{kj}(\hat{X}_{ji}) + \Upsilon_{ki}, \quad k = 1, 2, \ldots, p, \quad i = 1, 2, \ldots, n, \quad (2.3)$$

where $H_{ki} = \hat{X}_k'(t_i)$ and $\hat{X}_{ji} = \hat{X}_j(t_i)$, Υ_{ki} is the sum of measurement errors and estimation errors of $\hat{X}_k'(t)$ and $\hat{X}_k(t)$ from Step I. In model (2.3), the response variables and the covariates are derived from the nonparametric smoothing estimates of the state variables and their derivatives, respectively. Moreover, the resulting error terms Υ_{ki} are not iid, but dependent. Thus, this is not a standard sparse additive regression model studied in the literature (Meier, van de Geer, and Buhlmann 2009; Ravikumar et al. 2009; Huang, Horowitz, and Wei 2010; Cantoni, Flemming, and Ronchetti 2011). That is why we call it as a "pseudo"-sparse additive (PSA) model. Since $\hat{X}_k(t)$ and $\hat{X}_k'(t)$ in the above model are estimated continuously at any time point t from Step I, we may augment more time points than the original observation times for the next step analysis. In fact, other investigators (D'Haeseleer et al. 1999; Wessels, van Someren, and Reinders 2001; Bansal, Della Gatta, and di Bernardo 2006) have used this data augmentation strategy for ODE parameter estimation. We adopt a similar idea here.

Remark 1. Decoupled property. Note that the substitution approach in model (2.3) allows us to decouple the p-dimensional ODE model into p one-dimensional ODEs independently (Voit and Almeida 2004; Jia, Stephanopoulos, and Gunawan 2011; Lu et al. 2011), so that we can deal with the variable selection problem for these ODEs one by one separately. This is a unique feature of the two-stage smoothing-based estimation method for ODEs (Varah 1982; Brunel 2008; Chen and Wu 2008a,b; Liang and Wu 2008; Wu, Xue, and Kumar 2012).

We adopt a similar idea from Huang, Horowitz, and Wei (2010) and apply truncated series expansions with B-spline bases to approximate the additive components in model (2.3). Let $t_0 = \xi_0 < \xi_1 < \cdots < \xi_{K_n} < \xi_{K_n+1} = T$ be a partition of the interval $[t_0, T]$, where $K_n = O(n^{\varpi})$ $(0 < \varpi < 0.5)$ is a positive integer such that $\max_{0\le m\le K_n} |\xi_{m+1} - \xi_m| = O(n^{-\varpi})$. Let $\mathcal{S}_n$ be the space of polynomial splines on $[t_0, T]$ of degree $l \ge 1$ consisting of functions s satisfying: (i) s is a polynomial of degree l on the subintervals $I_m = [\xi_m, \xi_{m+1})$, $m = 0, \ldots, K_n - 1$, and $I_{K_n} = [\xi_{K_n}, \xi_{K_n+1}]$; (ii) for $l \ge 2$ and $0 \le l' \le l-2$, s is l' times continuously differentiable on $[t_0, T]$. Then there exists a normalized B-spline basis $\{\phi_m, 1 \le m \le m_n\}$ on $[t_0, T]$ for $\mathcal{S}_n$, where $m_n \equiv K_n + l$ such that, for any $f_{kj}^* \in \mathcal{S}_n$, it can be expressed as

$$f_{kj}^*(x) = \sum_{m=1}^{m_n} \beta_{kjm}\phi_m(x), \quad k, j = 1, 2, \ldots, p, \tag{2.4}$$

where β_{kjm} are spline coefficients. Here, we propose to conservatively choose the number of basis functions, m_n, as large as possible (more than enough or undersmoothing). Note that it is computationally prohibited to select different m_n's for different functions $f_{kj}(\cdot)$ when p is large. To deal with this problem, we will reapply the LASSO approach to shrink unnecessary basis coefficients into zero in Step V. Replacing f_{kj} by its B-spline approximation in (2.4), model (2.3) can be expressed as

$$H_{ki} = \mu_k + \sum_{j=1}^{p}\sum_{m=1}^{m_n} \beta_{kjm}\phi_m(\hat{X}_{ji}) + \Upsilon_{ki}^*, \quad k = 1, \ldots, p, \quad i = 1, \ldots, n, \tag{2.5}$$

where Υ_{ki}^* is the sum of Υ_{ki} and the approximation errors of the additive regression functions by splines. Let $\boldsymbol{\beta}_{kj} = (\beta_{kj1}, \ldots, \beta_{kjm_n})^T$ $(k, j = 1, \ldots, p)$ and $\boldsymbol{\beta}_k = (\boldsymbol{\beta}_{k1}^T, \ldots, \boldsymbol{\beta}_{kp}^T)^T$. Then we have p groups of parameters and our purpose is to select nonzero groups, that is, nonzero $\boldsymbol{\beta}_{kj}$, $k, j = 1, \ldots, p$.

2.3 Step III. Group LASSO

For model (1.3), we need a constraint to deal with unidentifiability problem, that is, $Ef_{kj}(X_j(t)) = 0$ $(j = 1, \ldots, p)$. Thus, for model (2.5), we impose the constraints $\sum_{i=1}^n \sum_{m=1}^{m_n} \beta_{kjm}\phi_m(\hat{X}_{ji}) = 0$, $1 \le j \le p$. We may also use the centralization of the response and the basis functions to remove the restrictions. Let $\bar{\phi}_{jm} = \frac{1}{n}\sum_{i=1}^n \phi_m(\hat{X}_{ji})$ and $\psi_m(x) \equiv \psi_{jm}(x) = \phi_m(x) - \bar{\phi}_{jm}$. Write $Z_{ij} = \{\psi_1(\hat{X}_{ji}), \ldots, \psi_{m_n}(\hat{X}_{ji})\}^T$, $Z_j = (Z_{1j}, \ldots, Z_{nj})^T$, and $\mathbf{Z} = (Z_1, \ldots, Z_p)$. Let $\bar{H}_k = \frac{1}{n}\sum_{i=1}^n H_{ki}$ and $\mathbf{H}_k = (H_{k1} - \bar{H}_k, \ldots, H_{kn} - \bar{H}_k)^T$. Similar to Brunel (2008) and Wu, Xue, and Kumar (2012), a weight function with boundary restrictions should be imposed to achieve a better convergence rate for parameter estimation. Let $\mathbf{D}_{k1} = \mathrm{diag}\{d_{k1}(t_1), \ldots, d_{k1}(t_n)\}$, $\mathbf{D}_{k2} = \mathrm{diag}\{d_{k2}(t_1), \ldots, d_{k2}(t_n)\}$, and $\mathbf{D}_{k3} = \mathrm{diag}\{d_{k3}(t_1), \ldots, d_{k3}(t_n)\}$, where $d_{k1}(t)$, $d_{k2}(t)$, and $d_{k3}(t)$ are prescribed nonnegative weight functions on $[t_0, T]$ with boundary conditions $d_{k1}(t_0) = d_{k1}(T) = 0$, $d_{k2}(t_0) = d_{k2}(T) = 0$ and $d_{k3}(t_0) = d_{k3}(T) = 0$. More discussions on how to select the weight function can be found in Brunel (2008) and Wu, Xue, and Kumar (2012). With these notations, we can obtain the group LASSO estimator $\tilde{\boldsymbol{\beta}}_k$ by minimizing the following penalized weighted least-square criterion:

$$L_{k1}(\boldsymbol{\beta}_k; \lambda_{k1}) = (\mathbf{H}_k - \mathbf{Z}\boldsymbol{\beta}_k)^T \mathbf{D}_{k1} (\mathbf{H}_k - \mathbf{Z}\boldsymbol{\beta}_k) + \lambda_{k1}\sum_{j=1}^{p} \|\boldsymbol{\beta}_{kj}\|_2, \tag{2.6}$$

where λ_{k1} is a penalty parameter, which can be determined by Bayesian information criterion (BIC) or extended Bayes information criterion (EBIC; Chen and Chen 2008). Here, we have dropped μ_k in the arguments of L_{k1} with the centering $\hat{\mu}_k = \bar{H}_k$. Based on the group LASSO estimator $\tilde{\boldsymbol{\beta}}_k$, we can also obtain the estimates of the nonparametric functions, $\tilde{f}_{kj}(x) = \sum_{m=1}^{m_n} \tilde{\beta}_{kjm}\psi_m(x)$, $1 \le j \le p$.

2.4 Step IV. Adaptive Group LASSO

The above group LASSO penalty treats coefficients from each group equally, which is not optimal. To allow different amounts of shrinkage for different coefficients, an adaptive group LASSO is necessary. In this step, we perform the adaptive group LASSO based on the results from Step III by setting $w_{kj} = \|\tilde{\boldsymbol{\beta}}_{kj}\|_2^{-1}$ if $\|\tilde{\boldsymbol{\beta}}_{kj}\|_2 > 0$, otherwise $w_{kj} = \infty$. Then we obtain the adaptive group LASSO estimator $\hat{\boldsymbol{\beta}}_k$ by minimizing the penalized weighted least-square criterion,

$$L_{k2}(\boldsymbol{\beta}_k; \lambda_{k2}) = (\mathbf{H}_k - \mathbf{Z}\boldsymbol{\beta}_k)^T \mathbf{D}_{k2} (\mathbf{H}_k - \mathbf{Z}\boldsymbol{\beta}_k) + \lambda_{k2}\sum_{j=1}^{p} w_{kj}\|\boldsymbol{\beta}_{kj}\|_2 \tag{2.7}$$

with a penalty parameter λ_{k2}, which can also be determined by BIC or EBIC (Chen and Chen 2008). Then we obtain the adaptive group LASSO estimates of μ_k and f_{kj}, $\hat{\mu}_k = \bar{H}_k \equiv \frac{1}{n}\sum_{i=1}^n H_{ki}$ and $\hat{f}_{kj}(x) = \sum_{m=1}^{m_n} \hat{\beta}_{kjm}\psi_m(x)$, $1 \le j \le p$.

2.5 Step V. Regular LASSO for Shrinking Basis Coefficients

In Step II, we approximate each of the nonparametric functions in the PSA model intentionally using a larger number of basis functions (undersmoothing). Thus, some of these basis functions may not be necessary. In this step, we reapply the regular LASSO or adaptive LASSO to the final model selected from the adaptive group LASSO in Step IV to shrink the coefficients of unnecessary basis functions into zero, so that we can obtain a final parsimonious model. The minimization criterion for the adaptive LASSO is

$$L_{k3}(\boldsymbol{\beta}_k; \lambda_{k3}) = \left(\mathbf{H}_k^{(s)} - \mathbf{Z}^{(s)}\boldsymbol{\beta}_k\right)^T \mathbf{D}_{k3}\left(\mathbf{H}_k^{(s)} - \mathbf{Z}^{(s)}\boldsymbol{\beta}_k\right) + \lambda_{k3}\sum_{j=1}^{s}\sum_{m=1}^{m_n} w_{kjm}\left|\beta_{kjm}^{(s)}\right|, \tag{2.8}$$

where the superscript "(s)" stands for the corresponding quantities for groups picked up from Step IV and s is the total number of groups. The weight w_{kjm} is set as $|\hat{\beta}_{kjm}^{(s)}|^{-1}$ if $|\hat{\beta}_{kjm}^{(s)}| > 0$ and $w_{kjm} = \infty$, otherwise.

3. THEORETICAL RESULTS

In this section, we establish the asymptotic properties of the proposed group LASSO estimator in Step III and the adaptive

group LASSO estimator in Step IV for the PSA model derived from a set of ODEs in the last section. This is challenging since we need to integrate the asymptotic results from the two-step smoothing-based estimation method for ODE models (Varah 1982; Brunel 2008; Chen and Wu 2008a,b; Liang and Wu 2008; Wu, Xue, and Kumar 2012) and the sparse additive models (Meier, van de Geer, and Buhlmann 2009; Ravikumar et al. 2009; Huang, Horowitz, and Wei 2010; Cantoni, Flemming, and Ronchetti 2011) together.

Let r be a nonnegative integer and $\zeta \in (0, 1]$ such that $\varrho = r + \zeta > 0.5$. Let $\mathcal{F}$ be the collection of functions f on $[t_0, T]$ whose rth derivative, $f^{(r)}$ exists and satisfies the Lipschitz condition of order ζ: $|f^{(r)}(s) - f^{(r)}(t)| \le C|s - t|^{\zeta}$ for $s, t \in [t_0, T]$ with a general positive constant C. In model (1.3), without loss of generality, suppose that the first q components are nonzero, that is, $f_{kj}(x) \neq 0$, $1 \le j \le q$, but $f_{kj}(x) \equiv 0$, $q + 1 \le j \le p$. Let $A_1 = \{1, \ldots, q\}$, $A_0 = \{q + 1, \ldots, p\}$, $\tilde{A}_1 = \{j :, \|\tilde{\beta}_{kj}\|_2 \neq 0, 1 \le j \le p\}$, and $\tilde{A}_2 = A_1 \cup \tilde{A}_1$. Let $|A|$ be the cardinality for any index set A. Define $\|f\|_2 = [\int_a^b f^2(x)dx]^{1/2}$ for any function $f(x)$ at $x \in [a, b]$, whenever the integral exists. For $1 \le k \le p$, we make the following assumptions:

Assumption A.

(A1) For $\kappa = \max_{0 \le j \le K_k}(\tau_{k,j+1} - \tau_{k,j})$, there exists a constant $M > 0$ such that $\frac{\kappa}{\min_{0 \le j \le K_k}(\tau_{k,j+1} - \tau_{k,j})} \le M$ and $\max_{0 \le j \le K_k} |\frac{\tau_{k,j+1} - \tau_{k,j}}{\tau_{k,j} - \tau_{k,j-1}} - 1| = o(1)$.

(A2) $K_k \sim cn^{\vartheta}$ with $1/(2\nu + 3) \le \vartheta < 1$ and $\lambda_k = O(n^{\pi})$ with $\pi \le \nu/(2\nu + 3)$.

(A3) $X_k(t) \in C^{\nu+1}[t_0, T]$ with $\nu \ge 2$.

(A4) $K^* = (K_k + \nu - 1)(\lambda_k \tilde{c}_1)^{1/4} n^{-1/4} < 1$ for some constant $\tilde{c}_1$.

(A5) Random design points $t_1, \ldots, t_n$ are iid with a cumulative distribution function $Q(t)$ and a positive and continuous derivative density $\rho(t)$. Moreover, $\rho(t)$ is bounded away from 0 and $+\infty$ and has a bounded and continuous first-order derivative.

Assumption B.

(B1) $l + 1 \ge \varrho$.

(B2) The number of nonzero components q is fixed and there is a constant $c_f > 0$ such that $\min_{1 \le j \le q} \|f_{kj}\|_2 \ge c_f$ for those nonzero f_{kj}'s.

(B3) The random variables $\varepsilon_k(t_i)$ $(i = 1, \ldots, n)$ are iid with $\mathrm{E}[\varepsilon_k(t_i)] = 0$ and $\mathrm{var}[\varepsilon_k(t_i)] = \sigma_k^2$ $(0 < \sigma_k^2 < \infty)$. Furthermore, their tail probabilities satisfy $\mathrm{P}\{|\varepsilon_k(t_i)| > x\} \le K \exp(-Cx^2)$, $i = 1, \ldots, n$, for all $x \ge 0$ and for some constants C and K.

(B4) $\mathrm{E}[f_{kj}(X_j(t))] = 0$ and $f_{kj} \in \mathcal{F}$ for $k, j = 1, \ldots, p$.

(B5) $\nu \ge 3\varrho$.

(B6) Both the weight functions $d_{k1}(.)$ and $d_{k2}(.)$ are bounded and nonnegative on $[t_0, T]$ with $d_{k1}(t_0) = d_{k1}(T) = 0$ and $d_{k2}(t_0) = d_{k2}(T) = 0$. For simplicity, assume that $\|d_{k1}\|_{\infty} \le 1$ and $\|d_{k2}\|_{\infty} \le 1$. Moreover, both $d_{k1}(t)$ and $d_{k2}(t)$ have bounded and continuous first-order derivatives.

Note that Assumption A is required to derive the local properties for penalized splines estimates, which were also used by Claeskens, Krivobokova, and Opsomer (2009) and Wu, Xue, and Kumar (2012). Assumption B1 is required for the nonparametric smoothing approaches (Huang 2003; Xue, Miao, and Wu 2010). Assumptions B2, B3, and B4 are standard conditions for nonparametric additive models (Huang, Horowitz, and Wei 2010). Assumption B5 means that the smoothness degrees for the state variables $X_k(.)$ are higher than those for the additive functions $f_{kj}(.)$, which is required to control the error of the first-step nonparametric smoothing to achieve the same order of the error rate in the PSA model in Step II. Assumption B6 is for dealing with the boundary effect for the derivative estimation, so that the proposed adaptive group LASSO estimator can achieve the optimal nonparametric convergence rate.

Theorem 1. Suppose that Assumptions A and B hold and $\lambda_{k1} \ge C\sqrt{n \log(pm_n)}$ for a sufficiently large constant C. Then we have

(i) With probability approaching to 1, $|\tilde{A}_1| \le M_1|A_1| = M_1 q$ for a finite constant $M_1 > 1$.

(ii) If $m_n^2 \log(pm_n)/n \to 0$ and $(\lambda_{k1}^2 m_n^2)/n^2 \to 0$ as $n \to \infty$, then all the nonzero $\beta_{k,j}$, $1 \le j \le q$, are selected with probability approaching to 1.

(iii) $\sum_{j=1}^p \|\tilde{\beta}_{kj} - \beta_{kj}\|_2^2 = O_p(\frac{m_n^2 \log(pm_n)}{n}) + O_p(\frac{m_n}{n}) + O(\frac{1}{m_n^{2\varrho-1}}) + O(\frac{m_n^2 \lambda_{k1}^2}{n^2})$.

Theorem 2. Suppose that Assumptions A and B hold and that $\lambda_{k1} \ge C\sqrt{n \log(pm_n)}$ for a sufficiently large constant C. Then we have the following results:

(i) Let $\tilde{A}_f = \{j : \|\tilde{f}_{k,j}\|_2 > 0, 1 \le j \le p\}$. There is a constant $M_1 > 1$ such that, with probability approaching to 1, $|\tilde{A}_f| \le M_1 q$.

(ii) If $m_n \log(pm_n))/n \to 0$ and $(\lambda_{k1}^2 m_n)/n^2 \to 0$ as $n \to \infty$, then all the nonzero additive components f_{kj}, $1 \le j \le q$, are selected with probability approaching to 1.

(iii) $\|\widetilde{f}_{kj} - f_{kj}\|_2^2 = O_p(\frac{m_n \log(pm_n)}{n}) + O_p(\frac{1}{n}) + O(\frac{1}{m_n^{2\varrho}}) + O(\frac{m_n \lambda_{k1}^2}{n^2})$, $j \in \tilde{A}_2$.

Corollary 1. Suppose that Assumptions A and B hold. If $\lambda_{k1} \asymp \sqrt{n \log(pm_n)}$ and $m_n \asymp n^{1/(2\varrho+1)}$, we have

(i) If $n^{-2\varrho/(2\varrho+1)} \log(p) \to 0$ as $n \to \infty$, then with probability approaching to 1, all the nonzero components f_{kj}, $1 \le j \le q$, are selected and the number of selected components is no more than $M_1 q$.

(ii) $\|\widetilde{f}_{kj} - f_{kj}\|_2^2 = O_p(n^{-2\varrho/(2\varrho+1)} \log(pm_n))$, $j \in \tilde{A}_2$.

Theorem 3. Suppose Assumptions A and B hold. If $\lambda_{k1} \asymp \sqrt{n \log(pm_n)}$, $m_n \asymp n^{1/(2\varrho+1)}$, and $\lambda_{k2} \le O(n^{1/2})$ and satisfies $\frac{\lambda_{k2}}{n^{(8\varrho+3)/(8\varrho+4)}} = o(1)$ and $\frac{n^{1/(4\varrho+2)} \log^{1/2}(pm_n)}{\lambda_{k2}} = o(1)$, then $\mathrm{P}(\|\hat{f}_{kj} - f_{kj}\|_2 > 0, j \in A_1 \text{ and } \|\hat{f}_{kj}\|_2 = 0, j \in A_0) \to 1$, that is, the adaptive group LASSO consistently selects the nonzero components. In addition, $\sum_{j=1}^q \|\hat{f}_{kj} - f_{kj}\|_2^2 = O_p(n^{-2\varrho/(2\varrho+1)})$.

All detailed proofs of Theorems 1–3, Corollary 1 are given in the Appendix.

Remark 2. We note that, for the λ_{k1}, λ_{k2}, and m_n given in Theorem 3, the number of zero components can be as large as $\exp(o(n^{2\varrho/(2\varrho+1)}))$, which is much larger than n. Thus, under the conditions of the theorems, the proposed adaptive group LASSO estimator is selection consistent and achieves the optimal rate convergence even when p is much larger than n.

Remark 3. All of these theoretical results (Theorems 1–3 and Corollary 1) can be extended to the case of fixed design points $t_1, \ldots, t_n$, in which case we can assume that there exists a distribution function $Q(t)$ with a positive and continuous derivative density $\rho(t)$ such that for the empirical distribution $Q_n(t)$, $\sup_{t\in[t_0,T]} |Q_n(t) - Q(t)| = o(K_k^{-1})$. This assumption was also adopted by Zhou, Shen, and Wolfe (1998) and Zhou and Wolfe (2000).

For the PSA models derived from a set of nonlinear/nonparametric ODEs, we have achieved similar theoretical results (Theorems 1–3 and Corollary 1) to those obtained by Huang, Horowitz, and Wei (2010) for nonparametric additive regression models. To prove these results, we develop an important lemma (Lemma A.1 in the Appendix) on the convergence rate of the projection involving the weighted residuals of the derivative estimate of the state variables in the ODEs. In addition, we used the weight functions, $\mathbf{D}_{k1}$ and $\mathbf{D}_{k2}$, in the group LASSO objective function (2.6) and the adaptive group LASSO objective function (2.7), respectively. Under a parametric ODE model setting, Wu, Xue, and Kumar (2012) used the weighted function with similar boundary conditions for estimation of constant coefficients in ODE models. They found that the parametric estimator has different convergence rates for different boundary conditions of the weight function, that is, if the weight function is zero at both boundaries, the standard root n rate for the constant ODE parameter estimates can be achieved; otherwise, only an optimal nonparametric convergence rate can be reached. In the situation of the PSA models, we have achieved the optimal nonparametric convergence rate for the adaptive group LASSO estimator (Theorem 3) of the nonparametric functions in the SA-ODE model under the zero boundary assumption for the weight function. If the zero boundary assumption of the weight function does not hold, a lower than the optimal nonparametric convergence rate is expected although we did not provide a detailed proof for this claim. Here, we adopt similar ideas in the proofs of Theorems 1–3 and Corollary 1 to those in Huang, Horowitz, and Wei (2010), but we have to tackle some challenges in the detailed proofs due to the difference between the PSA model and the sparse additive regression model in Huang, Horowitz, and Wei (2010).

4. SIMULATION STUDIES

In this section, we design simulation experiments to validate the proposed variable selection method for the SA-ODE model. We consider a true SA-ODE model with eight coupled ODEs as a dynamic network (a higher dimensional SA-ODE model is prohibited for simulation studies due to the limitation of current computational power):

$$X_1' = f_{1,1}(X_1) + f_{1,2}(X_2), \quad X_2' = f_{2,1}(X_1) + f_{2,2}(X_3),$$
$$X_3' = f_{3,1}(X_2) + f_{3,2}(X_5), \tag{4.1}$$
$$X_4' = f_{4,1}(X_8), \quad X_5' = f_{5,1}(X_4) + f_{5,2}(X_6) + f_{5,3}(X_8),$$
$$X_6' = f_{6,1}(X_5), \tag{4.2}$$
$$X_7' = f_{7,1}(X_6) + f_{7,2}(X_8), \quad X_8' = f_{8,1}(X_4) + f_{8,2}(X_6) \tag{4.3}$$

and we set

$$f_{1,1}(x) = -\sin(2x), \quad f_{1,2}(x) = 0.1x^2 - 1, \quad f_{2,1}(x) = x,$$
$$f_{2,2}(x) = e^{-x}, \tag{4.4}$$
$$f_{3,1}(x) = x^{1.5} - 2, \quad f_{3,2}(x) = 10\sin(2x)/(10 - \cos(2x)),$$
$$f_{4,1}(x) = 0.1x^2 - 0.5, \tag{4.5}$$
$$f_{5,1}(x) = 10\cos(2x), \quad f_{5,2}(x) = 10\cos(2x) + 10x \times \sin(2x),$$
$$f_{5,3}(x) = x^{1/3} + x^{1/5}, \tag{4.6}$$
$$f_{6,1}(x) = -x \times \cos(2x), \quad f_{7,1}(x) = 0.1x + x^2,$$
$$f_{7,2}(x) = -0.01x^3, \tag{4.7}$$
$$f_{8,1}(x) = \sqrt{|x|} \times \cos(2x), \quad f_{8,2}(x) = 0.01x^3 + 0.01x^2. \tag{4.8}$$

We assume the measurement model as

$$Y_k(t_{ij}) = X_k(t_{ij}) + \varepsilon_k(t_{ij}), \quad i = 1, 2, \ldots, n, \quad j = 1, 2, \ldots, n_k,$$
$$k = 1, 2, \ldots, 8. \tag{4.9}$$

Note that, to mimic the real data example in the next section, we also allow to have replicates of measurements, $j = 1, 2, \ldots, n_k$ at each time point for the kth equation, and the number of replicates may be different for different equations. In our simulation studies, we assume the numbers of replicates for the eight equations as $(n_1, n_2, n_3, n_4, n_5, n_6, n_7, n_8) = (50, 60, 70, 40, 50, 30, 45, 35)$, respectively; and the number of measurement time points is taken as $n = 15$ and 150 (equally spaced), respectively. If the replicates are repeated measures or longitudinal data as in our real data example in next section, we can use the nonparametric mixed-effects smoothing approach (Wu and Zhang 2006) for the first-step nonparametric smoothing, which is more efficient (Lu et al. 2011). In our simulation studies, we also consider the case of longitudinal replicates. We first generated the mean trajectory of the observational data, $\bar{X}_k(t)$, by numerically solving the ODE models (4.1)–(4.3) using the initial values of the state variables (in log-scale), $X_k(0) = X_{k0}$, sampled from a normal distribution with a mean 8 and a standard deviation 5. Then, assuming the real observational data to have a random departure from the mean trajectory at time t_i, b_{ki}, which is assumed to follow a standard normal distribution. In addition, the measurement error $\varepsilon_k(t_{ij})$ is assumed to follow a normal distribution with mean zero and variance σ^2, and we took $\sigma^2 = 0.01$ and 0.1, respectively, in our simulation studies. In summary, the observational data were generated as $y_k(t_{ij}) = \bar{X}_k(t_{ij}) + b_{ki} + \varepsilon_k(t_{ij})$. The number of simulation runs is taken as 100.

We evaluate the performance of the proposed adaptive group LASSO for the PSA models derived from a set of ODEs based on the simulated data that are described above. In Step II, the number of basis functions for approximating the nonparametric components was taken as 12 and the data augmentation strategy (D'Haeseleer et al. 1999; Wessels, van Someren, and Reinders

Table 1. Simulation results for the adaptive group LASSO in Step IV. The numbers are the averages of true positive rate (TP%) and false positive rate (FP%) from 100 simulation replicates for different scenarios of sample sizes and measurement errors. Standard deviations are given in parenthesis

δ^2	n	Eq. no.	TP%	FP%	n	Eq. no.	TP%	FP%
0.1	15				150			
		(1)	51(0.57)	45(0.64)		(1)	74(0.39)	30(0.27)
		(2)	55(0.48)	44(0.37)		(2)	78(0.27)	27(0.34)
		(3)	61(0.56)	39(0.42)		(3)	80(0.37)	22(0.4)
		(4)	84(0.39)	45(0.53)		(4)	100(0.0)	17(0.33)
		(5)	61(0.72)	38(0.61)		(5)	78(0.31)	26(0.33)
		(6)	70(0.42)	33(0.39)		(6)	94(0.12)	25(0.27)
		(7)	72(0.37)	39(0.48)		(7)	93(0.22)	21(0.33)
		(8)	75(0.62)	37(0.69)		(8)	96(0.11)	19(0.17)
0.01	15				150			
		(1)	71(0.83)	29(0.21)		(1)	88(0.22)	7(0.09)
		(2)	74(0.51)	33(0.31)		(2)	95(0.13)	12(0.16)
		(3)	68(0.44)	25(0.25)		(3)	94(0.17)	8(0.20)
		(4)	89(0.26)	28(0.63)		(4)	100(0.0)	10(0.34)
		(5)	68(0.29)	29(0.79)		(5)	91(0.12)	16(0.14)
		(6)	79(0.47)	36(0.58)		(6)	100(0.0)	9(0.18)
		(7)	80(0.48)	31(0.44)		(7)	97(0.11)	13(0.12)
		(8)	88(0.33)	28(0.83)		(8)	100(0.0)	11(0.23)

2001; Bansal, Della Gatta, and di Bernardo 2006; Lu et al. 2011) was used (2000 data points were taken from the smoothed estimates in Step I). The simulation results are reported in Table 1. From Table 1, we can see that, when the sample size is smaller ($n = 15$) and the measurement error is larger ($\sigma^2 = 0.1$), the true positive rate (TP) of the variable selection by the proposed adaptive group LASSO method ranges from 51% to 84%, and the false positive rate (FP) ranges from 33% to 45% for the eight ODEs, respectively. However, when the sample size is increased ($n = 150$), the minimum TP for the eight ODEs is increased to 74% and the FP is decreased to 17%–30%. When the measurement error is reduced from $\sigma^2 = 0.1$ to $\sigma^2 = 0.01$ for the sample size $n = 15$, the minimum TP for the eight ODEs is 68% and the FP is decreased to 25%–36%. When both the sample size is increased ($n = 150$) and the measurement error is reduced ($\sigma^2 = 0.01$), the TP ranges 88%–100% and the FP is further decreased to 7%–16%. These simulation results show that the proposed adaptive LASSO method is reasonably good and tends to perform perfectly when the measurement error is small and the sample size is large.

To illustrate the effect of Step V (the regular LASSO for shrinking basis coefficients) in the proposed variable selection procedure in Section 2, we plot the estimated nonparametric functions from Step IV (adaptive group LASSO) and Step V as well as the true function from one simulation run in Figure 1. From this figure, we can see that the two estimated nonparametric functions have similar trends and can capture the complex nonlinear functions, but the adaptive group LASSO estimates from Step IV are too wiggly (dotted lines), which is presumably due to too many basis functions included (undersmoothing). That is why Step V is necessary to shrink some of the unnecessary basis coefficients to zero using the regular LASSO, which produces better (smoother) estimates (dashed lines in Figure 1). We can also see that Step V estimates perform better at boundaries and inflection points of the curves compared to Step IV estimates, although this observation is only based on one simulation case. But in general, Step V estimates should be better for more smooth functions compared with those of Step IV, since Step V is designed to correct the undersmoothing problem for Step IV. To quantify the gains from Step V, a comparison of residual sum of squares (RSS) between Step IV and V for each of 15 nonlinear functions are listed in Table 2. The RSS for each function f_{ij} is defined as $\text{RSS} = \sum_{n_{ij}=1}^{n_{ss}} [f_{ij}(x_{n_{ij}}) - \widehat{f_{ij}(x_{n_{ij}})}]^2$, where $\widehat{f_{ij}(x_{n_{ij}})}$ is estimated from Step IV or V and $f_{ij}(x_{n_{ij}})$ is the true function value. n_{ss} is the sample size (here we use 2000). Table 2 shows that the RSS is reduced by Step V from Step IV in most cases. However, for two functions, f_{22} and f_{51}, out of 15 functions, the RSS from Step V is larger than that from Step IV.

5. APPLICATION: IDENTIFICATION OF NONLINEAR DYNAMIC GENE REGULATORY NETWORKS

In this section, we apply the proposed method to identify a nonlinear dynamic gene regulatory network based on time course microarray data for T-cell activation. The central event in generation of an immune response is the activation of T-lymphocytes (T-cells). Activated T-cells proliferate and produce cytokines involved in the regulation of effector cells such as B cells and macrophages, which are the primary mediators of the immune response. T-cell activation is initiated by the interaction between the T-cell receptor (TCR) complex and the antigenic peptide presented on the surface of an antigen-presenting cell. This event triggers a network of signaling molecules, including kinases, phosphatases, and adaptor proteins that couple the stimulatory signal received from the TCR to gene transcription events in the nucleus (Ley et al. 1991; Iwashima et al. 1994).

To better understand the gene regulation network (GRN) during T-cell activation, Rangel et al. (2004) performed two experiments to characterize the response of a human T-cell line

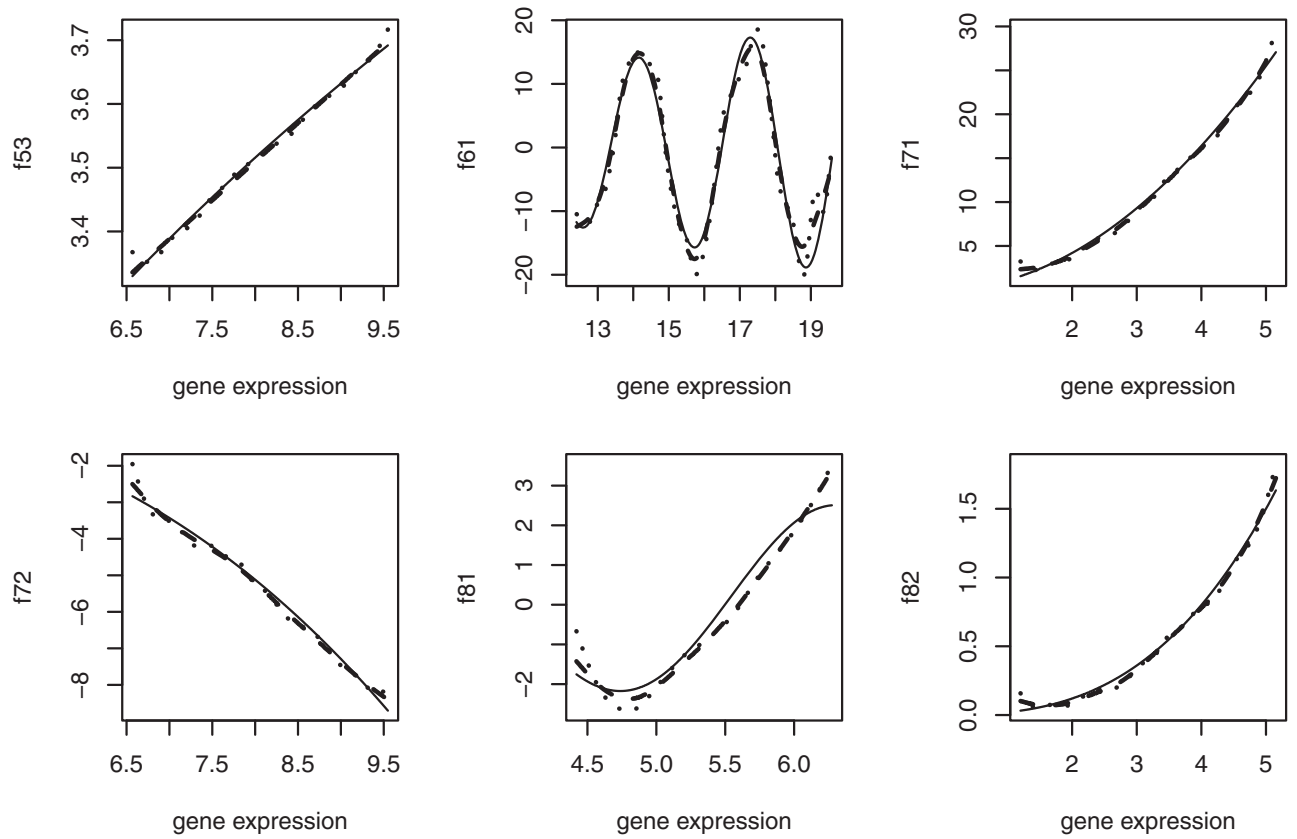

Figure 1. Simulation example: a comparison of the true functions (solid lines) and estimated functions from the adaptive group LASSO in Step IV (dotted lines) and from Step V after further basis coefficient shrinking (dashed lines) from one simulation run.

(Jurkat) to phorbol myristyl acetate (PMA) and ionomicin treatment. In the first experiment, they monitored the expression of 88 genes using cDNA array technology across 10 unequally spaced time points (0, 2, 4, 6, 8, 18, 24, 32, 48, 72 hr) and each gene was replicated 34 times. In the second experiment, an identical experimental protocol was used, but additional genes were added to the arrays and each gene was replicated 10 times. Genes that displayed very poor reproducibility between the two experiments were removed and the expression data for 58 genes were considered for final analysis by Rangel et al. (2004). Details of data collection and preprocessing can be found in Rangel et al. (2004). Rangel et al. (2004) applied a linear SSM to identify the transcriptional network based on this highly replicated expression profiling data. Beal et al. (2005) proposed a variational Bayesian treatment for identifying suitable dimension of hidden state space in SSM. In this article, we intend to apply the proposed SA-ODE model and the proposed variable selection method in the previous sections to establish a nonlinear dynamic regulatory network among 58 genes for the T-cell activation process. Through this application example, we will demonstrate that some of the gene regulation effects are essentially nonlinear and the linear network model may not be sufficient to capture the nonlinear features of the dynamic network.

We consider the 58 T-cell activation genes that were identified as reproducible from the two experiments by Rangel et al. (2004). For consistency, we only use the expression data from the first experiment with each gene replicated 34 times at 10 time points. The 34 replicates for each gene showed a similar expression pattern during the T-cell activation experiments. It is legitimate to smooth the data for each of the genes using the nonparametric mixed-effects smoothing splines technique (Lu et al. 2011; Wu and Zhang 2006) to obtain the estimates of the mean

Table 2. Comparisons of residual sum of squares (RSS) from Steps IV and V

	f11	**f12**	**f21** ($\times 10^3$)	**f22**	**f31** ($\times 10$)	**f32** ($\times 10^3$)	**f41** ($\times 10^3$)	**f51** ($\times 10^3$)
Step IV	7.0	53.1	8.2	1.5	34.3	7.0	9.0	1.1
Step V	6.8	9.9	8.0	2.7	8.6	5.8	1.2	2.8
	f52 ($\times 10^6$)	**f53** ($\times 10^{-2}$)	**f61** ($\times 10^5$)	**f71** ($\times 10^6$)	**f72** ($\times 10$)	**f81** ($\times 10^2$)	**f82** ($\times 10^3$)	
Step IV	8.6	35.3	1.5	1.2	19.2	8.0	4.0	
Step V	8.1	9.0	1.4	1.1	8.4	5.1	3.9	

NOTE: The bold formatting represents function $f_{i,j}(x)$ from Equations (4.4)–(4.8).

expression curves and their derivative curves for each gene, respectively. At the second step, these estimates were plugged in ODE model (1.3) to form the PSA model (2.3). The penalized pseudo-least-square methods, the group LASSO, and adaptive group LASSO approaches were used to identify significant regulations (connections) among the 58 genes with potential nonlinear regulation effects. We obtained the fitted curves for all genes by integrating 58 genes concurrently with parameters estimated from the adaptive grouped LASSO approach in Step IV and after shrinking the basis coefficients using regular LASSO in Step V, respectively.

We plot the fitted expression curves (dashed lines) for nine genes from Step V of the proposed procedure, overlaid with the raw data (dots) and the smoothed mean curves (solid lines) from Step I in Figure 2. For comparison purpose, the estimation results from the linear ODE model (Lu et al. 2011) were also obtained and plotted (dotted lines). From this figure, we can see that the proposed nonparametric additive ODE models fit the data better (also closer to the smoothed mean curves), compared with the linear ODE model fit. We also notice that the proposed SA-ODE models not only can fit the simple monotonic curves well, but also can flexibly fit complex nonlinear curves reasonably good. For each of 58 genes, we list the regulatory genes, identified by the proposed method in Table 3. One important feature of the identified GRN is that each of these 58 genes is regulated by only a few other genes (ranging from 1 to 8 genes), which reflects the fact of sparseness of the network connections.

In agreement with the findings from Rangel et al. (2004), our model identified the important adaptor molecule in TCR signaling pathway, FYB (gene 45 in Table 3) as one of the genes having the highest number of outward connections. In addition to the six genes, found by Rangel et al. (2004), that are regulated by FYB, we identified three more FYB-regulated genes that carry functions in proliferation (gene 2), interference (gene 44), and apoptosis (gene 49). This is presumably due

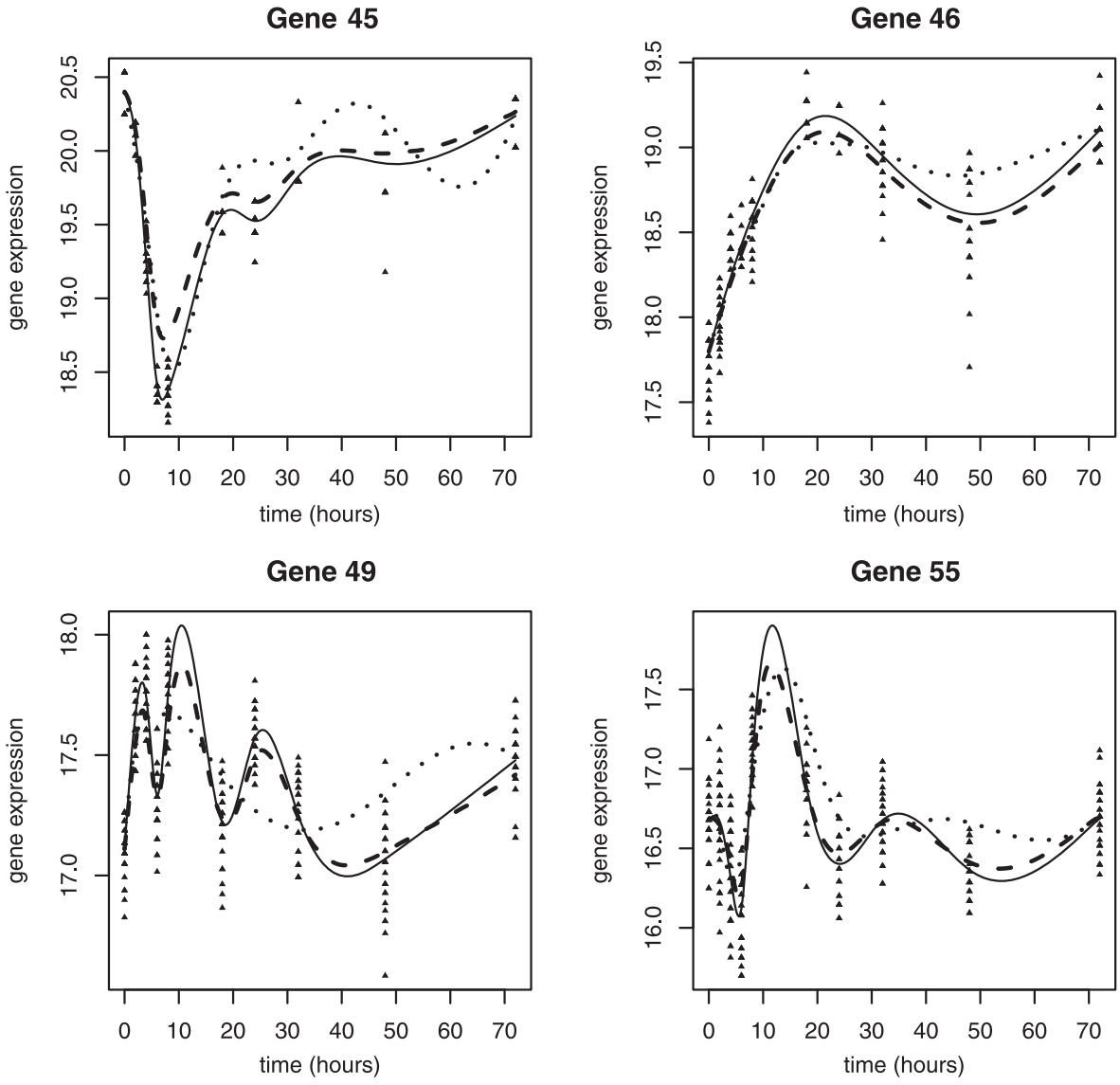

Figure 2. Real data analysis for T-cell activation experimental data (dots): the estimation results for gene FYB (gene 45) and 9 genes regulated by gene FYB, which include nonparametric smoothing estimates of mean expression curves from Step I (solid lines), the estimated curves (dashed lines) based on the proposed SA-ODE model from Step V of the proposed procedure, and the estimated curves (dotted lines) using the linear ODE model fit from Lu et al. (2011).

Table 3. T-cell activation gene regulatory network: variable selection results from the adaptive group LASSO. For a particular gene listed in Columns 1 and 4, Columns 2 and 5 provide the list of genes that have a significant regulation effects (inward influence) on this gene, and Columns 3 and 6 provide the list of genes that this particular gene has a significant regulation effect on outward influence

Gene	Inward influence genes	Outward influence genes	Gene	Inward influence genes	Outward influence genes
1	29, 32, 43		30	6, 46, 52	
2	32, 41, 45		31	25	6, 10, 23, 27, 35, 46, 48
3	29, 35, 41		32	48	1, 2, 5, 21, 24, 26, 56, 57
4	25		33	27, 28, 41	11, 23
5	32, 35		34	6, 27	
6	6, 25, 31	6, 9, 16, 26, 30, 34, 42, 52	35	20, 31	3, 5, 9, 14, 18, 20, 21, 38
7	45, 52, 57	11, 23	36	37	23
8	25		37	25	12, 13, 16, 22, 27, 36, 40, 56
9	6, 17, 27, 28, 35, 45		38	11, 28, 29, 35	
10	31, 51, 52		39	25	53
11	7, 13, 33, 46, 57	14, 18, 21, 38	40	37, 46	
12	29, 37		41	20, 55	2, 3, 15, 17, 25, 33, 42, 50, 58
13	37	11, 14, 15, 23, 50	42	6, 41, 46	
14	11, 13, 20, 28, 29, 35		43	27	1
15	13, 41, 45		44	28, 29, 45, 57	
16	6, 27, 37		45	52	2, 7, 9, 15, 18, 44, 46, 49, 55
17	41, 48	9, 54	46	31, 45	11, 23, 24, 30, 40, 42, 58
18	11, 35, 45, 54		47	27	
19	25	58	48	31, 48	17, 28, 32, 48
20	29, 35, 55	14, 25, 35, 41, 57	49	35, 45, 52	
21	11, 28, 32, 35		50	13, 41, 57	
22	27, 37, 52		51	25	10
23	7, 13, 31, 33, 36, 46, 57		52	6	7, 10, 22, 29, 30, 45, 49, 53, 55
24	32, 46, 57		53	39, 52	
25	20, 28, 41	4, 6, 8, 19, 26, 31, 37, 39, 51	54	17	
26	6, 25, 32		55	27, 45, 52	20, 41
27	31, 37	9, 16, 22, 33, 34, 43, 47, 55, 58	56	32, 37	
28	48, 57	9, 14, 21, 25, 33, 38, 44	57	20, 32	7, 11, 23, 24, 28, 44, 50
29	52	1, 3, 12, 14, 20, 38, 44	58	19, 27, 41, 46	

to that the proposed nonparametric additive ODE model and the variable selection method allow us to identify significant nonlinear regulation effects compared with the linear SSM used by Rangel et al. (2004). As another interesting finding based on the proposed model, we identified the direct regulation of integrin (gene 15) by FYB, which has already been corroborated by the experimental study (Peterson et al. 2001). On the contrary, the linear SSM used by Rangel et al. (2004) only found the indirect regulation mediated by IL3Rα (gene 55) and TRAF5 (gene 3). The estimated regulation functions of FYB (gene 45 in Table 3) to other genes are shown in Figure 3. Our results show that the gene FYB always has a positive regulation on genes 9, 18, 46, and 55, but a negative effect on gene 7, which agree with the findings from Rangel et al. (2004). For gene 15 that was found to be indirectly regulated by FYB in Rangel et al. (2004), we discover that the direct regulation effect is positive when the FYB expression level is low and turns to be negative when the FYB expression level is high. Similarly, for the three newly discovered FYB-regulated genes, genes 2, 44, and 49, by our method, we found that the regulation effect changes from positive to negative when the FYB expression is high. This demonstrates that the proposed nonparametric additive ODE model may help us to discover not only nonlinear regulation effects, but also varying effects due to the expression level of the regulator genes.

We use R function "igraph" to plot the obtained GRN in Figure 4 to visualize the complete network. From this figure, we can clearly see that there are 14 "big" regulator genes that regulate more than five other genes for each of them (see also Table 3). Due to space limitation, we will report more detailed annotations and biological implications of this established dynamic GRN for T-cell activation somewhere else. Notice that any statistical methods and modeling approaches can only identify potential regulation effects in actual GRNs. Usually some of these findings can be confirmed by existing literature and some others may need new experiments to validate. We expect that the proposed SA-ODE model and the variable selection methods provide a flexible tool to explore and identify additive nonlinear regulation effects in the GRNs.

6. CONCLUDING REMARKS

In this article, we have proposed a sparse additive ordinary differential equation (SA-ODE) model for dynamic GRNs to capture the nonlinear regulation effects. A five-step variable selection and estimation procedure by coupling the two-step smoothing-based ODE estimation method and LASSO techniques is developed. The asymptotic properties of the proposed methods are established. Simulation studies are conducted to demonstrate the good performance of the proposed

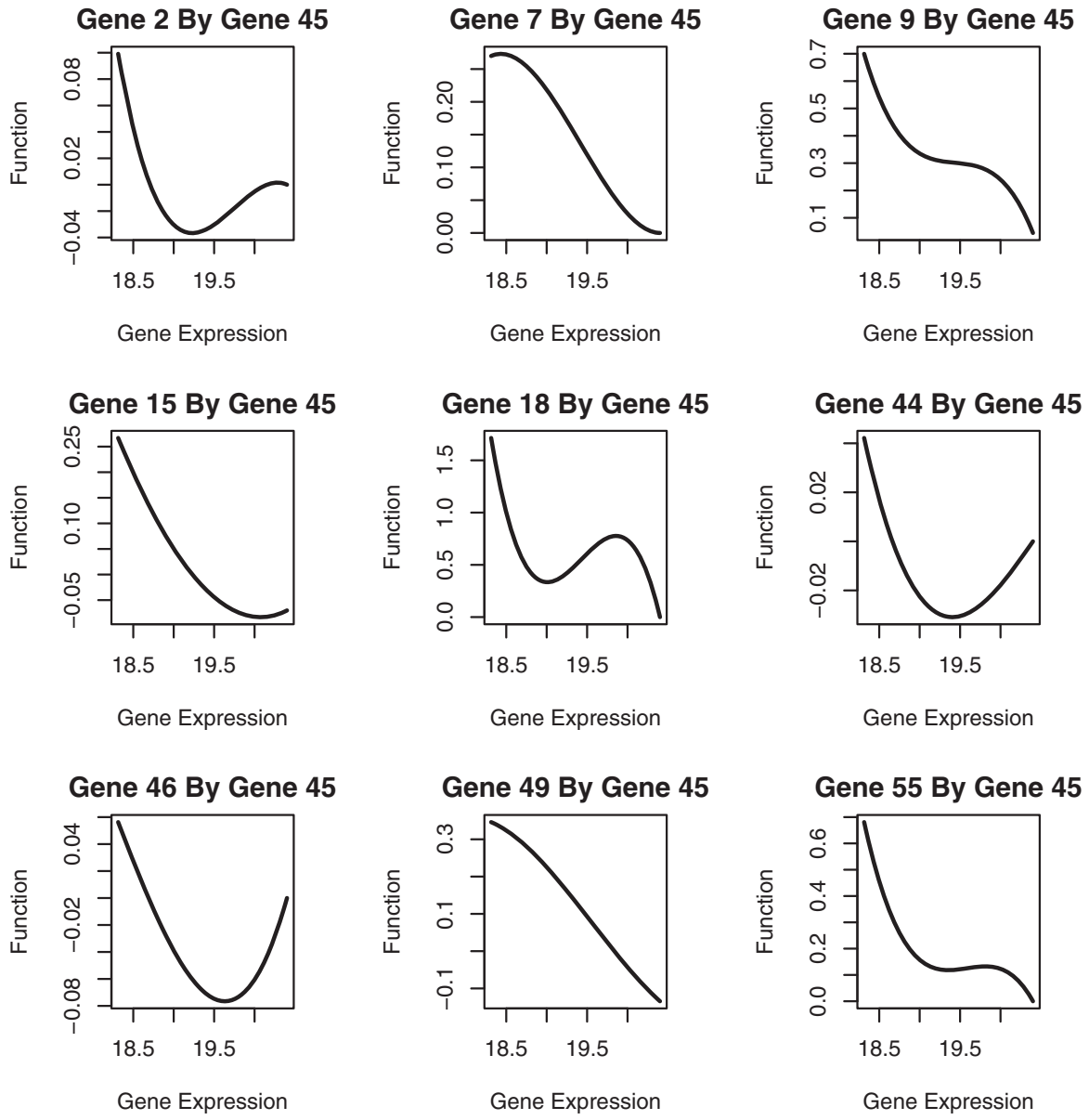

Figure 3. Estimated regulation functions of genes influenced by gene 45 (FYB).

variable selection method. We have successfully applied the proposed method to construct a nonlinear dynamic GRN for T-cell activation. We discovered additional new regulation effects in the complicated nonlinear network consisting of 58 genes, compared with the existing linear network modeling approach. The real data analysis results also show that the proposed SA-ODE model not only can capture complex nonlinear regulation effects, but also it can identify the varying-effects due to the expression levels of regulator genes.

In the theoretical development in Section 3, we have shown that the adaptive group LASSO selects the correct model with probability approaching 1 and achieves the optimal nonparametric convergence rate for the proposed SA-ODE model, which are similar to those obtained by Huang, Horowitz, and Wei (2010) for a nonparametric additive regression model. Notice that we convert the proposed SA-ODE model into a PSA model, which is different from the additive regression model in Huang, Horowitz, and Wei (2010) in the sense that both response variables and covariates in the PSA model are derived from the first-step nonparametric smoothing of state variables and their derivatives, instead of observed data, and the error terms in the PSA model are not iid but dependent. To tackle the challenges in this complex model, we used a weight function with boundary restrictions in the penalized least-square (LASSO) criteria and established a critical lemma to achieve the optimal convergence rate.

Alternative models for dynamic GRNs include boolean network, Bayesian network, hidden Markov models among others (Liang, Fuhrman, and Somogyi 1998; Murphy and Mian 1999; Akutsu, Miyano, and Kuhara 2000; Friedman et al. 2000; Hartemink et al. 2001; Shmulevich et al. 2002; Zou and Conzen 2005; Gupta, Qu, and Ibrahim 2007; Martin et al. 2007; Hirose et al. 2008; Kojima et al. 2009; Shimamura et al. 2009; Song, Kolar, and Xing 2009). It is interesting to compare these modeling approaches with the proposed ODE modeling method from the perspectives of computational cost and likelihood to recover

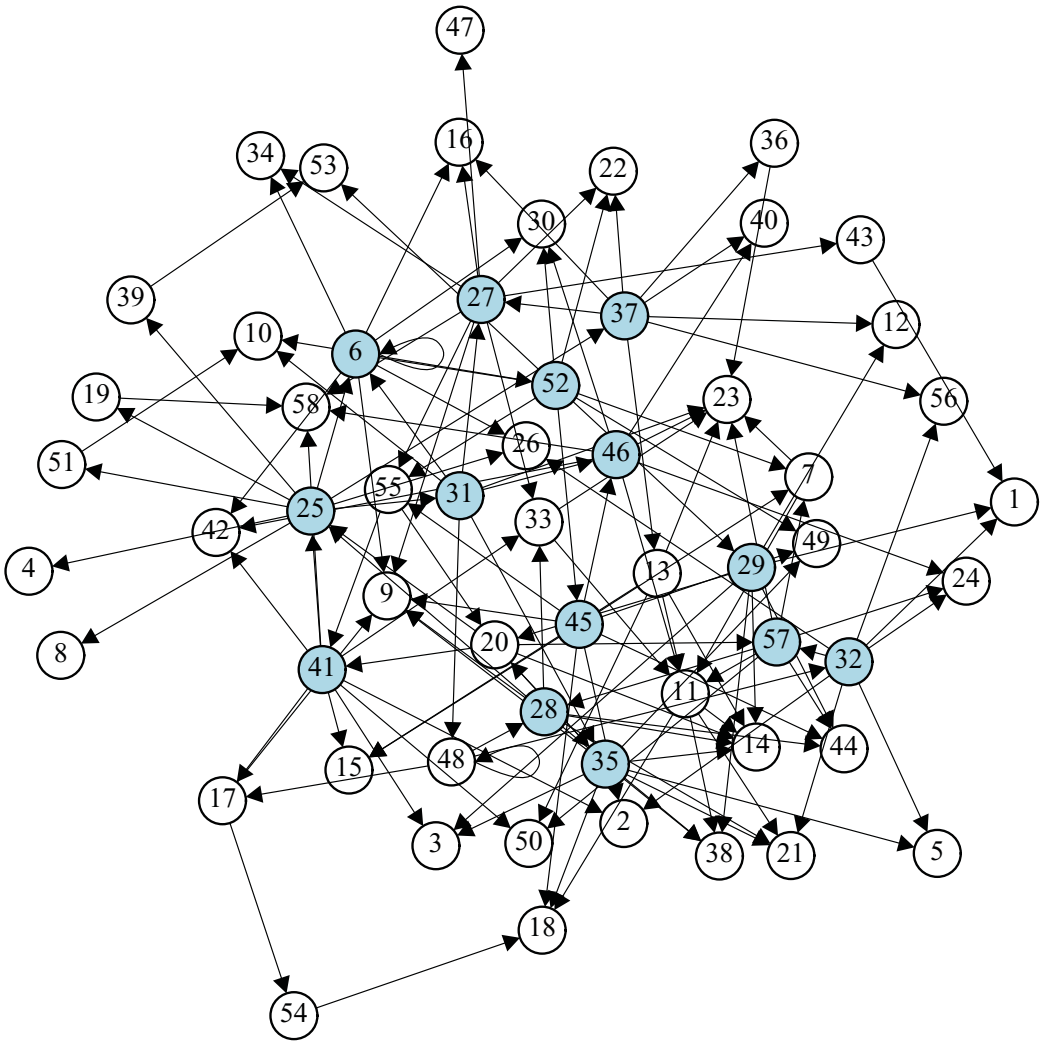

Figure 4. Graph of T-cell activation GRN formed by 58 genes. Each gene is represented as a node. Arrow stands for the direction of influence. The 14 "big" regulators are in blue.

the true network. But the computational cost for this comparison is beyond the limit of our current computational power. We employed the two-stage smoothing-based ODE estimation method for the SA-ODE model to reduce the computational cost and simplify the implementation of LASSO-based variable selection methods. One limitation of the two-stage smoothing-based ODE estimation method is its requirement of direct observations of all state variables in the system. It is still a challenging problem to perform variable selection for high-dimensional ODE models with partially observed state variables. Some other more efficient estimation approaches for ODE models such as the numerical discretization-based estimation method (Wu, Xue, and Kumar 2012) and generalized profiling approach (Ramsay et al. 2007) may also be considered for high-dimensional ODE variable selection. We believe that it is worth future exploration along a similar line, although the computational cost and implementation details need to be carefully considered. In the proposed SA-ODE model (1.3), we only consider the nonparametric additive structure of related variables. It could be extended to a nonparametric structure involving interactions (Radchenko and James 2010) since co-regulations are very often in gene regulatory networks. We are currently working on this problem and will report the results elsewhere.

APPENDIX: PROOFS

We can prove the theoretical results in Theorems 1–3 and Corollary 1 similarly to Huang, Horowitz, and Wei (2010). But there are more technical challenges that we need to deal with. The major differences between the PSA model (2.3) and the additive regression model in Huang, Horowitz, and Wei (2010) include: (i) both response variables and covariates in the PSA model are derived from the first-step nonparametric smoothing of state variables and their derivatives, instead of observed data, and (ii) the error terms, Υ_{ki} in the PSA model, are not iid but dependent. To tackle these problems, we establish the following lemma. For $1 \le k \le p$, let $T_{jm} = n^{-1/2} m_n^{1/2} \sum_{i=1}^{n} \psi_m(X_j(t_i)) d_{k1}^{1/2}(t_i)(\hat{X}'_k(t) - E[\hat{X}'_k(t)])$ for $1 \le j \le p$, $1 \le m \le m_n$ (for notational simplicity, we suppress the dependence of k), and $T_n = \max_{1 \le j \le p, 1 \le m \le m_n} |T_{jm}|$.

Lemma A.1. Under Assumptions A, B2–B4, and B6 in Section 3, we have $E(T_n) = O(1)\sqrt{\log(pm_n)}$.

Proof. From (2.2), we know that T_{jm} can be expressed as

$$\begin{aligned} T_{jm} &= n^{-1/2}m_n^{1/2}\sum_{i=1}^{n}\psi_m(X_j(t_i))d_{k1}^{1/2}(t_i)[\mathbf{N}'_{k,\nu+1}(t_i)]^T \\ &\quad\times\left(\mathbf{N}_k^T\mathbf{N}_k+\lambda_k\mathbf{V}_k\right)^{-1}\mathbf{N}_k^T\boldsymbol{\varepsilon}_k \\ &= n^{-1/2}m_n^{1/2}\sum_{\omega=1}^{n}\left[\sum_{i=1}^{n}\psi_m(X_j(t_i))d_{k1}^{1/2}(t_i)g_\omega(t_i)\right]\varepsilon_k(t_\omega), \end{aligned}$$

with $\boldsymbol{\varepsilon}_k=(\varepsilon_k(t_1),\ldots,\varepsilon_k(t_n))^T$ and $g_\omega(t_i)$ being the ωth component of the n-dimensional vector $[\mathbf{N}'_{k,\nu+1}(t_i)]^T(\mathbf{N}_k^T\mathbf{N}_k+\lambda_k\mathbf{V}_k)^{-1}\mathbf{N}_k^T$. Under Assumption B3, $\varepsilon_k(t_i)$'s are iid and sub-Gaussian. Moreover, for given t_i's, the coefficient $\sum_{i=1}^{n}\psi_m(X_j(t_i))d_{k1}^{1/2}(t_i)g_\omega(t_i)$ of $\varepsilon_k(t_\omega)$ is given. Therefore, conditional on t_i's, T_{jm}'s are sub-Gaussian.

Let $s_{jm}^2=\mathrm{var}(T_{jm}|t_1,\ldots,t_n)$ and $s_n^2=\max_{1\le j\le p,1\le m\le m_n}s_{jm}^2$. Then by the maximal inequality for sub-Gaussian random variables (Lemmas 2.2.1 and 2.2.2, van der Vaart and Wellner 1996), we have $\mathrm{E}(\max_{1\le j\le p,1\le m\le m_n}|T_{jm}||t_1,\ldots,t_n)\le C_1s_n\sqrt{\log(pm_n)}$, where $C_1>0$ is a constant. Therefore,

$$\mathrm{E}\Big(\max_{1\le j\le p,1\le m\le m_n}|T_{jm}|\Big)\le C_1\sqrt{\log(pm_n)}\mathrm{E}(s_n). \tag{A.1}$$

Now we discuss the order of $\mathrm{E}(s_n)$. Similar to Wu, Xue, and Kumar (2012), we consider the integral approximation of T_{jm} because of the boundary effects. By the strong law of large number, we have

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^{n}\psi_m(X_j(t_i))d_{k1}^{1/2}(t_i)(\hat{X}'_k(t_i)-\mathrm{E}[\hat{X}'_k(t_i)]) \\ &\quad=\int_{t_0}^{T}\psi_m(X_j(t))d_{k1}^{1/2}(t)\rho(t)(\hat{X}'_k(t)-\mathrm{E}[\hat{X}'_k(t)])dt+o(1). \end{aligned} \tag{A.2}$$

From the integration by parts, it follows that

$$\begin{aligned} &\int_{t_0}^{T}\psi_m(X_j(t))d_{k1}^{1/2}(t)\rho(t)(\hat{X}'_k(t)-\mathrm{E}[\hat{X}'_k(t)])dt \\ &\quad=\psi_m(X_j(t))d_{k1}^{1/2}(t)\rho(t)(\hat{X}_k(t)-\mathrm{E}[\hat{X}_k(t)])|_{t_0}^{T} \\ &\qquad-\int_{t_0}^{T}(\hat{X}_k(t)-\mathrm{E}[\hat{X}_k(t)])\frac{d}{dt}\big[\psi_m(X_j(t))d_{k1}^{1/2}(t)\rho(t)\big]dt \\ &\quad=-\int_{t_0}^{T}(\hat{X}_k(t)-\mathrm{E}[\hat{X}_k(t)])\frac{d}{dt}\big[\psi_m(X_j(t))d_{k1}^{1/2}(t)\rho(t)\big]dt, \end{aligned}$$

where $\mathrm{E}[\hat{X}'_k(t)]=(\mathrm{E}[\hat{X}_k(t)])'$ and the second equality holds because of the boundary condition $d_{k1}(t_0)=d_{k1}(T)=0$. Denote $A(t)=\psi_m(X_j(t))d_{k1}^{1/2}(t)\rho(t)$ and

$$\begin{aligned} \Sigma_X=\int_{t_0}^{T}\int_{t_0}^{T}A'(s)\mathrm{cov}\{&\hat{X}_k(s)-\mathrm{E}[\hat{X}_k(s)], \\ &\hat{X}_k(t)-\mathrm{E}[\hat{X}_k(t)]|t_1,\ldots,t_n\}A'(t)dsdt. \end{aligned}$$

By the Hölder's inequality, we have

$$\begin{aligned} \Sigma_X\le\|A'(s)A'(t)\|_\infty\int_{t_0}^{T}\int_{t_0}^{T}&\mathrm{cov}\{\hat{X}_k(s)-\mathrm{E}[\hat{X}_k(s)], \\ &\hat{X}_k(t)-\mathrm{E}[\hat{X}_k(t)]|t_1,\ldots,t_n\}dsdt. \end{aligned} \tag{A.3}$$

From the proof of Lemma 6 in Wu, Xue, and Kumar (2012), we know that

$$\begin{aligned} &\int_{t_0}^{T}\int_{t_0}^{T}\mathrm{cov}\{\hat{X}_k(s)-\mathrm{E}[\hat{X}_k(s)],\hat{X}_k(t)-\mathrm{E}[\hat{X}_k(t)]|t_1,\ldots,t_n\}dsdt \\ &\quad\le C_2n^{-1} \end{aligned} \tag{A.4}$$

for some positive constant C_2. By the extremal equality, it follows that $\|A'(s)A'(t)\|_\infty=\sup\{|\int_{t_0}^{T}\int_{t_0}^{T}A'(s)A'(t)g(s,t)dsdt|:g\in L^1(\mu,\mu),\|g\|_1\le1\}$, where $L^1(\mu,\mu)$ is the bivariate Lebesgue function space with L^1-norm. Continuing to apply integration by parts and the boundary condition $d_{k1}(t_0)=d_{k1}(T)=0$, we have

$$\begin{aligned} &\int_{t_0}^{T}\int_{t_0}^{T}A'(s)A'(t)g(s,t)dsdt \\ &\quad=\int_{t_0}^{T}\left\{A'(s)[g(t,s)A(t)]|_{t=t_0}^{t=T}-\int_{t_0}^{T}A(t)\frac{\partial g(t,s)}{\partial t}dt\right\}ds \\ &\quad=-\int_{t_0}^{T}\int_{t_0}^{T}A'(s)A(t)\frac{\partial g(t,s)}{\partial t}dsdt \\ &\quad=-\int_{t_0}^{T}A(t)dt\int_{t_0}^{T}\frac{\partial g(t,s)}{\partial t}dA(s) \\ &\quad=\int_{t_0}^{T}\int_{t_0}^{T}A(s)A(t)\frac{\partial^2g(t,s)}{\partial t\partial s}dsdt. \end{aligned}$$

Then $|\int_{t_0}^{T}\int_{t_0}^{T}A'(s)A'(t)g(s,t)dsdt|\le C_3\mathrm{cov}[\psi_m(X_j(s)),\psi_m(X_j(t))]$ for some positive constant C_3. By the properties of B-splines, we have

$$\mathrm{cov}[\psi_m(X_j(s)),\psi_m(X_j(t))]\le C_4m_n^{-1} \tag{A.5}$$

for some positive constant C_4. Combining (A.3)–(A.5) together, it follows that $\Sigma_X\le Cn^{-1}m_n^{-1}$ for some positive constant C. Then from (A.2), we have that $s_{jm}^2=\mathrm{var}(T_{jm}|t_1,\ldots,t_n)\le C$. Therefore $\mathrm{E}s_n\le(\mathrm{E}s_n^2)^{1/2}\le C$ for some positive constant C. Thus from (A.1), we have

$$\mathrm{E}\Big(\max_{1\le j\le p,1\le m\le m_n}|T_{jm}|\Big)\le C\sqrt{\log(pm_n)}$$

for some positive constant C. □

Proof of Theorem 1. The proof of part (i) is similar in spirit to the proof of Theorem 1 in Huang, Horowitz, and Wei (2010). But some technical challenges have to be tackled since here we need to consider the estimation errors of $\hat{X}'_k(t)$ and $\hat{X}_k(t)$ (involving the measurement error) and the approximation error of the additive regression functions by splines. Specifically, from (2.3) and Taylor expansion, we have

$$\begin{aligned} \Upsilon_{ki}^* &= \hat{X}'_k(t_i)-X'_k(t_i)+\sum_{j=1}^{q}f_{kj}(X_j(t_i))-\sum_{j=1}^{q}f_{kj}^*(\hat{X}_j(t_i)) \\ &= \hat{X}'_k(t_i)-X'_k(t_i)+\sum_{j=1}^{q}f_{kj}(X_j(t_i))-\sum_{j=1}^{q}f_{kj}(\hat{X}_j(t_i)) \\ &\quad+\sum_{j=1}^{q}f_{kj}(\hat{X}_j(t_i))-\sum_{j=1}^{q}f_{kj}^*(\hat{X}_j(t_i)) \\ &= \hat{X}'_k(t_i)-X'_k(t_i)+\sum_{j=1}^{q}f'_{kj}(X_{ji})(X_j(t_i)-\hat{X}_j(t_i)) \\ &\quad+\sum_{j=1}^{q}[f_{kj}(\hat{X}_j(t_i))-f_{kj}^*(\hat{X}_j(t_i))] \\ &= \hat{X}'_k(t_i)-\mathrm{E}[\hat{X}'_k(t_i)]-\sum_{j=1}^{q}f'_{kj}(X_{ji})\{\hat{X}_j(t_i)-\mathrm{E}[\hat{X}_j(t_i)]\} \\ &\quad+\mathrm{E}[\hat{X}'_k(t_i)]-X'_k(t_i)-\sum_{j=1}^{q}f'_{kj}(X_{ji})\mathrm{E}[\hat{X}_j(t_i)]-X_j(t_i)\} \\ &\quad+\sum_{j=1}^{q}[f_{kj}(\hat{X}_j(t_i))-f_{kj}^*(\hat{X}_j(t_i))], \end{aligned}$$

where X_{ji} locates between $\hat{X}_j(t_i)$ and $X_j(t_i)$.

For $\mathbf{D}_{k1}=\mathrm{diag}\{d_{k1}(t_1),\ldots,d_{k1}(t_n)\}$, we have $\mathbf{D}_{k1}^{1/2}=\mathrm{diag}\{d_{k1}^{1/2}(t_1),\ldots,d_{k1}^{1/2}(t_n)\}$. Let $\boldsymbol{\xi}_k=\mathbf{D}_{k1}^{1/2}(\boldsymbol{\Xi}_k+\boldsymbol{\Pi}_k+\boldsymbol{\delta}_k)$, where $\boldsymbol{\Xi}_k=$

$(\Xi_k(t_1),\ldots,\Xi_k(t_n))^T$, $\mathbf{\Pi}_k=(\Pi_{k1},\ldots,\Pi_{kn})^T$, and $\boldsymbol{\delta}_k=(\delta_{k1},\ldots,\delta_{kn})^T$ with $\Xi_k(t_i)=d_{k1}^{1/2}(t_i)(\hat{X}_k'(t_i)-\mathrm{E}[X_k'(t_i)]-\sum_{j=1}^q f_{kj}'(X_{ji})\{\hat{X}_j(t_i)-\mathrm{E}[\hat{X}_j(t_i)]\})$, $\Pi_{ki}=d_{k1}^{1/2}(t_i)(\mathrm{E}[X_k'(t_i)]-X_k'(t_i)-\sum_{j=1}^q f_{kj}'(X_{ji})\mathrm{E}[\hat{X}_j(t_i)]-X_j(t_i)\})$, $\delta_{ki}=d_{k1}^{1/2}(t_i)\sum_{j=1}^q[f_{kj}(\hat{X}_j(t_i))-f_{kj}^*(\hat{X}_j(t_i))]$. For any integer m, let

$$\chi_m=\max_{|A|=m}\max_{\|U_{A_k}\|_2=1,1\le k\le m}\frac{|\boldsymbol{\xi}_k^T V_A(\mathbf{S}_A)|}{\|V_A(\mathbf{S}_A)\|_2},$$

and

$$\chi_m^*=\max_{|A|=m}\max_{\|U_{A_k}\|_2=1,1\le k\le m}\frac{|\mathbf{\Xi}_k^T V_A(\mathbf{S}_A)|}{\|V_A(\mathbf{S}_A)\|_2},$$

where $V_A(\mathbf{S}_A)=\boldsymbol{\xi}_k^T[\mathbf{D}_{k1}^{1/2}\mathbf{Z}_A(\mathbf{Z}_A\mathbf{D}_{k1})\mathbf{Z}_A)^{-1}\mathbf{S}_A-(\mathrm{I}-\mathrm{P}_A)\mathbf{Z}\boldsymbol{\beta}]$ for $|A|=q_1=m\ge 0$ (here the inverse matrix can also be replaced by the generalized inverse), $\mathbf{S}_A=(S_{A_1}^T,\ldots,S_{A_m}^T)^T$, $S_{A_k}=\lambda_{k1}U_{A_k}$, and $\|U_{A_k}\|_2=1$. Then from the expression of $\hat{X}_k'(t)$ and $\hat{X}_k(t)$, we have $\mathbf{\Pi}_k=\mathbf{N}_k'(\mathbf{N}_k^T\mathbf{N}_k+\lambda_k\mathbf{V}_k)^{-1}\mathbf{N}_k^T\boldsymbol{\varepsilon}_k+\sum_{j=1}^q\mathrm{diag}\{f_{kj}'(X_{j1}),\ldots,f_{kj}'(X_{jn})\}\mathbf{N}_j(\mathbf{N}_j^T\mathbf{N}_j+\lambda_j\mathbf{V}_j)^{-1}\mathbf{N}_j^T\boldsymbol{\varepsilon}_j$ with $\mathbf{N}_k'=\{\mathbf{N}_{k,\nu+1}'(t_1),\ldots,\mathbf{N}_{k,\nu+1}'(t_n)\}^T$.

For a sufficiently large constant $C_1>0$, define

$$\Omega_{m_0}=\{(\mathbf{Z},\mathbf{\Xi}_k):\chi_m\le\sigma_k C_1\sqrt{(m\vee 1)m_n\log(pm_n)},\forall m\ge m_0\}$$

and

$$\Omega_{m_0}^*=\{(\mathbf{Z},\mathbf{\Xi}_k):\chi_m^*\le\sigma_k C_1\sqrt{(m\vee 1)m_n\log(pm_n)},\forall m\ge m_0\},$$

where $m_0\ge 0$. Similar to the proof of Theorem 2.1 of Wei and Huang (2010), it can follow that $(\mathbf{Z},\mathbf{\Xi}_k)\in\Omega_q\Rightarrow|\tilde{A}_1|\le M_1 q$ for a constant $M_1>1$ and

$$\mathrm{P}(\Omega_q^*)\to 1. \tag{A.6}$$

By the triangle and Cauchy–Schwarz inequalities,

$$\frac{|\boldsymbol{\xi}_k^T V_A(\mathbf{S}_A)|}{\|V_A(\mathbf{S}_A)\|_2}=\frac{|(\mathbf{\Xi}_k+\mathbf{\Pi}_k+\boldsymbol{\delta}_k)^T V_A(\mathbf{S}_A)|}{\|V_A(\mathbf{S}_A)\|_2}\le\frac{|\mathbf{\Xi}_k^T V_A(\mathbf{S}_A)|}{\|V_A(\mathbf{S}_A)\|_2}+\|\mathbf{\Pi}_k\|_2+\|\boldsymbol{\delta}_k\|_2. \tag{A.7}$$

Since $\|f_{kj}-f_{kj}^*\|_2=O(m_n^{-\varrho})=O(n^{-\varrho/(2\varrho+1)})$ for $m_n=n^{1/(2\varrho+1)}$, we have that, for all $m\ge q$ and n sufficiently large,

$$\begin{aligned}\|\boldsymbol{\delta}_k\|_2&\le C_2\sqrt{nqm_n^{-2\varrho}}=C_2q^{1/2}n^{1/(4\varrho+2)}\\&\le\sigma_k C_1\sqrt{(m\vee 1)m_n\log(pm_n)},\end{aligned} \tag{A.8}$$

for some constant $C_2>0$. From Lemma 1(i) of Wu, Xue, and Kumar (2012) and Theorem 2(a) of Claeskens, Krivobokova, and Opsomer (2009), we have $\mathrm{E}[\hat{X}_k'(t)]-X_k'(t)=O(\kappa^\nu)+O(\lambda_k n^{-1}\kappa^{-3})=O(n^{-\nu/(2\nu+3)})$ and $\mathrm{E}[\hat{X}_k(t)]-X_k(t)=O(\kappa^{\nu+1})+O(\lambda_k n^{-1}\kappa^{-2})=O(n^{-(\nu+1)/(2\nu+3)})$ for $\kappa=n^{-1/(2\nu+3)}$ and $\lambda_k=O(n^{\nu/(2\nu+3)})$. Then $\Pi_{ki}=O(qn^{-\nu/(2\nu+3)})$. Under Assumption B5, it follows that $3/(4\nu+6)\le 1/(4\varrho+2)$. Then for all $m\ge q$ and n sufficiently large, we have

$$\begin{aligned}\|\mathbf{\Pi}_k\|_2&\le C_3\sqrt{nqn^{-2\nu/(2\nu+3)}}=C_3q^{1/2}n^{3/(4\nu+6)}\\&\le\sigma_k C_1\sqrt{(m\vee 1)m_n\log(pm_n)},\end{aligned} \tag{A.9}$$

for some constant $C_3>0$. Finally, it follows from (A.6) to (A.9) that $\mathrm{P}(\Omega_q)\to 1$. This completes the proof of part (i) of Theorem 1.

Before proving part (ii), we first prove part (iii) of Theorem 1. By the definition of $\tilde{\boldsymbol{\beta}}_k=(\tilde{\beta}_{k1}^T,\ldots,\tilde{\beta}_{kp}^T)^T$,

$$\begin{aligned}&(\mathbf{H}_k-\mathbf{Z}\tilde{\boldsymbol{\beta}}_k)^T\mathbf{D}_{k1}(\mathbf{H}_k-\mathbf{Z}\tilde{\boldsymbol{\beta}}_k)+\lambda_{k1}\sum_{j=1}^p\|\tilde{\boldsymbol{\beta}}_{kj}\|_2\\&\le(\mathbf{H}_k-\mathbf{Z}\boldsymbol{\beta}_k)^T\mathbf{D}_{k1}(\mathbf{H}_k-\mathbf{Z}\boldsymbol{\beta}_k)+\lambda_{k1}\sum_{j=1}^p\|\boldsymbol{\beta}_{kj}\|_2.\end{aligned} \tag{A.10}$$

Let $A_2=\{j:\|\beta_{kj}\|_2\ne 0\text{ or }\|\tilde{\beta}_{kj}\|_2\ne 0\}$ and $d_{k2}=|A_2|$. By part (i), $d_{k2}=O_p(q)$. By (A.10) and the definition of A_2,

$$\begin{aligned}&(\mathbf{H}_k-\mathbf{Z}_{A_2}\tilde{\boldsymbol{\beta}}_{kA_2})^T\mathbf{D}_{k1}(\mathbf{H}_k-\mathbf{Z}_{A_2}\tilde{\boldsymbol{\beta}}_{kA_2})+\lambda_{k1}\sum_{j\in A_2}\|\tilde{\boldsymbol{\beta}}_{kj}\|_2\\&\le(\mathbf{H}_k-\mathbf{Z}_{A_2}\boldsymbol{\beta}_{kA_2})^T\mathbf{D}_{k1}(\mathbf{H}_k-\mathbf{Z}_{A_2}\boldsymbol{\beta}_{kA_2})+\lambda_{k1}\sum_{j\in A_2}\|\boldsymbol{\beta}_{kj}\|_2.\end{aligned} \tag{A.11}$$

Let $\boldsymbol{\eta}_k=\mathbf{D}_{k1}^{1/2}(\mathbf{H}_k-\mathbf{Z}\boldsymbol{\beta}_k)$ and $\boldsymbol{\nu}_k=\mathbf{D}_{k1}^{1/2}\mathbf{Z}_{A_2}(\tilde{\boldsymbol{\beta}}_{kA_2}-\boldsymbol{\beta}_{kA_2})$. Write $\mathbf{D}_{k1}^{1/2}(\mathbf{H}_k-\mathbf{Z}_{A_2}\tilde{\boldsymbol{\beta}}_{kA_2})=\mathbf{D}_{k1}^{1/2}(\mathbf{H}_k-\mathbf{Z}\boldsymbol{\beta}_k)-\mathbf{D}_{k1}^{1/2}\mathbf{Z}_{A_2}(\tilde{\boldsymbol{\beta}}_{kA_2}-\boldsymbol{\beta}_{kA_2})=\boldsymbol{\eta}_k-\boldsymbol{\nu}_k$. We have $(\mathbf{H}_k-\mathbf{Z}_{A_2}\tilde{\boldsymbol{\beta}}_{kA_2})^T\mathbf{D}_{k1}(\mathbf{H}_k-\mathbf{Z}_{A_2}\tilde{\boldsymbol{\beta}}_{kA_2})=\boldsymbol{\eta}_k^T\boldsymbol{\eta}_k-2\boldsymbol{\eta}_k^T\boldsymbol{\nu}_k+\boldsymbol{\nu}_k^T\boldsymbol{\nu}_k$. (A.11) can therefore be expressed as $\boldsymbol{\nu}_k^T\boldsymbol{\nu}_k-2\boldsymbol{\eta}_k^T\boldsymbol{\nu}_k\le\lambda_{k1}\sum_{j\in A_1}(\|\boldsymbol{\beta}_{kj}\|_2-\|\tilde{\boldsymbol{\beta}}_{kj}\|_2)$. Note that $|\sum_{j\in A_1}(\|\boldsymbol{\beta}_{kj}\|_2-\|\tilde{\boldsymbol{\beta}}_{kj}\|_2)|\le\sqrt{|A_1|}\|\tilde{\boldsymbol{\beta}}_{kA_1}-\boldsymbol{\beta}_{kA_1}\|_2\le\sqrt{|A_1|}\|\tilde{\boldsymbol{\beta}}_{kA_2}-\boldsymbol{\beta}_{kA_2}\|_2$. These two expressions indicate that

$$\boldsymbol{\nu}_k^T\boldsymbol{\nu}_k-2\boldsymbol{\eta}_k^T\boldsymbol{\nu}_k\le\lambda_{k1}\sqrt{|A_1|}\|\tilde{\boldsymbol{\beta}}_{kA_2}-\boldsymbol{\beta}_{kA_2}\|_2. \tag{A.12}$$

Let $\boldsymbol{\eta}_k^*$ be the projection of $\boldsymbol{\eta}_k$ on the span of $\mathbf{D}_{k1}^{1/2}\mathbf{Z}_{A_2}$, that is, $\boldsymbol{\eta}_k^*=\mathbf{D}_{k1}^{1/2}\mathbf{Z}_{A_2}(\mathbf{Z}_{A_2}^T\mathbf{D}_{k1}\mathbf{Z}_{A_2})^{-1}\mathbf{Z}_{A_2}^T\mathbf{D}_{k1}^{1/2}\boldsymbol{\eta}_k$. By Cauchy–Schwarz inequality and $ab\le a^2+b^2/4$, we have that

$$2|\boldsymbol{\eta}_k^T\boldsymbol{\nu}_k|\le 2\|\boldsymbol{\eta}_k^*\|_2\cdot\|\boldsymbol{\nu}_k\|_2\le 2\|\boldsymbol{\eta}_k^*\|_2^2+\frac{1}{2}\|\boldsymbol{\nu}_k\|_2^2. \tag{A.13}$$

From (A.12) and (A.13), we have

$$\boldsymbol{\nu}_k^T\boldsymbol{\nu}_k\le 4\|\boldsymbol{\eta}_k^*\|_2^2+2\lambda_{k1}\sqrt{|A_1|}\|\tilde{\boldsymbol{\beta}}_{kA_2}-\boldsymbol{\beta}_{kA_2}\|_2. \tag{A.14}$$

By the definition of $\mathbf{D}_{k1}$, we know that $\mathbf{Z}_{A_2}^T\mathbf{D}_{k1}\mathbf{Z}_{A_2}/n$ is a nonnegative definite matrix and its smallest positive eigenvalue c_{n*} exists. Under Assumptions A5 and B3, by Lemma 3 in Huang, Horowitz, and Wei (2010) and part (i), we have $c_{n*}\asymp m_n^{-1}$ with probability converging to 1. Since $\|\boldsymbol{\nu}_k\|_2^2\ge nc_{n*}\|\tilde{\beta}_{kA_2}-\boldsymbol{\beta}_{kA_2}\|_2^2$ and $2ab\le a^2+b^2$, from (A.14), we have

$$\begin{aligned}nc_{n*}\|\tilde{\boldsymbol{\beta}}_{kA_2}-\boldsymbol{\beta}_{kA_2}\|_2^2&\le 4\|\boldsymbol{\eta}_k^*\|_2^2+\frac{(2\lambda_{k1}\sqrt{|A_1|})^2}{2nc_{n*}}\\&\quad+\frac{1}{2}nc_{n*}\|\tilde{\boldsymbol{\beta}}_{kA_2}-\boldsymbol{\beta}_{kA_2}\|_2^2.\end{aligned}$$

It follows that

$$\|\tilde{\boldsymbol{\beta}}_{kA_2}-\boldsymbol{\beta}_{kA_2}\|_2^2\le\frac{8\|\boldsymbol{\eta}_k^*\|_2^2}{nc_{n*}}+\frac{4\lambda_{k1}^2|A_1|}{n^2c_{n*}^2}. \tag{A.15}$$

Let $f_k(\hat{X}(t_i))=\sum_{j=1}^p f_{kj}(\hat{X}_j(t_i))$, $f_{kA}(\hat{X}(t_i))=\sum_{j\in A}f_{kj}(\hat{X}_j(t_i))$, $f_k^*(\hat{X}(t_i))=\sum_{j=1}^p f_{kj}^*(\hat{X}_j(t_i))$, and $f_{kA}^*(\hat{X}(t_i))=\sum_{j\in A}f_{kj}^*(\hat{X}_j(t_i))$ with $f^*(.)$ defined by (2.4). For each component η_i of $\boldsymbol{\eta}_k$, we have

$$\begin{aligned}\eta_i&=d_{k1}^{1/2}(t_i)\left[H_k(t_i)-\bar{H}_k-\sum_{j\in A_2}\mathbf{Z}_{ij}^T\boldsymbol{\beta}_{kj}\right]\\&=d_{k1}^{1/2}(t_i)\{[\hat{X}_k'(t_i)-X_k'(t_i)]+[X_k'(t_i)-\mu_k-f_k(X(t_i))]\\&\quad+(\mu_k-\bar{H}_k)+[f_{kA_2}(X(t_i))-f_{k,A_2}(\hat{X}(t_i))]\\&\quad+[f_{kA_2}(\hat{X}(t_i))-f_{kA_2}^*(\hat{X}(t_i))]\}\\&=d_{k1}^{1/2}(t_i)\{[\Gamma_k(t_i)+(\mu_k-\bar{H}_k)+[f_{kA_2}(X(t_i))-f_{kA_2}(\hat{X}(t_i))]\\&\quad+[f_{kA_2}(\hat{X}(t_i))-f_{kA_2}^*(\hat{X}(t_i))]\}.\end{aligned}$$

Since $|\mu_k-\bar{H}_k|^2=O_p(n^{-1})$ and $\|f_{kj}-f_{kj}^*\|_\infty=O(m_n^{-\varrho})$, we have

$$\|\boldsymbol{\eta}_k^*\|_2^2\le 2\|\mathbf{\Gamma}_k^*\|_2^2+2\|\mathbf{\Lambda}_k^*\|_2^2+O_p(1)+O\big(nd_{n2}m_n^{-2\varrho}\big), \tag{A.16}$$

where $\mathbf{\Gamma}_k^*$ and $\mathbf{\Lambda}_k^*$ are projections of $\mathbf{\Gamma}_k=\{\Gamma_k(t_1),\ldots,\Gamma_k(t_n)\}^T$ and $\mathbf{\Lambda}_k=\{f_{kA_2}(\hat{X}(t_1)),\ldots,f_{kA_2}(\hat{X}(t_n))\}^T$ on the span of $\mathbf{D}_{k1}^{1/2}\mathbf{Z}_{A_2}$, respectively. We have $\|\mathbf{\Gamma}_k^*\|_2^2=\|(\mathbf{Z}_{A_2}^T\mathbf{D}_{k1}\mathbf{Z}_{A_2})^{-1/2}\mathbf{Z}_{A_2}^T\mathbf{D}_{k1}^{1/2}\mathbf{\Gamma}_k\|_2^2\le$

$1/nc_{n*}\|\mathbf{Z}_{A_2}^T\mathbf{D}_{k1}^{1/2}\boldsymbol{\Gamma}_k\|_2^2$. Now

$$\max_{A:|A|\le d_{n2}}\left\|\mathbf{Z}_{A_2}^T\mathbf{D}_{k1}^{1/2}\boldsymbol{\Gamma}_k\right\|_2^2 = \max_{A:|A|\le d_{n2}}\sum_{j\in A}\left\|\mathbf{Z}_j^T\mathbf{D}_{k1}^{1/2}\boldsymbol{\Gamma}_k\right\|_2^2 \le d_{n2}m_n \max_{1\le j\le p,1\le m\le m_n}\left|\mathcal{Z}_{jm}^T\mathbf{D}_{k1}^{1/2}\boldsymbol{\Gamma}_k\right|^2,$$

where $\mathcal{Z}_{jm} = \{\psi_m(\hat{X}_j(t_1)),\ldots,\psi_m(\hat{X}_j(t_i))\}^T$. By Lemma A.1, we have

$$\begin{aligned}&\max_{1\le j\le p,1\le m\le m_n}\left|\mathcal{Z}_{jm}^T\mathbf{D}_{k1}^{1/2}\boldsymbol{\Gamma}_k\right|^2\\ &\quad= nm_n^{-1}\max_{1\le j\le p,1\le m\le m_n}\left|(m_n/n)^{1/2}\mathcal{Z}_{jm}^T\mathbf{D}_{k1}^{1/2}\boldsymbol{\Gamma}_k\right|^2\\ &\quad= O_p(1)nm_n^{-1}\log(pm_n).\end{aligned}$$

It follows that,

$$\begin{aligned}\|\boldsymbol{\Gamma}_k^*\|_2^2 &= O_p(1)\frac{d_{n2}\log(pm_n)}{c_{n*}} \text{ and similarly}\\ \|\boldsymbol{\Lambda}_k^*\|_2^2 &= O_p(1)\frac{d_{n2}\log(pm_n)}{c_{n*}}.\end{aligned}\tag{A.17}$$

Combining (A.15)–(A.17), we get

$$\begin{aligned}\|\tilde{\boldsymbol{\beta}}_{kA_2}-\boldsymbol{\beta}_{kA_2}\|_2^2 = O_p\left(\frac{d_{n2}\log(pm_n)}{nc_{n*}^2}\right)+O_p\left(\frac{1}{nc_{n*}}\right)\\ +O\left(\frac{d_{n2}m_n^{-2\varrho}}{c_{n*}}\right)+\frac{4\lambda_{k1}^2|A_1|}{n^2c_{n*}^2}.\end{aligned}$$

Since $d_{n2}=O_p(q)$, $c_{n*}\asymp m_n^{-1}$ with probability converging to 1, we have

$$\begin{aligned}\|\tilde{\boldsymbol{\beta}}_{kA_2}-\boldsymbol{\beta}_{kA_2}\|_2^2 = O_p\left(\frac{m_n^2\log(pm_n)}{n}\right)+O_p\left(\frac{m_n}{n}\right)\\ +O\left(\frac{1}{m_n^{2\varrho-1}}\right)+O\left(\frac{m_n^2\lambda_{k1}^2}{n^2}\right).\end{aligned}\tag{A.18}$$

This completes the proof of part (iii).

We now prove part (ii). Under Assumption B2, $\|f_{kj}\|_2\ge c_f$, $1\le j\le q$, $\|f_{kj}-f_{kj}^*\|_2=O(m_n^{-\varrho})$, and $\|f_{kj}^*\|_2\ge\|f_{kj}\|_2-\|f_{kj}-f_{kj}^*\|_2$, we have $\|f_{kj}^*\|_2\ge 0.5c_f$ for n sufficiently large. By a result of de Boor (2001), see also (12) of Stone (1986), there are positive constants c_1 and c_2 such that $c_1m_n^{-1}\|\boldsymbol{\beta}_{kj}\|_2^2\le\|f_{kj}^*\|_2^2\le c_2m_n^{-1}\|\boldsymbol{\beta}_{kj}\|_2^2$. It follows that $\|\boldsymbol{\beta}_{kj}\|_2^2\ge c_2^{-1}m_n\|f_{kj}^*\|_2^2\ge 0.25c_2^{-1}c_f^2m_n$. Therefore, if $\|\boldsymbol{\beta}_{kj}\|_2\ne 0$ but $\|\tilde{\boldsymbol{\beta}}_{kj}\|_2=0$, then $\|\tilde{\boldsymbol{\beta}}_{kj}-\boldsymbol{\beta}_{kj}\|_2^2\ge 0.25c_2^{-1}c_f^2m_n$, which contradicts part (iii) since $m_n^2\log(pm_n)/n\to 0$ and $(\lambda_{k1}^2m_n^2)/n^2\to 0$ as $n\to\infty$. □

Proof of Theorem 2. By the definition of f_{kj}, $1\le j\le p$, parts (i) and (ii) follow from parts (i) and (ii) of Theorem 1 directly. Now, consider part (iii). By the properties of spline (de Boor 2001), there exist positive constants c_1 and c_2 such that $c_1m_n^{-1}\|\tilde{\boldsymbol{\beta}}_{kj}-\boldsymbol{\beta}_{kj}\|_2^2\le\|\tilde{f}_{kj}-f_{kj}^*\|_2^2\le c_2m_n^{-1}\|\tilde{\boldsymbol{\beta}}_{kj}-\boldsymbol{\beta}_{kj}\|_2^2$. Thus,

$$\begin{aligned}\|\tilde{f}_{kj}-f_{kj}^*\|_2^2 = O_p\left(\frac{m_n\log(pm_n)}{n}\right)+O_p\left(\frac{1}{n}\right)\\ +O\left(\frac{1}{m_n^{2\varrho}}\right)+O\left(\frac{4m_n\lambda_{k1}^2}{n^2}\right).\end{aligned}\tag{A.19}$$

By Assumption B3, $\|f_{kj}-f_{kj}^*\|_2^2=O(m_n^{-2\varrho})$. Part (iii) follows. □

Corollary 1 follows from Theorem 2 directly. The proof of Theorem 3 essentially follows the proof of Corollary 2 in Huang, Horowitz, and Wei (2010) by similar changes to those given in the proof of part (iii) of Theorem 1, and we omit it here.

[Received January 2012. Revised November 2012.]

REFERENCES

Akutsu, T., Miyano, S., and Kuhara, S. (2000), "Inferring Qualitative Relations in Genetic Networks and Metabolic Pathways," *Bioinformatics*, 16, 727–734. [700,710]

Bansal, M., Della Gatta, G., and di Bernardo, D. (2006), "Inference of Gene Regulatory Networks and Compound Mode of Action From Time Course Gene Expression Profiles," *Bioinformatics*, 22, 815–822. [702,706]

Bard, Y. (1974), *Nonlinear Parameter Estimation*, New York: Academic Press. [701]

Beal, M. J., Falciani, F., Ghahramani, Z., Rangel, C., and Wild, D. L. (2005), "A Bayesian Approach to Reconstructing Genetic Regulatory Networks With Hidden Factors," *Bioinformatics*, 21, 349–356. [707]

Bornholdt, S. (2008), "Boolean Network Models of Cellular Regulation: Prospects and Limitations," *Journal of the Royal Society Interface*, 5, S85–S94. [700]

Brunel, N. (2008), "Parameter Estimation of ODE's via Nonparametric Estimators," *Electronic Journal of Statistics*, 2, 1242–1267. [701,702,703]

Cantoni, E., Flemming, J. M., and Ronchetti, E. (2011), "Variable Selection in Additive Models by Nonnegative Garrote," *Statistical Modelling*, 11, 237–252. [701,702,704]

Carthew, R. W., and Sontheimer, E. J. (2009), "Origins and Mechanisms of miRNAs and siRNAs," *Cell*, 136, 642–655. [700]

Chen, J., and Chen, Z. (2008), "Extended Bayesian Information Criteria for Model Selection With Large Model Spaces," *Biometrika*, 95, 759–771. [703]

Chen, J., and Wu, H. (2008a), "Estimation of Time-Varying Parameters in Deterministic Dynamic Models," *Statistica Sinica*, 18, 987–1006. [701,702,704]

——— (2008b), "Efficient Local Estimation for Time-Varying Coefficients in Deterministic Dynamic Models With Applications to HIV-1 Dynamics," *Journal of the American Statistical Association*, 103, 369–384. [701,702,704]

Claeskens, G., Krivobokova, T., and Opsomer, J. D. (2009), "Asymptotic Properties of Penalized Spline Estimators," *Biometrika*, 96, 529–544. [702,704,713]

Craven, P., and Wahba, G. (1979), "Smoothing Noisy Data With Spline Functions: Estimating the Correct Degree of Smoothing by the Method of Generalized Cross-Validation," *Numerische Mathematik*, 31, 337–403. [702]

de Boor, C. (2001), *A Practical Guide to Splines* (revised ed.), New York: Springer-Verlag. [702,714]

DeJong, H. (2002), "Modeling and Simulation of Genetic Regulatory Systems: A Literature Review," *Journal of Computational Biology*, 9, 67–103. [700]

D'Haeseleer, P., Wen, X., Fuhrman, S., and Somogyi, R. (1999), "Linear Modeling of mRNA Expression Levels During CNS Development and Injury," *Pacific Symposium on Biocomputing*, 4, 41–52. [702,705]

Donnet, S., and Samson, A. (2007), "Estimation of Parameters in Incomplete Data Models Defined by Dynamical Systems," *Journal of Statistical Planning and Inference*, 137, 2815–2831. [701]

Fan, J., and Li, R. (2001), "Variable Selection via Nonconcave Penalized Likelihood and Its Oracle Properties," *Journal of the American Statistical Association*, 96, 1348–1360. [701]

Frank, I., and Friedman, J. (1993), "A Statistical View of Some Chemometrics Regression Tools" (with discussion), *Technometrics*, 35, 109–148. [701]

Friedman, N., Linial, M., Nachman, I., and Peer, D. (2000), "Using Bayesian Networks to Analyze Expression Data," *Journal of Computational Biology*, 7, 601–620. [700,710]

Guo, F., Hanneke, S., Fu, W., and Xing, E. P. (2007), "Recovering Temporally Rewiring Networks: A Model-Based Approach," in *Proceedings of the 24th International Conference on Machine Learning (ICML)*, 321–328. [700]

Gupta, M., and Ibrahim, J. G. (2007), "Variable Selection in Regression Mixture Modeling for the Discovery of Gene Regulatory Networks," *Journal of the American Statistical Association*, 102, 867–880. [700]

Gupta, M., Qu, P., and Ibrahim, J. G. (2007), "A Temporal Hidden Markov Regression Model for the Analysis of Gene Regulatory Networks," *Biostatistics*, 8, 805–820. [700,710]

Hanneke, S., and Xing, E. P. (2006), "Discrete Temporal Models of Social Networks," in *Proceedings of the Workshop on Statistical Network Analysis, the 23rd International Conference on Machine Learning (ICML-SNA)*, 115–125. [700]

Hartemink, A., Gifford, D., Jaakkola, T., and Young, R. (2001), "Using Graphical Models and Genomic Expression Data to Statistically Validate Models of Genetic Regulatory Networks," *Pacific Symposium on Biocomputing*, 6, 422–433. [700,710]

Hecker, M., Lambeck, S., Toepfer, S., van Someren, E., and Guthke, R. (2009), "Gene Regulatory Network Inference: Data Integration in Dynamic Models—A Review," *Biosystems*, 96, 86–103. [700]

Heckerman, D. (1996), "A Tutorial on Learning With Bayesian Networks," Technical Report No. MSR-TR–95-06, Microsoft Research. [700]

Heckman, N. E., and Ramsay, J. O. (2000), "Penalized Regression With Model-Based Penalties," *Canadian Journal of Statistics*, 28, 241–258. [701]

Hemker, P. (1972), "Numerical Methods for Differential Equations in System Simulation and in Parameter Estimation," in *Analysis and Simulation of Biochemical Systems*, eds. H. C. Hemker and B. Hess, Amsterdam: North-Holland Publishing, pp. 59–80. [701]

Hirose, O., Yoshida, R., Imoto, S., Yamaguchi, R., Higuchi, T., Charnock-Jones, D. S., Print, C., and Miyano, S. (2008), "Statistical Inference of Transcriptional Module-Based Gene Networks From Time Course Gene Expression Profiles by Using State Space Models," *Bioinformatics*, 24, 932–942. [700,710]

Holter, N. S., Maritan, A., Cieplak, M., Fedoroff, N. V., and Banavar, J. R. (2001), "Dynamic Modeling of Gene Expression Data," in *Proceedings of the National Academy of Sciences of the United States of America*, 98, 1693–1698. [700]

Huang, J., Horowitz, J. L., and Wei, F. (2010), "Variable Selection in Nonparametric Additive Models," *The Annals of Statistics*, 38, 2282–2313. [701,702,704,705,710,711,712,713,714]

Huang, J. Z. (2003), "Local Asymptotics for Polynomial Spline Regression," *The Annals of Statistics*, 31, 1600–1635. [704]

Huang, Y., Liu, D., and Wu, H. (2006), "Hierarchical Bayesian Methods for Estimation of Parameters in a Longitudinal HIV Dynamic System," *Biometrics*, 62, 413–423. [701]

Imoto, S., Sunyong, K., Goto, T., Aburatani, S., Tashiro, K., Kuhara, S., and Miyano, S. (2003), "Bayesian Network and Nonparametric Heteroscedastic Regression for Nonlinear Modeling of Genetic Network," *Journal of Bioinformatics and Computational Biology*, 6, 231–252. [700]

Iwashima, M., Irving, B. A., van Oers, N. S., Chan, A. C., and Weiss, A. (1994), "Sequential Interactions of the TCR With Two Distinct Cytoplasmic Tyrosine Kinases," *Science*, 263, 1136–1139. [706]

Jia, G., Stephanopoulos, G. N., and Gunawan, R. (2011), "Parameter Estimation of Kinetic Models From Metabolic Profiles: Two-Phase Dynamic Decoupling Method," *Bioinformatics*, 27, 1964–1970. [701,702]

Kauffman, S. A. (1969), "Metabolic Stability and Epigenesis in Randomly Constructed Genetic Nets," *Journal of Theoretical Biology*, 22, 437–467. [700]

Kim, H., Lee, J. K., and Park, T. (2009), "Inference of Large-Scale Gene Regulatory Networks Using Regression-Based Network Approach," *Journal of Bioinformatics and Computational Biology*, 7, 717–735. [700]

Kojima, K., Yamaguchi, R., Imoto, S., Yamauchi, M., Nagasaki, M., Yoshida, R., Shimamura, T., Ueno, K., Higuchi, T., Gotoh, N., and Miyano, S. (2009), "A State Space Representation of VAR Models With Sparse Learning for Dynamic Gene Networks," *Genome Informatics*, 22, 56–68. [700,710]

Kolar, M., Song, L., Ahmed, A., and Xing, E. P. (2010), "Estimating Time-Varying Networks," *Annals of Applied Statistics*, 4, 94–123. [700]

Ley, S. C., Davies, A. A., Druker, B., and Crumpton, M. J. (1991), "The T-Cell Receptor/CD3 Complex and CD2 Stimulate the Tyrosine Phosphorylation of Indistinguishable Patterns of Polypeptides in the Human T-Leukemic Cell-Line Jurkat," *European Journal of Immunology*, 21, 2203–2209. [706]

Li, Y., and Ruppert, D. (2008), "On the Asymptotics of Penalized Splines," *Biometrika*, 95, 415–436. [702]

Li, Z., Osborne, M. R., and Pravan, T. (2005), "Parameter Estimation in Ordinary Differential Equations," *IMA Journal of Numerical Analysis*, 25, 264–285. [701]

Liang, H., and Wu, H. (2008), "Parameter Estimation for Differential Equation Models Using a Framework of Measurement Error in Regression Models," *Journal of the American Statistical Association*, 103, 1570–1583. [701,702,704]

Liang, S., Fuhrman, S., and Somogyi, R. (1998), "REVEAL, a General Reverse Engineering Algorithm for Inference of Genetic Network Architectures," *Proceedings of the Pacific Symposium on Biocomputing*, 3, 18–29. [700,710]

Lu, T., Liang, H., Li, H., and Wu, H. (2011), "High Dimensional ODEs Coupled With Mixed-Effects Modeling Techniques for Dynamic Gene Regulatory Network Identification," *Journal of the American Statistical Association*, 106, 1242–1258. [701,702,705,707,708]

Martin, S., Zhang, Z., Martino, A., and Faulon, J. L. (2007), "Boolean Dynamics of Genetic Regulatory Networks Inferred From Microarray Time Series Data," *Bioinformatics*, 23, 866–874. [700,710]

Meier, L., van de Geer, S., and Buhlmann, P. (2009), "High-Dimensional Additive Modeling," *The Annals of Statistics*, 37, 3779–3821. [701,702,704]

Murphy, K., and Mian, S. (1999), "Modelling Gene Expression Data Using Dynamic Bayesian Networks," Technical Report, University of California, Berkeley. [700,710]

Needham, C. J., Bradford, J. R., Bulpitt, A. J., and Westhead, D. R. (2007), "A Primer on Learning in Bayesian Networks for Computational Biology," *PLoS Computational Biology*, 3, 1409–1416. [700]

Peterson, E. J., Woods, M. L., Dmowski, S. A., Derimanov, G., Jordan, M. S., Wu, J. N., Myung, P. S., Liu, Q., Pribila, J. T., Freedman, B. D., Shimizu, Y., and Koretzky, G. A. (2001), "Coupling of the TCR to Integrin Activation by SLAP-130/Fyb," *Science*, 293, 2263–2265. [709]

Poyton, A. A., Varziri, M. S., McAuley, K. B., McLellan, P. J., and Ramsay, J. O. (2006), "Parameter Estimation in Continuous-Time Dynamic Models Using Principal Differential Analysis," *Computers and Chemical Engineering*, 30, 698–708. [701]

Putter, H., Heisterkamp, S. H., Lange, J. M. A., and de Wolf, F. (2002), "A Bayesian Approach to Parameter Estimation in HIV Dynamical Models," *Statistics in Medicine*, 21, 2199–2214. [701]

Qi, X., and Zhao, H. (2010), "Asymptotic Efficiency and Finite-Sample Properties of the Generalized Profiling Estimation of the Parameters in Ordinary Differential Equations," *The Annals of Statistics*, 38, 435–481. [701]

Radchenko, P., and James, G. M. (2010), "Variable Selection Using Adaptive Nonlinear Interaction Structures in High Dimensions," *Journal of the American Statistical Association*, 105, 1541–1553. [711]

Ramsay, J. O. (1996), "Principal Differential Analysis: Data Reduction by Differential Operators," *Journal of the Royal Statistical Society,* Series B, 58, 495–508. [701]

Ramsay, J. O., Hooker, G., Campbell, D., and Cao, J. (2007), "Parameter Estimation for Differential Equations: A Generalized Smoothing Approach" (with discussion), *Journal of the Royal Statistical Society,* Series B, 69, 741–796. [701,711]

Ramsay, J. O., and Silverman, B. W. (2005), *Functional Data Analysis* (2nd ed.), New York: Springer. [701]

Rangel, C., Angus, J., Ghahramani, Z., Lioumi, M., Sotheran, E., Gaiba, A., Wild, D. L., and Falciani, F. (2004), "Modeling T-Cell Activation Using Gene Expression Profiling and State-Space Models," *Bioinformatics*, 20, 1361–1372. [706,707,708]

Ravikumar, P., Lafferty, J., Liu, H., and Wasserman, L. (2009), "Sparse Additive Models," *Journal of the Royal Statistical Society,* Series B, 71, 1009–1030. [701,702,704]

Ruppert, D., Wand, M. P., and Carroll, R. J. (2003), *Semiparametric Regression*, Cambridge: Cambridge University Press. [702]

Sakamoto, E., and Iba, H. (2001), "Inferring a Systems of Differential Equations for a Gene Regulatory Network by Using Genetic Programming," in *Proceedings of the IEEE Congress on Evolutionary Computation*, pp. 720–726. [701]

Schumaker, L. L. (1981), *Spline Functions: Basic Theory*, New York: Wiley. [702]

Shimamura, T., Imoto, S., Yamaguchi, R., Fujita, A., Nagasaki, M., and Miyano, S. (2009), "Recursive Regularization for Inferring Gene Networks From Time-Course Gene Expression Profiles," *BMC Systems Biology*, 3, 41–54. [700,710]

Shmulevich, I., Dougherty, E. R., Kim, S., and Zhang, W. (2002), "Probabilistic Boolean Networks: A Rule-Based Uncertainty Model for Gene Regulatory Networks," *Bioinformatics*, 18, 261–274. [700,710]

Shojaie, A., Basu, S., and Michailidis, G. (2012), "Adaptive Thresholding for Reconstructing Regulatory Networks From Time-Course Gene Expression Data," *Statistics in Biosciences*, 4, 66–83. [700]

Shojaie, A., and Michailidis, G. (2009), "Analysis of Gene Sets Based on the Underlying Regulatory Network," *Journal of Computational Biology*, 16, 407–426. [700]

Song, L., Kolar, M., and Xing, E. P. (2009), "Time-Varying Dynamic Bayesian Networks," in *Advances in Neural Information Processing Systems 22*, eds. Y. Bengio, D. Schuurmans, J. Lafferty, C. K. I. Williams, A. Culotta, NIPS Foundation, pp. 1732–1740. [700,710]

Spieth, C., Hassis, N., and Streichert, F. (2006), "Comparing Mathematical Models on the Problem of Network Inference," in *Proceedings of 8th Annual Conference on Genetic and evolutionary computation*, pp. 279–285. [701]

Steuer, R., Kurths, J., Daub, C. O., Weise, J., and Selbig, J. (2002), "The Mutual Information: Detecting and Evaluating Dependencies Between Variables," *Bioinformatics*, 18, S231–S240. [700]

Stone, C. J. (1986), "The Dimensionality Reduction Principle for Generalized Additive Models," *The Annals of Statistics*, 14, 590–606. [714]

Stuart, J. M., Segal, E., Koller, D., and Kim, S. K. (2003), "A Gene-Coexpression Network for Global Discovery of Conserved Genetic Modules," *Science*, 302, 249–255. [700]

Thomas, R. (1973), "Boolean Formalization of Genetic Control Circuits," *Journal of Theoretical Biology*, 42, 563–585. [700]

Tibshirani, R. J. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society,* Series B, 58, 267–288. [701]

van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer. [712]

Varah, J. M. (1982), "A Spline Least Squares Method for Numerical Parameter Estimation in Differential Equations," *SIAM Journal on Scientific and Statistical Computing*, 3, 28–46. [701,702,704]

Varziri, M. S., Poyton, A. A., McAuley, K. B., McLellan, P. J., and Ramsay, J. O. (2008), "Selecting Optimal Weighting Factors in iPDA for Parameter Estimation in Continuous-Time Dynamic Models," *Computers and Chemical Engineering*, 32, 3011–3022. [701]

Voit, E. O. (2000), *Computational Analysis of Biochemical Systems: A Practical Guide for Biochemists and Molecular Biologists*, Cambridge: Cambridge University Press. [700]

Voit, E. O., and Almeida, J. (2004), "Decoupling Dynamical Systems for Pathway Identification From Metabolic Profiles," *Bioinformatics*, 22, 1670–1681. [701,702]

Wang, H., and Leng, C. (2008), "A Note on the Adaptive Group Lasso," *Computational Statistics and Data Analysis*, 52, 5277–5286. [701,702]

Weaver, D. C., Workman, C. T., and Stormo, G. D. (1999), "Modeling Regulatory Networks With Weight Matrices," *Pacific Symposium on Biocomputing*, 4, 112–123. [701]

Wei, F., and Huang, J. (2010), "Consistent Group Selection in High-Dimensional Linear Regression," *Bernoulli*, 16, 1369–1384. [713]

Werhli, A. V., and Husmeier, D. (2007), "Reconstructing Gene Regulatory Networks With Bayesian Networks by Combining Expression Data With Multiple Sources of Prior Knowledge," *Statistical Applications in Genetics and Molecular Biology*, 6, 1–47. [700]

Wessels, L. F., van Someren, E. P., and Reinders, M. J. (2001), "A Comparison of Genetic Network Models," *Pacific Symposium on Biocomputing*, 6, 508–519. [702,706]

Wu, H., Xue, H., and Kumar, A. (2012), "Numerical Algorithm-Based Estimation Methods for ODE Models via Penalized Spline Smoothing," *Biometrics*, 68, 344–352. [701,702,703,704,705,711,712,713]

Wu, H., and Zhang, J. T. (2006), *Nonparametric Regression Methods for Longitudinal Data Analysis*, Hoboken, NJ: Wiley. [702,705,707]

Xue, H., Miao, H., and Wu, H. (2010), "Sieve Estimation of Constant And Time-Varying Coefficients in Nonlinear Ordinary Differential Equation Models by Considering Both Numerical Error and Measurement Error," *The Annals of Statistics*, 38, 2351–2387. [701,704]

Yeung, M. K. S., Tegner, J., and Collins, J. J. (2002), "Reverse Engineering Gene Networks Using Singular Value Decomposition and Robust Regression," *Proceedings of the National Academy of Sciences of the United States of America*, 99, 6163–6168. [700]

Yuan, M., and Lin, Y. (2006), "Model Selection and Estimation in Regression With Grouped Variables," *Journal of the Royal Statistical Society,* Series B, 68, 49–67. [701,702]

Zhou, S., Shen, X., and Wolfe, D. (1998), "Local Asymptotics for Regression Splines and Confidence Regions," *The Annals of Statistics*, 26, 1760–1782. [705]

Zhou, S., and Wolfe, D. (2000), "On Derivative Regression in Spline Estimation," *Statistica Sinica*, 10, 93–108. [705]

Zou, H. (2006), "The Adaptive LASSO and Its Oracle Properties," *Journal of the American Statistical Association*, 101, 1418–1429. [701,702]

Zou, H., and Hastie, T. (2006), "Regularization and Variable Selection via the Elastic Net," *Journal of the Royal Statistical Society,* Series B, 67, 301–320. [701]

Zou, M., and Conzen, S. D. (2005), "A New Dynamic Bayesian Network (DBN) Approach for Identifying Gene Regulatory Networks From Time Course Microarray Data," *Bioinformatics*, 21, 71–79. [700,710]