Network Reconstruction From High Dimensional Ordinary Differential Equations

Shizhe Chen, Ali Shojaie, and Daniela M. Witten*

Department of Biostatistics
University of Washington
Box 357232
Seattle, WA 98195-7232

*Shizhe Chen is Graduate Student, Department of Biostatistics, University of Washington, WA 98195 (e-mail: shizhe.chen@gmail.com); Ali Shojaie is Associate Professor, Department of Biostatistics, and Adjunct Associate Professor, Department of Statistics, University of Washington, WA 98195 (e-mail: ashojaie@u.washington.edu); and Daniela M. Witten is Associate Professor, Departments of Biostatistics and Statistics, University of Washington, WA 98195 (e-mail: dwitten@u.washington.edu). We thank the associate editor and two anonymous reviewers for helpful comments. We thank the authors of Brunel et al. (2014), Hall and Ma (2014), Henderson and Michailidis (2014), and Wu et al. (2014) for sharing their code for their proposals, and for responding to our inquiries. We thank the Allen Institute for Brain Science for providing the data set analyzed in Section 6.2. A.S. was supported by NSF grant DMS-1561814 and NIH grants 1K01HL124050-01A1 and 1R01GM114029-01A1, and D.W. was supported by NIH Grant DP5OD009145, NSF CAREER Award DMS-1252624, and an Alfred P. Sloan Foundation Research Fellowship.

Abstract

We consider the task of learning a dynamical system from high-dimensional time-course data. For instance, we might wish to estimate a gene regulatory network from gene expression data measured at discrete time points. We model the dynamical system non-parametrically as a system of additive ordinary differential equations. Most existing methods for parameter estimation in ordinary differential equations estimate the derivatives from noisy observations. This is known to be challenging and inefficient. We propose a novel approach that does not involve derivative estimation. We show that the proposed method can consistently recover the true network structure even in high dimensions, and we demonstrate empirical improvement over competing approaches.

Keywords Additive model; Group lasso; High dimensionality; Ordinary differential equation; Variable selection consistency

1. INTRODUCTION

Ordinary differential equations (ODEs) have been widely used to model dynamical systems in many fields, including chemical engineering (Biegler et al., 1986), genomics (Chou and Voit, 2009), neuroscience (Izhikevich, 2007), and infectious diseases (Wu, 2005). A system of ODEs takes the form

$$X'(t;\theta) \equiv \begin{bmatrix} \frac{dX_1(t;\theta)}{dt} \\ \vdots \\ \frac{dX_p(t;\theta)}{dt} \end{bmatrix} = \begin{bmatrix} f_1(X(t;\theta),\theta) \\ \vdots \\ f_p(X(t;\theta),\theta) \end{bmatrix} \equiv f(X(t;\theta),\theta); \quad t \in [0,1], \tag{1}$$

where $X(t;\theta) = (X_1(t;\theta), \ldots, X_p(t;\theta))^{\mathrm{T}}$ denotes a set of variables, and the form of the functions $f = (f_1, \ldots, f_p)^{\mathrm{T}}$ may be known or unknown. In (1), t indexes time. Typically, there is also an initial condition of the form $X(0;\theta) = C$, where C is a p-vector. In practice, the system (1) is often observed on discrete time points subject to measurement errors. Let $Y_i \in \mathbb{R}^p$ be the measurement of the system at time t_i such that

$$Y_i = X(t_i;\theta^*) + \epsilon_i, \quad i = 1, \ldots, n, \tag{2}$$

where θ^* denotes the true set of parameter values and the random p-vector ϵ_i represents independent measurement errors. In what follows, for notational simplicity, we sometimes suppress the dependence of $X(t;\theta)$ on θ, i.e., $X(t) \equiv X(t;\theta)$ in (1) and $X^*(t) \equiv X(t;\theta^*)$ in (2).

In the context of high-dimensional time-course data arising from biology, it can be of interest to recover the structure of a system of ODEs — that is, to determine which features regulate each

other. If f_j in (1) is a function of X_k, then we say that X_k *regulates* X_j in the sense that X_k controls the changes of X_j through its derivative X_j'. For instance, biologists might want to infer gene regulatory networks from noisy time-course gene expression data. In this case, the number of variables p exceeds the number of time points n; we refer to this as the high-dimensional setting.

In high-dimensional statistics, sparsity-inducing penalties such as the lasso (Tibshirani, 1996) and the group lasso (Yuan and Lin, 2006) have been well-studied. Such penalties have also been extensively used to recover the structure of probabilistic graphical models (e.g., Yuan and Lin, 2007; Friedman et al., 2008; Meinshausen and Bühlmann, 2010; Voorman et al., 2014). However, model selection in high-dimensional ODEs remains a relatively open problem, with the exception of some notable recent work (Lu et al., 2011; Henderson and Michailidis, 2014; Wu et al., 2014). In fact, the tasks of parameter estimation and model selection in ODEs from noisy data are very challenging, even in the classical statistical setting where $n > p$ (see e.g., Ramsay et al., 2007; Brunel, 2008; Liang and Wu, 2008; Qi and Zhao, 2010; Xue et al., 2010; Gugushvili and Klaassen, 2012; Hall and Ma, 2014; Zhang et al., 2015). Moreover, the problem of high-dimensionality is compounded if the form of the function f in (1) is unknown, leading to both statistical and computational issues.

In this paper, we propose an efficient procedure for structure recovery of an ODE system of the form (1) from noisy observations of the form (2), in the setting where the functional form of f is unknown. In Section 2, we review existing methods. In Section 3, we propose a new structure recovery procedure. In Section 4, we study the theoretical properties of our proposal. In Section 5, we apply our procedure to simulated data. In Section 6, we apply it to *in silico* gene expression data generated by GeneNetWeaver (Schaffter et al., 2011) and to calcium imaging data. We conclude with a discussion in Section 7. Proofs and additional details are provided in the supplementary

material.

2. LITERATURE REVIEW

In this section, we review existing statistical methods for parameter estimation and/or model selection in ODEs. Most of the methods reviewed in this section are proposed for the low-dimensional setting. Even though they may not be directly applicable to the high-dimensional setting, they lay the foundation for the development of model selection procedures in high-dimensional additive ODEs.

2.1 Notation

Without loss of generality, assume that $0 = t_1 < t_2 < \ldots < t_n = 1$. We let Y_{ij} indicate the observation of the jth variable at the ith time point, t_i. We use $\mathcal{X}(h)$ to denote a nonparametric class of functions on $[0, 1]$ indexed by some smoothing parameter(s) h. We use $Z(\cdot)$ to represent an arbitrary function belonging to $\mathcal{X}(\cdot)$. We use $\|\cdot\|_2$ to denote the ℓ_2-norm of a vector or a matrix, and $|||f|||$ to denote the ℓ_2-norm of a function f on the interval $[0, 1]$, i.e. $|||f|||^2 \equiv \int_0^1 f^2(t)\,dt$. We use an asterisk to denote true values—for instance, θ^* denotes the true value of θ in (1). We use $\Lambda_{\min}(A)$ and $\Lambda_{\max}(A)$ to denote the minimum and maximum eigenvalues of a square matrix A, respectively.

2.2 Methods that assume a known form of f

2.2.1 Gold standard approach

To begin, we suppose that the function f in (1) takes a known form. Benson (1979) and Biegler et al. (1986) proposed to estimate the unknown parameter θ^* in (2) by solving the problem

$$\hat{\theta}^{\text{gold}} = \arg\min_{\theta} \sum_{i=1}^{n} \|Y_i - X(t_i;\theta)\|_2^2 \tag{3a}$$

$$\text{subject to} \quad X'(t;\theta) = f(X(t;\theta),\theta), \quad t \in [0,1]. \tag{3b}$$

Note that $X(\cdot;\theta)$ in (3) is a fixed function given θ, although an analytic expression may not be available. The resulting estimator $\hat{\theta}^{\text{gold}}$ has appealing theoretical properties: for instance, when the measurement errors ϵ_i in (2) are Gaussian, then $\hat{\theta}^{\text{gold}}$ is the maximum likelihood estimator, and is $\sqrt{n}$-consistent. In this sense, (3) can thus be considered the *gold standard* approach. However, solving (3) is often computationally challenging.

2.2.2 Two-step collocation methods

In order to overcome the computational challenges associated with solving (3), *collocation* methods have been employed by a number of authors (Varah, 1982; Ellner et al., 2002; Ramsay et al., 2007; Brunel, 2008; Cao and Zhao, 2008; Liang and Wu, 2008; Cao et al., 2011; Lu et al., 2011; Gugushvili and Klaassen, 2012; Brunel et al., 2014; Hall and Ma, 2014; Henderson and Michailidis, 2014; Wu et al., 2014; Dattner and Klaassen, 2015; Zhang et al., 2015).

The two-step collocation procedure first proposed by Varah (1982) involves fitting a smoothing estimate $\hat{X}(\cdot;h)$ to the observations $Y_1,\ldots,Y_n$ in (2) with a smoothing parameter h, and then

plugging $\hat{X}(\cdot; h)$ and its derivative with respect to t into (1) in order to estimate θ. This amounts to solving the optimization problem

$$\hat{\theta}^{\text{TS}} = \arg\min_{\theta} \int_0^1 \left\| \hat{X}'(t; h) - f\big(\hat{X}(t; h), \theta\big) \right\|_2^2 dt, \tag{4a}$$

where

$$\hat{X}(\cdot; h) = \arg\min_{Z(\cdot) \in \mathcal{X}(h)} \sum_{i=1}^{n} \|Y_i - Z(t_i)\|_2^2. \tag{4b}$$

The two-step procedure (4) has a clear advantage over the gold standard approach (3) because the former decouples the estimation of θ and X. However, this advantage comes at a cost: due to the presence of $\hat{X}'$ in (4a), the properties of the estimator $\hat{\theta}^{\text{TS}}$ in (4) rely heavily on the smoothing estimates obtained in (4b), and $\sqrt{n}$-consistency has only been shown for certain values of the smoothing parameter h that are hard to choose in practice (Brunel, 2008; Liang and Wu, 2008; Gugushvili and Klaassen, 2012).

Dattner and Klaassen (2015) proposed an improvement to (4) for a special case of (1). To be more specific, they assume that $f_j(X(t), \theta)$ in (1) is a linear function of θ, which leads to

$$X'(t) \equiv \begin{bmatrix} \frac{dX_1(t)}{dt} \\ \vdots \\ \frac{dX_p(t)}{dt} \end{bmatrix} = \begin{bmatrix} g_1^{\mathrm{T}}(X(t))\theta \\ \vdots \\ g_p^{\mathrm{T}}(X(t))\theta \end{bmatrix} \equiv g(X(t))\theta; \quad t \in [0, 1], \tag{5}$$

where $g(X(t))$ is a known function of $X(t)$. Integrating both sides of (5) gives

$$X(t) = \left\{ \int_0^t g(X(u))\, du \right\} \theta + C, \tag{6}$$

where $C \equiv X(0;\theta)$. The unknown parameter θ^* is estimated by solving

$$\hat{\theta}^{\text{LM}} = \arg\min_{\theta} \int_0^1 \left\| \hat{X}(t;h) - \left\{ \int_0^t g\big(\hat{X}(u;h)\big)\, du \right\} \theta - C \right\|_2^2 dt, \tag{7a}$$

where

$$\hat{X}(\cdot;h) = \arg\min_{Z(\cdot)\in\mathcal{X}(h)} \sum_{i=1}^{n} \|Y_i - Z(t_i)\|_2^2. \tag{7b}$$

The optimization problem (7a) has an analytical solution, given the smoothing estimates from (7b). Compared with the two-step procedure (4), this approach requires an estimate of the integral, $\int_0^t g\big(\hat{X}(u;h)\big)\, du$ in (7a), rather than an estimate of the derivative, $\hat{X}'(t;h)$. This has profound effects on the asymptotic behaviour of the estimator $\hat{\theta}^{\text{LM}}$. $\sqrt{n}$-consistency of $\hat{\theta}^{\text{LM}}$ has been established under mild conditions, and it has been found that the choice of smoothing parameter h is less crucial than for other methods (Gugushvili and Klaassen, 2012).

Recently, Brunel et al. (2014) and Hall and Ma (2014) have considered alternatives to the loss function in (4a). Let $\mathbb{C}^1(0,1)$ be the set of functions that are first-order differentiable on $(0,1)$ and equal zero on the boundary points 0 and 1. Then (1) implies that, for any $\phi \in \mathbb{C}^1(0,1)$,

$$\int_0^1 f(X(t),\theta)\phi(t)dt + \int_0^1 X(t)\phi'(t)dt = 0. \tag{8}$$

Equation (8) is referred to as the *variational formulation* of the ODE. A least squares loss based on (8) takes the form

$$\hat{\theta}^{\mathrm{V}} = \arg\min_{\theta} \frac{1}{L} \sum_{l=1}^{L} \left\| \int_0^1 f\big(\hat{X}(t;h),\theta\big)\phi_l(t)dt + \int_0^1 \hat{X}(t;h)\phi_l'(t)dt \right\|_2^2, \tag{9}$$

where $\hat{X}(t;h)$ is defined in (4b) and $\{\phi_l, l = 1,\ldots,L\}$ is a finite set of functions in $\mathbb{C}^1(0,1)$ (Brunel et al., 2014). In Hall and Ma (2014), the loss function is the sum of the loss functions in (4b) and (9), so that θ and the optimal bandwidth h are estimated simultaneously. It is immediately clear that the derivative $X'(\cdot;\theta)$ is not needed in (9), which can lead to substantial improvement compared to the two-step procedure in (4). A minor drawback of (9) is that the variational formulation (8) is enforced on a finite set of functions $\{\phi_l, l = 1,\ldots,L\}$ rather than on the whole class $\mathbb{C}^1(0,1)$. Under suitable assumptions, the estimator $\hat{\theta}^{\mathrm{V}}$ is $\sqrt{n}$-consistent (Brunel et al., 2014; Hall and Ma, 2014).

2.2.3 The generalized profiling method

Another collocation-based method is the generalized profiling method of Ramsay et al. (2007). Instead of the smoothing estimate $\hat{X}(\cdot;h)$ in (4b), the generalized profiling method uses a smoothing estimate $\check{X}(\cdot;h,\theta)$ that minimizes the weighted sum of a data-fitting loss and a model-fitting loss for any given θ. In greater detail,

$$\hat{\theta}_{\lambda}^{\mathrm{GP}} = \arg\min_{\theta} \sum_{i=1}^{n} \big\| Y_i - \check{X}(t_i;h,\theta) \big\|_2^2, \tag{10a}$$

where

$$\check{X}(\cdot; h, \theta) = \underset{Z(\cdot)\in\mathcal{X}(h)}{\arg\min} \frac{1}{n}\sum_{i=1}^{n} \|Y_i - Z(t_i)\|_2^2 + \lambda \int_0^1 \|Z'(t) - f(Z(t), \theta)\|_2^2 \, dt. \tag{10b}$$

In Ramsay et al. (2007), the authors solve (10a) iteratively for a non-decreasing sequence of λ's in (10b). $\sqrt{n}$-consistency of the limiting estimator was later established by Qi and Zhao (2010). Zhang et al. (2015) proposed a model selection procedure by applying an *ad hoc* lasso procedure (Wang and Leng, 2007) to the estimates from (10).

2.3 Methods that do not assume the form of f

A few authors have recently considered modeling large-scale dynamical systems from biology using ODEs (Henderson and Michailidis, 2014; Wu et al., 2014), under the assumption that the right-hand side of (1) is additive,

$$X_j'(t) = \theta_{j0} + \sum_{k=1}^{p} f_{jk}(X_k(t)), \quad \theta_{j0} \in \mathbb{R}. \tag{11}$$

Henderson and Michailidis (2014) and Wu et al. (2014) approximate the unknown f_{jk} with a truncated basis expansion. Consider a finite basis, $\psi(x) = (\psi_1(x), \ldots, \psi_M(x))^{\mathrm{T}}$, such that

$$f_{jk}(a_k) = \psi(a_k)^{\mathrm{T}}\theta_{jk} + \delta_{jk}(a_k), \quad \theta_{jk} \in \mathbb{R}^M, \tag{12}$$

where $\delta_{jk}(a_k)$ denotes the residual. Using (12), a system of additive ODEs of the form (11) can be written as

$$X_j'(t) = \theta_{j0} + \sum_{k=1}^{p} \psi(X_k(t))^{\mathrm{T}}\theta_{jk} + \sum_{k=1}^{p} \delta_{jk}(X_k(t)), \quad j = 1, \ldots, p. \tag{13}$$

Henderson and Michailidis (2014) and Wu et al. (2014) consider the problem of estimating and selecting the non-zero elements θ_{jk} in (13). Roughly speaking, they propose to solve optimization problems of the form

$$\begin{aligned} \hat{\theta}_j^{\mathrm{NP}} = \underset{\theta_{j0}\in\mathbb{R},\theta_{jk}\in\mathbb{R}^M}{\arg\min} \int_0^1 \left\| \hat{X}_j'(t;h) - \theta_{j0} - \sum_{k=1}^{p} \psi\big(\hat{X}_k(t;h)\big)^{\mathrm{T}}\theta_{jk} \right\|_2^2 dt \\ + \lambda_n \sum_{k=1}^{p} \left[\int_0^1 \{\psi\big(\hat{X}_k(t;h)\big)^{\mathrm{T}}\theta_{jk}\}^2 \, dt \right]^{1/2}, \end{aligned} \tag{14a}$$

for $j = 1, \ldots, p$, where

$$\hat{X}(\cdot;h) = \underset{Z(\cdot)\in\mathcal{X}(h)}{\arg\min} \sum_{i=1}^{n} \|Y_i - Z(t_i)\|_2^2. \tag{14b}$$

In (14a), a standardized group lasso penalty forces all elements in θ_{jk} to be either zero or non-zero when λ_n is large, thereby providing variable selection.

The proposals of Henderson and Michailidis (2014) and Wu et al. (2014) are slightly more involved than (14): an extra ℓ_2-penalty is applied to the θ_{jk}'s in (14a) in Henderson and Michailidis (2014), whereas in Wu et al. (2014) (14a) is followed by tuning parameter selection using Bayesian information criterion (BIC), an adaptive group lasso regression, and a regular lasso. We refer the reader to Henderson and Michailidis (2014) and Wu et al. (2014) for further details.

3. PROPOSED APPROACH

We consider the problem of model selection in high-dimensional ODEs. As in Henderson and Michailidis (2014) and Wu et al. (2014), we assume an additive ODE model (11). We use a finite basis $\psi(\cdot)$ to approximate the additive components f_{jk} as in (12), leading to an ODE system that is linear in the unknown parameters (13). Following the example of Dattner and Klaassen (2015), we exploit this linearity by integrating both sides of (13), which yields

$$X_j(t) = X_j(0) + \theta_{j0}t + \sum_{k=1}^{p} \Psi_k(t)^{\mathrm{T}}\theta_{jk} + \sum_{k=1}^{p} \int_0^t \delta_{jk}(X_k(u))\, du, \tag{15}$$

where $\Psi_k(t)$ denotes the integrated basis such that

$$\Psi_k(t) = (\Psi_{k1}(t), \ldots, \Psi_{kM}(t))^{\mathrm{T}} = \int_0^t \psi(X_k(u))\, du,\ k = 1, \ldots, p, \tag{16}$$

and $\Psi_0(t) = t$. Our method, called *Graph Reconstruction via Additive Differential Equations* (GRADE), then solves the following problem for $j = 1, \ldots, p$:

$$\begin{aligned} \hat{\theta}_j = \underset{C_{j0}\in\mathbb{R},\theta_{j0}\in\mathbb{R},\, \theta_{j1},\ldots,\theta_{jp}\in\mathbb{R}^M}{\arg\min} & \frac{1}{2n}\sum_{i=1}^{n}\left\{Y_{ij} - C_{j0} - \theta_{j0}\hat{\Psi}_0(t_i) - \sum_{k=1}^{p}\theta_{jk}^{\mathrm{T}}\hat{\Psi}_k(t_i)\right\}^2 \\ & + \lambda_{n,j}\sum_{k=1}^{p}\left[\frac{1}{n}\sum_{i=1}^{n}\left\{\theta_{jk}^{\mathrm{T}}\hat{\Psi}_k(t_i)\right\}^2\right]^{1/2}, \end{aligned} \tag{17a}$$

where

$$\hat{X}(\cdot; h) = \underset{Z(\cdot)\in\mathcal{X}(h)}{\arg\min} \sum_{i=1}^{n} \|Y_i - Z(t_i)\|_2^2, \tag{17b}$$

and

$$\hat{\Psi}_0(t) = t;\ \hat{\Psi}_k(t) = \int_0^t \psi(\hat{X}_k(u; h))\, du,\ k = 1, \ldots, p. \tag{17c}$$

In (17a), $\lambda_{n,j}$ is a non-negative sparsity-inducing tuning parameter. We may sometimes use $\lambda_{n,j} \equiv \lambda_n$ for $j = 1, \ldots, p$ for simplicity. If the true function f_{jk}^* in (11) is non-zero, we say that the kth

variable X_k^* is a true regulator of X_j^*. We let $S_j \equiv \{k : \|f_{jk}^*\|_2 \neq 0, k = 1, \ldots, p\}$ denote the set of true regulators. We let the estimated index set of regulators be $\hat{S}_j \equiv \{k : \|\hat{\theta}_{jk}\|_2 \neq 0, k = 1, \ldots, p\}$. We then reconstruct the network using $\hat{S}_j, j = 1, \ldots, p$.

Both (17a) and (17b) can be implemented efficiently using existing software (e.g., Loader, 2013; Meier, 2014). In our theoretical analysis in Section 4, we use local polynomial regression to obtain the smoothing estimate in (17b). We use generalized cross-validation (GCV) on the loss (17b) to select the smoothing tuning parameter h. We use BIC to select the number of bases M for ψ and $\hat{\Psi}$ in (17c), and the sparsity tuning parameter λ_n in (17a).

In some studies, time-course data is collected from multiple samples, or experiments. Let R denote the total number of experiments, and $Y^{(r)}$ the observations in the rth experiment. We assume that the same ODE system (13) applies across all experiments with the same true parameter θ_{jk}^*. We allow a different set of initial values for each experiment. Assume that each experiment consists of measurements on the same set of time points. This leads us to modify (17) as follows:

$$\begin{aligned}\hat{\theta}_j = \underset{C_{j0}^{(r)} \in \mathbb{R}, \theta_{j0} \in \mathbb{R},\, \theta_{j1}, \ldots, \theta_{jp} \in \mathbb{R}^M}{\arg\min} & \frac{1}{2Rn} \sum_{r=1}^{R} \sum_{i=1}^{n} \left\{ Y_{ij}^{(r)} - C_{j0}^{(r)} - \theta_{j0} \hat{\Psi}_0(t_i) - \sum_{k=1}^{p} \theta_{jk}^{\mathrm{T}} \hat{\Psi}_k^{(r)}(t_i) \right\}^2 \\ & + \lambda_n \sum_{k=1}^{p} \left[\frac{1}{Rn} \sum_{r=1}^{R} \sum_{i=1}^{n} \{\theta_{jk}^{\mathrm{T}} \hat{\Psi}_k^{(r)}(t_i)\}^2 \right]^{1/2},\end{aligned} \tag{18}$$

where

$$\hat{X}^{(r)}(\cdot; h) = \underset{Z(\cdot) \in \mathcal{X}(h)}{\arg\min} \sum_{i=1}^{n} \|Y_i^{(r)} - Z(t_i)\|_2^2,\ r = 1, \ldots, R,$$

$$\hat{\Psi}_0(t) = t;\ \hat{\Psi}_k^{(r)}(t) = \int_0^t \psi\big(\hat{X}_k^{(r)}(u; h)\big)\, du,\ k = 1, \ldots, p.$$

In Sections 4, 5.1, and 5.2, we will assume that only one experiment is available, so that our proposal takes the form (17). In Sections 5.3 and 6, we will apply our proposal to data from multiple experiments using (18).

Remark 1. To facilitate the comparison of GRADE (17) with other methods, we introduce an intermediate variable,

$$\tilde{X}_j(t; h, \theta) \equiv C_{j0} + \theta_{j0} t + \sum_{k=1}^{p} \theta_{jk}^{\mathrm{T}} \hat{\Psi}_k(t), \tag{19}$$

following from (15). Plugging (19) into the loss function in (17a) yields $\sum_{i=1}^{n} \left\{ Y_{ij} - \tilde{X}_j(t_i; h, \theta) \right\}^2$. In the gold standard (3), the ODE system (1) is strictly satisfied due to the constraint in (3b). In the two-step procedure (4a) and (14a), the smoothing estimate $\hat{X}(\cdot; h)$ does not satisfy (1). GRADE stands in between: the initial estimate $\hat{X}(\cdot; h)$ in (17b) is solely based on the observations, while the intermediate estimate $\tilde{X}(\cdot; h, \theta)$ is calculated by plugging $\hat{X}(\cdot; h)$ into the additive ODE (13).

4. THEORETICAL PROPERTIES

In this section, we establish variable selection consistency of the GRADE estimator (17). Technical proofs of the statements in this section are available in Section

The proposed method (17) differs from the standard sparse additive model (Ravikumar et al., 2009) in that the regressors $\hat{\Psi}_k(t)$ in (17c) are estimated from smoothing estimates $\hat{X}(\cdot; h)$ (17b) instead of the true trajectories X^* in (2). We use local polynomial regression to compute $\hat{X}(\cdot; h)$ in (17b) (see e.g., Equation 1.67 of Tsybakov, 2009 for details on parameterization). To establish variable selection consistency, it is necessary to obtain a bound for the difference between $\hat{X}(\cdot; h)$ and X^*. This is addressed in Theorem 1. Using the bound in Theorem 1, we then establish variable selection consistency of the estimator in (17) for high-dimensional ODEs in Theorem 2.

In this study, we assume that the measurement errors in (2) are normally distributed. Generalizations to bounded or sub-Gaussian errors are straightforward.

Assumption 1. The measurement errors in (2) are independent, and $\epsilon_{ij} \sim N(0, \sigma^2), i = 1, \ldots, n, j =$

$1, \ldots, p$.

We also require the true trajectories X_j^* in (2) to be smooth.

Assumption 2. Assume that the solutions $X_j^*, 1 \leq j \leq p$, belong to a Hölder class $\Sigma(\beta_1, L_1)$, where $\beta_1 \geq 3$.

In addition, we need some regularity assumptions to hold for the smoothing estimation (17b). These assumptions are common and not crucial to this study, and are hence deferred to Section

Theorem 1. *Suppose that Assumptions 1–2 and*

For the methods outlined in (14) *(Henderson and Michailidis, 2014; Wu et al., 2014), variable selection consistency depends on the convergence of* $\left|\left|\left|\hat{X}' - (X^*)'\right|\right|\right|$ *and* $\left|\left|\left|\hat{X} - X^*\right|\right|\right|$. *In contrast, our method depends only on the convergence rate of* $\left|\left|\left|\hat{X} - X^*\right|\right|\right|$. *It is known that the convergence of* $\left|\left|\left|\hat{X}' - (X^*)'\right|\right|\right|$ *is slower than that of* $\left|\left|\left|\hat{X} - X^*\right|\right|\right|$, *see e.g. Gugushvili and Klaassen (2012). As a result, the rate of convergence of* $\hat{\theta}_{jk}$ *from* (14) *is slower than that of our proposed method* (17).

In order to establish the main result, we need the following additional assumptions. Recall the definition of $\Psi_j(t)$ *from* (16)*; for convenience, we suppress the dependence of* $\Psi(t)$ *on* t *in what follows.*

Assumption 3. For $j = 1, \ldots, p$, $(X_j^*)'$ is an additive function of X_k^*, $k = 1, \ldots, p$. In other words,

$$\big(X_j^*\big)'(t) = \theta_{j0}^* + \sum_{k=1}^{p} f_{jk}^*\big(X_k^*(t)\big), \quad \theta_{j0}^* \in \mathbb{R},\ j = 1, \ldots, p, \tag{20}$$

where $\int_0^1 f_{jk}^*\big(X_k^*(t)\big)dt = 0$ for all j, k. Furthermore, the functions f_{jk}^* $(1 \leq j, k \leq p)$ belong to a Sobolev class $W(\beta_2, L_2)$ on a finite interval with $\beta_2 \geq 3$.

Assumption 4. The eigenvalues of $\int_0^1 \Psi_{S_j^0}\Psi_{S_j^0}^{\mathrm{T}}\,dt$ are bounded from above by $C_{\max}$ and bounded from below by a positive number $C_{\min}$, and for $k \notin S_j^0$, the eigenvalues of $\int_0^1 \Psi_k\Psi_k^{\mathrm{T}}\,dt$ are bounded from below by $C_{\min}$. In other words,

$$0 < C_{\min} \leq \Lambda_{\min}\left(\int_0^1 \Psi_{S_j^0}\Psi_{S_j^0}^{\mathrm{T}}\,dt\right) \leq \Lambda_{\max}\left(\int_0^1 \Psi_{S_j^0}\Psi_{S_j^0}^{\mathrm{T}}\,dt\right) \leq C_{\max}, \tag{21}$$

and

$$C_{\min} \leq \Lambda_{\min}\left(\int_0^1 \Psi_k\Psi_k^{\mathrm{T}}\,dt\right), \quad \text{for} \quad k \notin S_j^0. \tag{22}$$

Assumption 5. Assume that

$$\max_{k\notin S_j^0}\left\|\left(\int_0^1 \Psi_k\Psi_{S_j^0}^{\mathrm{T}}\,dt\right)\left(\int_0^1 \Psi_{S_j^0}\Psi_{S_j^0}^{\mathrm{T}}\,dt\right)^{-1}\right\|_2 \leq \xi. \tag{23}$$

The first part of Assumption 4 ensures identifiability among the $s_j + 1$ elements in the set $\{t, X^*_{S_j}\}$, and the second part ensures that Ψ_k is non-degenerate for $k \notin S_j^0$. Assumption 5 restricts the association between the elements in the set $\{t, X^*_{S_j}\}$ and the elements in the set $X^*_{S_j^c}$. Note that in order for the parameters in an additive model such as (13) *to be identifiable, there must be no concurvity among the variables (Buja et al., 1989). This is guaranteed by Assumptions 4 and 5, which appear often in the literature of lasso regression (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Ravikumar et al., 2009; Wainwright, 2009; Lee et al., 2013). We refer the readers to Miao et al. (2011) for a detailed discussion of the identifiability of the parameters in an ODE model.*

The next assumption characterizes the relationships between the quantities in Assumptions 4 and 5 and the sparsity tuning parameter λ_n in (17a)*. Similar assumptions have been made in*

lasso-type regression (Meinshausen and Bühlmann, 2006; Zhao and Yu, 2006; Ravikumar et al., 2009; Wainwright, 2009; Lee et al., 2013).

Assumption 6. Assume that

$$f_{\min} > \lambda_n \frac{4\sqrt{2sC_{\max}}}{C_{\min}} \quad \text{and} \quad \xi < \frac{1}{4}\sqrt{\frac{C_{\min}}{sC_{\max}}},$$

where $f_{\min} \equiv \min_{k \in S_j} \left\{ \int_0^1 \left[f_{jk}^*(X_k^*(t)) \right]^2 dt \right\}^{1/2}$ is the minimum regulatory effect.

Furthermore, we impose some regularity conditions on the bases $\psi(\cdot)$; these are deferred to Assumption

We arrive at the following theorem.

Theorem 2. *Suppose that Assumptions 1–6 and*

Because the regressors $\hat{\Psi}$ are estimated, establishing variable selection consistency requires extra attention. To prove Theorem 2, we must first establish variable selection consistency of group lasso regression with errors in variables. This generalizes the recent work on errors in variables for lasso regression (Loh and Wainwright, 2012). Theorem 2 ensures that the proposed method is able to recover the true graph exactly, given sufficiently dense observations in a finite time interval if the graph is sparse. The number of variables in the system can grow exponentially fast with respect to n, which means that the result holds for the "large p, small n" scenario.

Theorem 2 does not provide us with practical guidance for selecting the bandwidth h_n for the local polynomial regression estimator $\hat{X}_j$. The next result mirrors Theorem 2 for the bandwidths selected by cross-validation or GCV*, which converge to $h_n \propto n^{-1/(2\beta_1+1)}$ asymptotically (see Xia and Li, 2002; Tsybakov, 2009 for details).*

Proposition 1. *Suppose that Assumptions 1–6 and*

We note that selecting the values of M and λ_n that yield the rate specified in Proposition 1 is challenging in practice. The rate of convergence of the sparsity tuning parameter λ_n is slower in Proposition 1 compared to Theorem 2. This results in an increase in the minimum regulatory effect $f_{\min}$ because of the relation between $f_{\min}$ and λ_n in Assumption 6.

5. NUMERICAL EXPERIMENTS

We study the empirical performance of our proposal in three different scenarios in the following subsections. In what follows, given a set of initial conditions and a system of ODEs, numerical solutions of the ODEs are obtained using the Euler method with step size 0.001. *Observations are drawn from the solutions at an evenly-spaced time grid* $\{iT/n; i = 1, \ldots, n\}$ *with independent* $N(0, 1)$ *measurement errors, unless specified otherwise. To facilitate the comparison of GRADE with other methods, we fit the smoothing estimates* $\hat{X}$ *in* (17b) *using smoothing splines with bandwidth chosen by* GCV. *We use cubic splines with two internal knots as the basis functions in* (17c) *in Sections 5.1 and 5.3. Linear basis functions are used in Section 5.2. The integral* $\hat{\Psi}_k(t) = \int_0^t \psi\big(\hat{X}_k(u; h)\big)\, du$ *in* (17c) *is calculated numerically with step size* 0.01.

5.1 Variable selection in additive ODEs

In this simulation, we compare GRADE with NeRDS (Henderson and Michailidis, 2014) and SA-ODE (Wu et al., 2014) described in (14). *We consider the following system of additive ODEs, for*

$k = 1, \ldots, 5$*:*

$$\begin{cases} X'_{2k-1}(t) = \theta_{2k-1,0} + \psi(X_{2k-1}(t))^{\mathrm{T}}\theta_{2k-1,2k-1} + \psi(X_{2k}(t))^{\mathrm{T}}\theta_{2k-1,2k} \\ X'_{2k}(t) = \theta_{2k,0} + \psi(X_{2k-1}(t))^{\mathrm{T}}\theta_{2k,2k-1} + \psi(X_{2k}(t))^{\mathrm{T}}\theta_{2k,2k} \end{cases}, t \in [0, 20], \quad (24)$$

where $\psi(x) = (x, x^2, x^3)^{\mathrm{T}}$ *is the cubic monomial basis. The parameters and initial conditions are chosen so that the solution trajectories are identifiable under an additive model (Buja et al., 1989). Detailed specification of* (24) *can be found in Section*

After generating data according to (24) *and introducing noise, we apply GRADE, NeRDS, and SA-ODE to recover the directed graph encoded in* (24)*. Both NeRDS and SA-ODE are implemented using code provided by the authors. NeRDS and SA-ODE use smoothing splines to estimate* $\hat{X}$ *and* $\hat{X}'$ *in* (14b)*, and cubic splines with two internal knots as the basis* ψ *in* (14a)*. As mentioned briefly in Section 2, NeRDS applies an additional smoothing penalty which amounts to an* ℓ_2 *penalty on* θ_{jk} *in* (14a)*, controlled by a parameter selected using* GCV *(Henderson and Michailidis, 2014). We apply GRADE using the same smoothing estimates and basis functions as NeRDS and SA-ODE. To facilitate a direct comparison to NeRDS, we apply GRADE both with and without an additional* ℓ_2*-type penalty on the* θ_{jk}*'s in* (17a)*. We apply all methods for a range of values of the sparsity-inducing tuning parameter (e.g.,* λ_n *in* (17a)*), in order to yield a recovery curve of varying sparsity.*

We summarize the simulation results in Figure 1, where the numbers of true edges selected are displayed against the total numbers of selected edges over a range of sparsity tuning parameters. We see that GRADE outperforms the other two methods, which corroborates our theoretical findings in Section 4 that our proposed method is more efficient than methods such as NeRDS and SA-ODE which involve derivative estimation (see e.g., comments below Theorem 1).

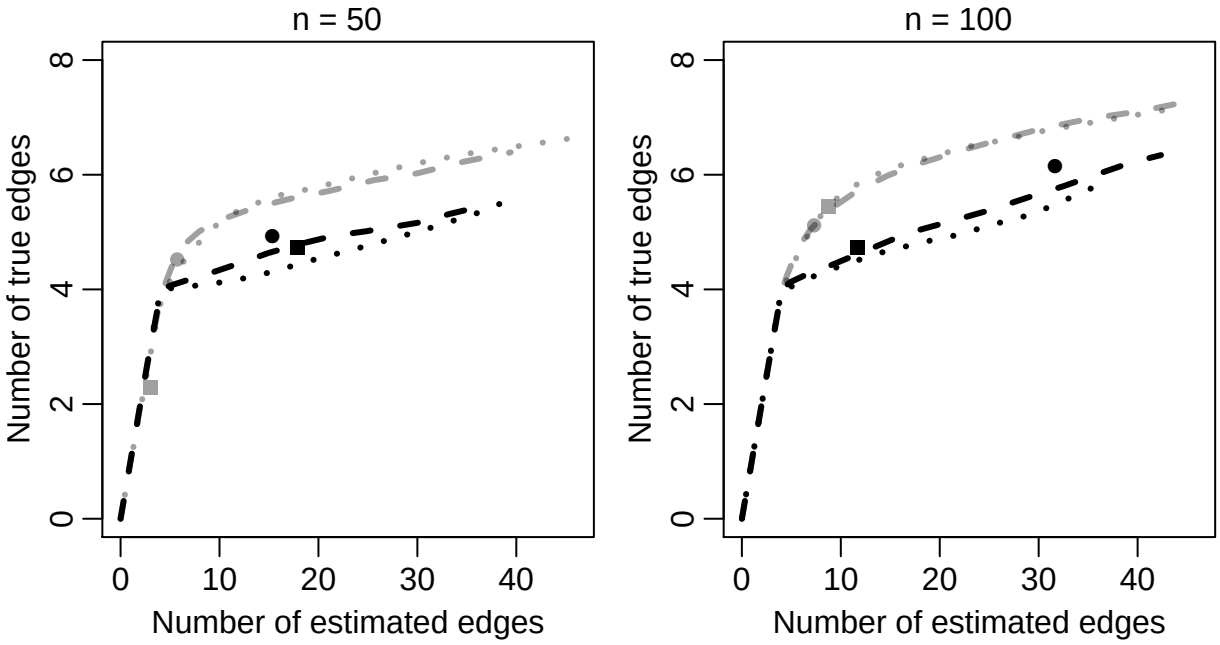

Figure 1: Performance of network recovery methods on the system of additive ODEs in (24), averaged over 400 simulations. The four curves represent SA-ODE (⋯), NeRDS (- -), and GRADE without (⋯) and with (- -) the additional smoothing penalty in (17a) used by NeRDS. Each point on the curves corresponds to average performance for a given sparsity tuning parameter λ_n in (14a) or (17a). The symbols indicate the sparsity tuning parameter λ_n selected using BIC (SA-ODE, ▪, and GRADE, ▪ and •) or GCV (NeRDS, •).

5.2 Variable selection in linear ODEs

In this simulation, we compare GRADE to two recent proposals by Brunel et al. (2014) and Hall and Ma (2014). Recall from Section 2.2.2 that Brunel et al. (2014) and Hall and Ma (2014) are proposed to estimate a few unknown parameters in an ODE system of known form. Hence, we consider a simple linear ODE system, for $k = 1, \ldots, 4$*,*

$$\begin{cases} X'_{2k-1}(t) = 2k\pi X_{2k}(t) \\ \\ X'_{2k}(t) = -2k\pi X_{2k-1}(t) \end{cases}, t \in [0, 1]. \tag{25}$$

For each $k = 1, \ldots, 4$*, we set the initial condition to be* $(X_{2k-1}(0), X_{2k}(0)) = (\sin(y_k), \cos(y_k))$ *where* $y_k \sim N(0, 1)$*. The solutions to* (25) *take the form of sine and cosine functions of frequencies ranging from* 2π *to* 8π*. The graph corresponding to* (25) *is sparse, with only eight directed edges out of* 64 *possible edges. We fit the model*

$$X'(t) = \Theta X(t) + C, \tag{26}$$

where Θ *is an unknown* 8×8 *matrix and* C *is an* 8*-vector. We apply the method in Brunel et al. (2014) using the code provided by the authors. We implement the method in Hall and Ma (2014) in R based on the authors' code in Fortran. Because the loss function in Hall and Ma (2014) is not convex, we use five sets of random initial values and report the best performance. Since both Brunel et al. (2014) and Hall and Ma (2014) yield dense estimates for* Θ *in* (26)*, in order to examine how well these methods recover the true graph, we threshold the estimates at a range of values in order to obtain a variable selection path. We apply GRADE using the linear basis function* $\psi(x) = x$*.*

Results are shown in Figure 2. We can see that GRADE outperforms the methods in Brunel et al. (2014) and Hall and Ma (2014). This is likely due to the fact that GRADE exploits the sparsity of the true graph with a sparsity-inducing penalty. In principle, Brunel et al. (2014) and Hall and Ma (2014) could be generalized in order to include penalties on the parameters. We leave this to future research.

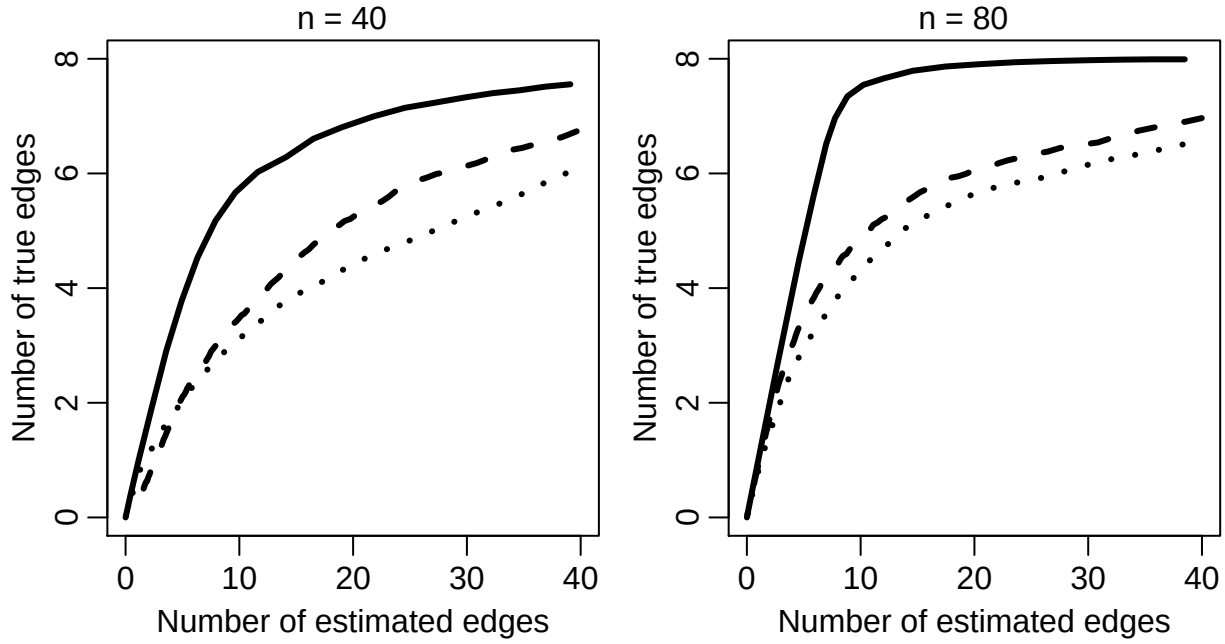

Figure 2: Network recovery on the system of linear ODEs (25), averaged over 200 simulated data sets. The three curves represent GRADE (—), Hall and Ma (2014) (····), Brunel et al. (2014) (- -).

5.3 Robustness of GRADE to the additivity assumption

The GRADE method assumes that the true underlying model is additive (Assumption 3). However, in many systems, the additivity assumption is violated; for instance, multiplicative effects may be present in gene regulatory networks (Ma et al., 2009). In this subsection, we investigate the performance of GRADE in a setting where the true model is non-additive. We consider the following system of ODEs, for $k = 1, \ldots, 5$*,*

$$\begin{cases} X'_{2k-1}(t) = f_{2k-1}\left(X_{2k-1}(t), X_{2k}(t)\right) \equiv 2X_{2k-1}(t) - vX_{2k-1}(t)X_{2k}(t) \\ X'_{2k}(t) = f_{2k}\left(X_{2k-1}(t), X_{2k}(t)\right) \equiv vX_{2k-1}(t)X_{2k}(t) - 2X_{2k}(t) \end{cases}, t \in [0, 5], \quad (27)$$

where v *is a positive constant. For each* $k = 1, \ldots, 5$*, the pair of equations* (27) *is a special case of the Lotka-Volterra equations (Volterra, 1928), which represent the dynamics between predators*

(X_{2k}) and prey (X_{2k-1}). The parameter v defines the interaction between the two populations. For $v \neq 0$, both X'_{2k-1} and X'_{2k} are non-additive functions of X_{2k-1} and X_{2k}. We define two types of directed edges, where $\mathcal{E}_1 \equiv \{(X_j, X_j), j = 1, \ldots, 10\}$ and $\mathcal{E}_2 \equiv \{(X_{2k-1}, X_{2k}), (X_{2k}, X_{2k-1}),\ k = 1, \ldots, 5\}$ represent the self-edges and non-self-edges, respectively. Figure 3(a) contains an illustration of the graph and edge types for each pair of equations. In what follows, we investigate how well GRADE recovers these two types of edges as we change the parameter v, i.e., as the additivity assumption is violated.

Since measurement error is not essential to the current discussion, we generate data according to (27) *without adding noise. To ensure that the trajectories are identifiable, we generate $R = 2$ sets of random initial values drawn from $N_{10}(0, 2I_{10})$, where I_{10} is a 10×10 identity matrix. In order to quantify the amount of signal in an edge that GRADE can detect, we introduce the quantity*

$$D_{j,k}(v) = \mathbb{E}\left[R \int_0^T \left\{\frac{\partial f_j}{\partial X_k}\left(t; X(0)\right)\right\}^2 dt\right], \tag{28}$$

where the expectation is taken with respect to the random initial values $X(0)$ and R is the number of initial values. The measure $D_{j,k}$ in (28) *is a loose analogy to $\left\{\int_0^1 \left[f^*_{jk}(X^*_k(t))\right]^2 dt\right\}^{1/2}$ used in Assumption 6. Note that if no edge is present from X_k to X_j, then $\partial f_j / \partial X_k \equiv 0$ and hence $D_{j,k}(v) = 0$. One immediately notes that, as R increases, the regulatory effect for a true edge increases proportionally to R, while the regulatory effect of a non-edge remains zero. For the self-edges in $\mathcal{E}_1$ and the non-self-edges in $\mathcal{E}_2$, we can define $D^{(1)}(v)$ and $D^{(2)}(v)$ as*

$$D^{(1)}(v) = \min_{k=1,\ldots,10} D_{k,k}(v), \quad \textit{and} \quad D^{(2)}(v) = \min_{k=1,\ldots,5} \{D_{2k-1,2k}(v), D_{2k,2k-1}(v)\}, \tag{29}$$

where we use the minimum because variable selection is limited by the minimum regulatory effect (see Assumption 6). With a slight abuse of definition, we refer to (29) *as the minimum regulatory effects in a non-additive model.*

We apply GRADE using the formulation in (18)*. The sparsity parameter* λ *is chosen so that there are* 20 *directed edges in the estimated network. We record the number of estimated edges that are in* $\mathcal{E}_1$ *and* $\mathcal{E}_2$*. The edge recovery performance is shown in Figure 3(b). In Figure 3(c), we display the minimum regulatory effects defined in* (29)*. Edge recovery and minimum regulatory effects show a similar trend as a function of* r *in* (27)*. This suggests that* (29)*, and thus* (28)*, is a reasonable measure of the additive components of the regulatory effect of the edges. The slight deviation between the trends reflects the fact that the measure defined in* (28) *is not a direct counterpart of* $\left\{\int_0^1 \left[f_{jk}^*(X_k^*(t))\right]^2 dt\right\}^{1/2}$ *in a non-additive model. The edge recovery improves when a larger value of* R *is used, though these results are omitted due to space constraints. Our results indicate that GRADE can recover the true graph even when the additivity assumption is violated, provided that the regulatory effects* (28) *for the true edges are sufficiently large.*

6. APPLICATIONS

6.1 Application to *in silico* gene expression data

GeneNetWeaver (GNW) provides an in silico benchmark for assessing the performance of network recovery methods (Schaffter et al., 2011), and was used in the third DREAM challenge (Marbach et al., 2009). GNW is based upon real gene regulatory networks of yeast and E. coli. It extracts sub-networks from the yeast or E. coli gene regulatory networks, and assigns a system of ODEs to the extracted network. This system of ODEs is non-additive, and includes unobserved variables

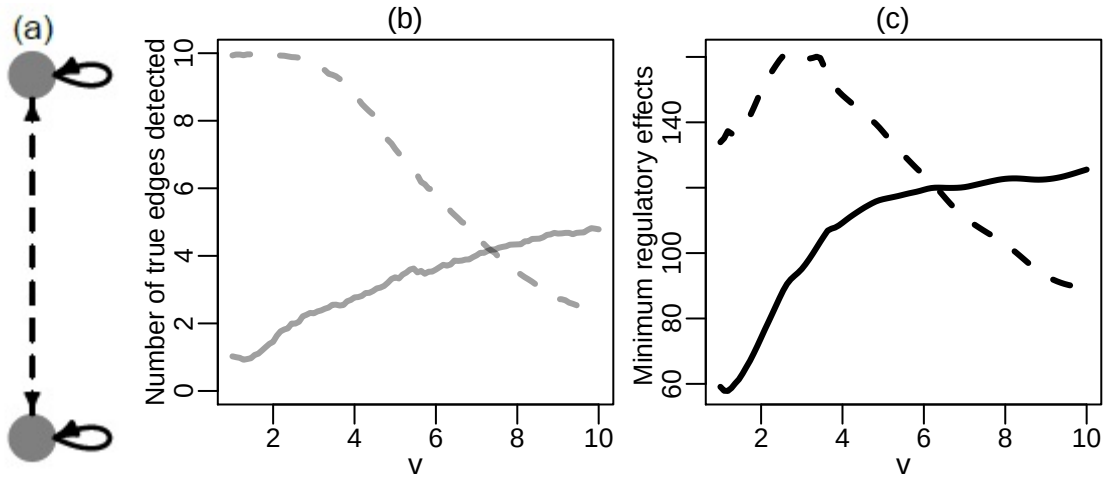

Figure 3: (a): The graph encoded by a pair of Lotka-Volterra equations as given in (27). Self-edges (—) and non-self-edges (- -) are shown. (b): Self-edge (—) and non-self-edge (- -) recovery of GRADE, averaged over 200 simulated data sets. (c): Minimum signals defined in (29), for self-edges, $D^{(1)}(\cdot)$ (—), and non-self-edges, $D^{(2)}(\cdot)$ (- -).

(Marbach et al., 2010). Therefore, the assumptions of GRADE are violated in the GNW data.

To mimic real-world laboratory experiments, GNW provides several data generation mechanisms. In this study, we consider data from the perturbation experiments. The perturbation experiments are similar to the data generating mechanisms used in Section 5.3, where initial conditions of the ODE system are perturbed in order to emulate the diversity of trajectories from multiple independent experiments.

We investigate ten networks from GNW that have been previously studied in Henderson and Michailidis (2014), of which five have 10 *nodes and five have* 100 *nodes. For each network, GNW provides one set of noiseless gene expression data consisting of* R *perturbation experiments where the trajectories are measured at* $n = 21$ *evenly-spaced time points in* $[0, 1]$*. Here* $R = 10$ *for the five 10-node networks and* $R = 100$ *for the five 100-node networks. As in Henderson and Michailidis (2014), we add independent* $N\left(0, 0.025^2\right)$ *measurement errors to the data at each timepoint.*

We apply NeRDS as described in Henderson and Michailidis (2014). We apply GRADE using the formulation (18) *to handle observations from multiple experiments, with the smoothing estimates* $\hat{X}$ *in* (17b) *fit using smoothing splines with bandwidth chosen by* GCV*, and using cubic splines with two internal knots as the basis functions in* (17c)*. The integral* $\hat{\Psi}_k(t) = \int_0^t \psi\big(\hat{X}_k(u;h)\big)\,du$ *in* (17c) *is calculated numerically with step size* 0.01*. Finally, we apply an additional* ℓ_2*-type penalty to the* θ_{jk}*'s in* (18) *in order to match the setup of NeRDS. The tuning parameter for this penalty is set to be* 0.1*.*

Results are shown in Table 1. Recall that the data generating mechanism violates crucial assumptions for both NeRDS and GRADE. We see in Table 1 that NeRDS outperforms GRADE in one network, while GRADE outperforms NeRDS in the other nine networks. This suggests that GRADE is a competitive exploratory tool for reconstructing gene regulatory networks.

Table 1: Area Under ROC Curves for NeRDS and GRADE

	$p = 10$		$p = 100$	
	NeRDS	GRADE	NeRDS	GRADE
Ecoli1	$0.450\ (0.438, 0.462)$	$\mathbf{0.545}\ (0.534, 0.557)$	$0.624\ (0.622, 0.627)$	$\mathbf{0.670}\ (0.667, 0.673)$
Ecoli2	$0.512\ (0.502, 0.523)$	$\mathbf{0.643}\ (0.634, 0.653)$	$0.637\ (0.635, 0.640)$	$\mathbf{0.653}\ (0.650, 0.656)$
Yeast1	$0.486\ (0.476, 0.495)$	$\mathbf{0.679}\ (0.666, 0.691)$	$0.610\ (0.607, 0.612)$	$\mathbf{0.636}\ (0.635, 0.638)$
Yeast2	$0.525\ (0.518, 0.532)$	$\mathbf{0.607}\ (0.600, 0.613)$	$0.568\ (0.566, 0.569)$	$\mathbf{0.584}\ (0.582, 0.585)$
Yeast3	$0.467\ (0.460, 0.474)$	$\mathbf{0.576}\ (0.566, 0.587)$	$\mathbf{0.617}\ (0.616, 0.619)$	$0.567\ (0.566, 0.568)$

The average area under the curves and 90% confidence intervals, over 100 simulated data sets. Networks and data generating mechanisms are described in Section 6.1. Boldface indicates the method with larger AUC.

6.2 Application to calcium imaging recordings

In this section, we consider the task of learning regulatory relationships among populations of neurons. We investigate the calcium imaging recording data from the Allen Brain Observatory project

conducted by the Allen Institute for Brain Science[1]*. Here, we investigate one of the experiments in the project. In this experiment, calcium fluorescence levels (a surrogate for neuronal activity) are recorded at* 30 *Hz on a region of the primary visual cortex while the subject mouse is shown forty visual stimuli. The forty visual stimuli are combinations of eight spatial orientations and five temporal frequencies. Each stimulus lasts for two seconds and is repeated 15 times. The recorded videos are processed by the Allen Institute to identify individual neurons. In this particular experiment, there are* 575 *neurons. Each neuron's activity is defined as the average calcium fluorescence level of the pixels that it covers in the video.*

It is known that the activities of individual neurons are noisy and sometimes misleading (Cunningham and Byron, 2014). As an alternative, neuronal populations can be studied (see e.g., Part Three of Gerstner et al., 2014). We define 25 neuronal populations by dividing the recording region into a 5×5 *grid, where each population contains roughly 20 neurons. We use GRADE to capture the functional connectivity among the* 25 *neuronal populations. Note that functional connectivity is distinct from physical connectivity. Functional connectivity involves the relationships among neuronal populations that can be observed through neuron activities and may change across stimuli, whereas physical connectivity consists of synaptic interactions.*

We estimate the functional connectivity corresponding to three different but related stimuli, consisting of frequencies of 1 *Hz,* 2 *Hz, and* 4 *Hz, each at a spatial orientation of* $90°$*. For each stimulus, we have calcium fluorescence levels of the* $p = 25$ *neuronal populations for each of* $R = 15$ *repetitions. Since each repetition spans two seconds and the calcium fluorescence is recorded at* 30 *Hz, there are* 60 *timepoints per repetition. We apply GRADE using the formulation*

[1]Website: ©2016 Allen Institute for Brain Science. Allen Brain Observatory [Internet]. Available from: `http://observatory.brain-map.org`.

in (18) *in order to reconstruct the functional connectivity under each of the three stimuli. We use smoothing splines with bandwidth* h *selected with* GCV *in order to estimate* $\hat{X}$ *in* (17b)*, and use cubic splines with* 4 *internal knots as the basis functions* $\psi(\cdot)$ *in* (17c)*. The sparsity parameter* $\lambda_{j,n}$ *for each nodewise regression in* (18) *is selected using* BIC *for each* $j = 1, \ldots, 25$*. For ease of visualization, we prefer a sparse network, and so we fit GRADE using tuning parameter values* $\alpha(\lambda_{1,n}, \ldots, \lambda_{p,n})$*, where the scalar* α *is selected so that each of the estimated networks contains approximately 25 edges.*

Estimated functional connectivities are shown in Figure 4. We see that, in all three networks, the 24*th neuronal population regulates many other neuronal populations, indicating that this region may contain neurons that are sensitive to this spatial orientation. Furthermore, we see that the adjacent connectivity networks in Figure 4 are somewhat similar to each other, whereas the networks at* 1 *Hz and* 4 *Hz have few similarities. This agrees with the observation in neuroscience that neurons in the mouse primary visual cortex are responsive to a somewhat narrow range of temporal frequencies near their peak frequencies (see, e.g., Gao et al., 2010).*

7. DISCUSSION

In this paper, we propose a new approach, GRADE, for estimating a system of high-dimensional additive ODEs. GRADE involves estimation of an integral rather than a derivative. We show that estimating the integral is superior to estimating the derivatives both theoretically and empirically. We leave an extension of our work to non-additive ODEs to future research.

In this paper, we have not addressed the issue of experimental design. Given a finite set of resources, one may choose to design an experiment to measure n *observations on a very dense time grid, or on a coarse time grid. Alternatively, one might choose to measure* n/R *observations for*

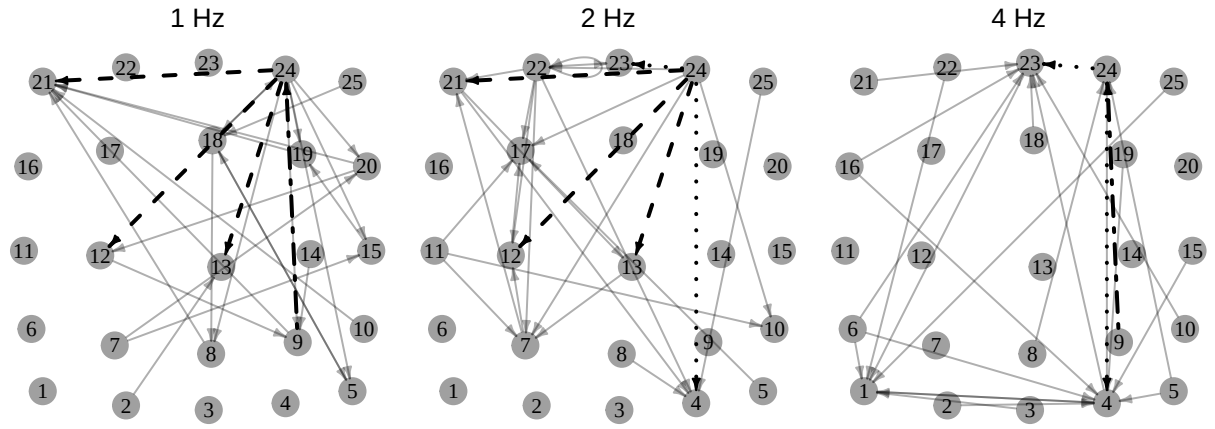

Figure 4: Estimated functional connectivities among neuronal populations from the calcium imaging data described in Section 6.2. Each node is positioned near the center of the neuronal population it represents, with jitter added for ease of display. The three edges are shared between the estimated networks at 1 Hz and 2 Hz; the two edges are shared between estimated networks at 2 Hz and 4 Hz; the single edge is shared between the estimated networks at 1 Hz and 4 Hz. For reference, given two Erdös-Rènyi graphs consisting of 25 nodes and 25 edges, the probability of having three or more shared edges is 0.07, and the probability of having two or more shared edges is 0.26.

R distinct experiments from a single ODE system (1)*, each with a different initial condition. This presents a trade-off that is especially interesting in the context of ODEs: using a dense time grid improves the quality of the smoothing estimates $\hat{X}$, as seen in Sections 5.1 and 5.2, while running multiple experiments enhance the identifiability of the true structure, as seen in Section 5.3. We leave a more detailed treatment of these issues to future work.*

8. SUPPLEMENTARY MATERIALS

Supplementary Material: *The supplementary material contains proofs and details on data generation used in the main paper. (pdf file)*

References

Benson, M. (1979). Parameter fitting in dynamic models. Ecological Modelling 6*(2), 97 – 115.*

Biegler, L. T., J. J. Damiano, and G. E. Blau (1986). Nonlinear parameter estimation: A case study comparison. AIChE Journal 32*(1), 29–45.*

Brunel, N. J.-B. (2008). Parameter estimation of ODE's via nonparametric estimators. Electron. J. Stat. 2*, 1242–1267.*

Brunel, N. J.-B., Q. Clairon, and F. d'Alché Buc (2014). Parametric estimation of ordinary differential equations with orthogonality conditions. J. Amer. Statist. Assoc. 109*(505), 173–185.*

Buja, A., T. J. Hastie, and R. J. Tibshirani (1989). Linear smoothers and additive models. Ann. Statist. 17*(2), 453–555.*

Cao, J., L. Wang, and J. Xu (2011). Robust estimation for ordinary differential equation models. Biometrics 67*(4), 1305–1313.*

Cao, J. and H. Zhao (2008). Estimating dynamic models for gene regulation networks. Bioinformatics 24*(14), 1619–1624.*

Chou, I.-C. and E. O. Voit (2009). Recent developments in parameter estimation and structure identification of biochemical and genomic systems. Math. Biosci. 219*(2), 57–83.*

Cunningham, J. P. and M. Y. Byron (2014). Dimensionality reduction for large-scale neural recordings. Nature neuroscience 17*(11), 1500–1509.*

Dattner, I. and C. A. J. Klaassen (2015). Optimal rate of direct estimators in systems of ordinary differential equations linear in functions of the parameters. Electron. J. Stat. 9*(2), 1939–1973.*

Ellner, S. P., Y. Seifu, and R. H. Smith (2002). Fitting population dynamic models to time-series data by gradient matching. Ecology 83*(8), 2256–2270.*

Friedman, J. H., T. J. Hastie, and R. J. Tibshirani (2008). Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9*(3), 432–441.*

Gao, E., G. C. DeAngelis, and A. Burkhalter (2010). Parallel input channels to mouse primary visual cortex. The Journal of neuroscience 30*(17), 5912–5926.*

Gerstner, W., W. M. Kistler, R. Naud, and L. Paninski (2014). Neuronal dynamics: From single neurons to networks and models of cognition. *Cambridge University Press.*

Gugushvili, S. and C. A. J. Klaassen (2012). $\sqrt{n}$-consistent parameter estimation for systems of ordinary differential equations: bypassing numerical integration via smoothing. Bernoulli 18*(3), 1061–1098.*

Hall, P. and Y. Ma (2014). Quick and easy one-step parameter estimation in differential equations. J. R. Stat. Soc. Ser. B. Stat. Methodol. 76*(4), 735–748.*

Henderson, J. and G. Michailidis (2014). Network reconstruction using nonparametric additive ode models. PLoS ONE 9*(4), e94003.*

Izhikevich, E. M. (2007). Dynamical systems in neuroscience: the geometry of excitability and bursting. *Computational Neuroscience. MIT Press, Cambridge, MA.*

Lee, J. D., Y. Sun, and J. E. Taylor (2013). On model selection consistency of penalized M-estimators: a geometric theory. In C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger (Eds.), Advances in Neural Information Processing Systems 26*, pp. 342–350. Curran Associates, Inc.*

Liang, H. and H. Wu (2008). Parameter estimation for differential equation models using a framework of measurement error in regression models. J. Amer. Statist. Assoc. 103*(484), 1570–1583.*

Loader, C. (2013). locfit: Local Regression, Likelihood and Density Estimation. *R package version 1.5-9.1.*

Loh, P.-L. and M. J. Wainwright (2012). High-dimensional regression with noisy and missing data: provable guarantees with nonconvexity. Ann. Statist. 40*(3), 1637–1664.*

Lu, T., H. Liang, H. Li, and H. Wu (2011). High-dimensional ODEs coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. J. Amer. Statist. Assoc. 106*(496), 1242–1258.*

Ma, W., A. Trusina, H. El-Samad, W. A. Lim, and C. Tang (2009). Defining network topologies that can achieve biochemical adaptation. Cell 138*(4), 760–773.*

Marbach, D., R. J. Prill, T. Schaffter, C. Mattiussi, D. Floreano, and G. Stolovitzky (2010). Revealing strengths and weaknesses of methods for gene network inference. Proceedings of the National Academy of Sciences 107*(14), 6286–6291.*

Marbach, D., T. Schaffter, C. Mattiussi, and D. Floreano (2009). Generating realistic in silico gene networks for performance assessment of reverse engineering methods. Journal of Computational Biology 16*(2), 229–239.*

Meier, L. (2014). grplasso: Fitting user specified models with group lasso penalty. *R package version 0.4-4.*

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. Ann. Statist. 34*(3), 1436–1462.*

Meinshausen, N. and P. Bühlmann (2010). Stability selection. J. R. Stat. Soc. Ser. B Stat. Methodol. 72*(4), 417–473.*

Miao, H., X. Xia, A. Perelson, and H. Wu (2011). On identifiability of nonlinear ODE models and applications in viral dynamics. SIAM Rev. 53*(1), 3–39.*

Qi, X. and H. Zhao (2010). Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations. Ann. Statist. 38*(1), 435–481.*

Ramsay, J. O., G. Hooker, D. Campbell, and J. Cao (2007). Parameter estimation for differential equations: a generalized smoothing approach. J. R. Stat. Soc. Ser. B Stat. Methodol. 69*(5), 741–796. With discussions and a reply by the authors.*

Ravikumar, P. K., J. Lafferty, H. Liu, and L. Wasserman (2009). Sparse additive models. J. R. Stat. Soc. Ser. B Stat. Methodol. 71*(5), 1009–1030.*

Schaffter, T., D. Marbach, and D. Floreano (2011). Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. Bioinformatics 27*(16), 2263–2270.*

Tibshirani, R. J. (1996). Regression shrinkage and selection via the lasso. J. Roy. Statist. Soc. Ser. B 58*(1), 267–288.*

Tsybakov, A. B. (2009). Introduction to nonparametric estimation. *Springer Series in Statistics. Springer, New York. Revised and extended from the 2004 French original, Translated by Vladimir Zaiats.*

Varah, J. M. (1982). A spline least squares method for numerical parameter estimation in differential equations. SIAM J. Sci. Statist. Comput. 3*(1), 28–46.*

Volterra, V. (1928). Variations and fluctuations of the number of individuals in animal species living together. J. Cons. Int. Explor. Mer 3*(1), 3–51.*

Voorman, A. L., A. Shojaie, and D. M. Witten (2014). Graph estimation with joint additive models. Biometrika 101*(1), 85–101.*

Wainwright, M. J. (2009). Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). IEEE Trans. Inform. Theory 55*(5), 2183–2202.*

Wang, H. and C. Leng (2007). Unified LASSO estimation by least squares approximation. J. Amer. Statist. Assoc. 102*(479), 1039–1048.*

Wu, H. (2005). Statistical methods for HIV dynamic studies in AIDS clinical trials. Stat. Methods Med. Res. 14*(2), 171–192.*

Wu, H., T. Lu, H. Xue, and H. Liang (2014). Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. J. Amer. Statist. Assoc. 109*(506), 700–716.*

Xia, Y. and W. K. Li (2002). Asymptotic behavior of bandwidth selected by the cross-validation method for local polynomial fitting. J. Multivariate Anal. 83*(2), 265–287.*

Xue, H., H. Miao, and H. Wu (2010). Sieve estimation of constant and time-varying coefficients in nonlinear ordinary differential equation models by considering both numerical error and measurement error. Ann. Statist. 38*(4), 2351–2387.*

Yuan, M. and Y. Lin (2006). Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B Stat. Methodol. 68*(1), 49–67.*

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

Zhang, X., J. Cao, and R. J. Carroll (2015). On the selection of ordinary differential equation models with application to predator-prey dynamical models. Biometrics 71*(1), 131–138.*

Zhao, P. and B. Yu (2006). On model selection consistency of Lasso. J. Mach. Learn. Res. 7*, 2541–2563.*