JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 117, NO. 540, 1711–1725: Theory and Methods
https://doi.org/10.1080/01621459.2021.1882466

Kernel Ordinary Differential Equations

Xiaowu Dai[a] and Lexin Li[a]

[a]Department of Economics and Simons Institute for the Theory of Computing, the University of California, Berkeley, Berkeley, CA; [b]Department of Biostatistics and Epidemiology, the University of California, Berkeley, Berkeley, CA

ABSTRACT

Ordinary differential equation (ODE) is widely used in modeling biological and physical processes in science. In this article, we propose a new reproducing kernel-based approach for estimation and inference of ODE given noisy observations. We do not assume the functional forms in ODE to be known, or restrict them to be linear or additive, and we allow pairwise interactions. We perform sparse estimation to select individual functionals, and construct confidence intervals for the estimated signal trajectories. We establish the estimation optimality and selection consistency of kernel ODE under both the low-dimensional and high-dimensional settings, where the number of unknown functionals can be smaller or larger than the sample size. Our proposal builds upon the smoothing spline analysis of variance (SS-ANOVA) framework, but tackles several important problems that are not yet fully addressed, and thus extends the scope of existing SS-ANOVA as well. We demonstrate the efficacy of our method through numerous ODE examples.

ARTICLE HISTORY
Received August 2020
Accepted January 2021

KEYWORDS
Component selection and smoothing operator; High dimensionality; Ordinary differential equations; Smoothing spline analysis of variance; Reproducing kernel Hilbert space

1. Introduction

Ordinary differential equation (ODE) has been widely used to model dynamic systems and biological and physical processes in a variety of scientific applications. Examples include infectious disease (Liang and Wu 2008), genomics (Cao and Zhao 2008; Chou and Voit 2009; Ma et al. 2009; Lu et al. 2011; Henderson and Michailidis 2014; Wu et al. 2014), neuroscience (Izhikevich 2007; Zhang et al. 2015, 2017; Cao, Sandstede, and Luo 2019), among many others. A system of ODEs take the form,

$$\frac{dx(t)}{dt} = \begin{pmatrix} \frac{dx_1(t)}{dt} \\ \vdots \\ \frac{dx_p(t)}{dt} \end{pmatrix} = \begin{pmatrix} F_1(x(t)) \\ \vdots \\ F_p(x(t)) \end{pmatrix} = F(x(t)), \qquad (1)$$

where $x(t) = (x_1(t), \ldots, x_p(t))^\top \in \mathbb{R}^p$ denotes the system of p variables of interest, $F = \{F_1, \ldots, F_p\}$ denotes the set of unknown functionals that characterize the regulatory relations among $x(t)$, and t indexes time in an interval standardized to $\mathcal{T} = [0, 1]$. Typically, the system (1) is observed on discrete time points $\{t_1, \ldots, t_n\}$ with measurement errors,

$$y_i = x(t_i) + \epsilon_i, \quad i = 1, \ldots, n, \qquad (2)$$

where $y_i = (y_{i1}, \ldots, y_{ip})^\top \in \mathbb{R}^p$ denotes the observed data, $\epsilon_i = (\epsilon_{i1}, \ldots, \epsilon_{ip})^\top \in \mathbb{R}^p$ denotes the vector of measurement errors that are usually assumed to follow independent normal distribution with mean 0 and variance σ_j^2, $j = 1, \ldots, p$, and n denotes the number of time points. Besides, an initial condition $x(0) \in \mathbb{R}^p$ is usually given for the system (1).

In a biological or physical system, a central question of interest is to uncover the structure of the system of ODEs in terms of which variables regulate which other variables, given the observed noisy time-course data $\{y_i\}_{i=1}^n$. Specifically, we say that x_k regulates x_j, if F_j is a functional of x_k. In other words, x_k controls the change of x_j through the functional F_j on the derivative dx_j/dt. Therefore, the functionals $F = \{F_1, \ldots, F_p\}$ encode the regulatory relations of interest, and are often assumed to take the form,

$$F_j(x(t)) = \theta_{j0} + \sum_{k=1}^{p} F_{jk}(x_k(t)) + \sum_{k \neq l, k=1}^{p} \sum_{l=1}^{p} F_{jkl}(x_k(t), x_l(t)),$$
$$j = 1, \ldots, p, \qquad (3)$$

where $\theta_{j0} \in \mathbb{R}$ denotes the intercept, and F_{jk} and F_{jkl} represent the main effect and two-way interaction, respectively. Higher order interactions are possible, but two-way interactions are the most common structure studied in ODE (Ma et al. 2009; Zhang et al. 2015).

There have been numerous pioneering works studying statistical modeling of ODEs. However, nearly all existing solutions constrain the forms of F. Broadly speaking, there are three categories of functional forms imposed. The first category considers linear functionals for F. For instance, Lu et al. (2011) studied a system of linear ODEs to model dynamic gene regulatory networks. Zhang et al. (2015) extended the linear ODE to include the interactions to model brain connectivity networks. The model of Zhang et al. (2015), other than differentiating between the variables that encode the neuronal activities and

CONTACT Lexin Li lexinli@berkeley.edu Department of Biostatistics and Epidemiology, the University of California, Berkeley, Berkeley, CA 94720.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

the ones that represent the stimulus signals, is in effect of the form,

$$F_j(x(t)) = \theta_{j0} + \sum_{k=1}^{p} \theta_{jk} x_k(t) + \sum_{k\neq l,k=1}^{p} \sum_{l=1}^{p} \theta_{jkl} x_k(t) x_l(t), \quad j = 1, \ldots, p, \tag{4}$$

whereas the model of Lu et al. (2011) is similar to (4) but focuses on the main-effect terms only. In both cases, F_j takes a linear form. Dattner and Klaassen (2015) further extended the functional F_j in (4) to a generalized linear form, but without the interactions, that is,

$$F_j(x(t)) = \theta_{j0} + \psi_j(x(t))^\top \theta_j, \quad j = 1, \ldots, p, \tag{5}$$

where $\theta_{j0} \in \mathbb{R}$, $\theta_j \in \mathbb{R}^d$, and $\psi_j(x) = (\psi_{j1}(x), \ldots, \psi_{jd}(x))^\top \in \mathbb{R}^d$ is a finite set of *known* basis functions. The second category considers additive functionals for F. Particularly, Henderson and Michailidis (2014), Wu et al. (2014), and Chen, Shojaie, and Witten (2017) considered the generalized additive model for F_j,

$$\begin{aligned} F_j(x(t)) &= \theta_{j0} + \sum_{k=1}^{p} F_{jk}(x_k(t)) \\ &= \theta_{j0} + \sum_{k=1}^{p} \left\{ \psi(x_k(t))^\top \theta_{jk} + \delta_{jk}(x_k(t)) \right\}, \\ & \qquad j = 1, \ldots, p, \end{aligned} \tag{6}$$

where $\theta_{j0} \in \mathbb{R}$, $\theta_{jk} \in \mathbb{R}^d$, $\psi(x) = (\psi_1(x), \ldots, \psi_d(x))^\top \in \mathbb{R}^d$ is a finite set of common basis functions, and $\delta_{jk} \in \mathbb{R}$ is the residual function. Different from Dattner and Klaassen (2015), the residual δ_{jk} is unknown. The functional F_j in (6) takes an additive form. Finally, there is a category of ODE solutions focusing on the scenario where the functional forms for F are *known* (González, Vujačić, and Wit 2014; Zhang, Cao, and Carroll 2015; Mikkelsen and Hansen 2017).

These works have laid a solid foundation for statistical modeling of ODE. However, in plenty of scientific applications, the forms of the functionals F are unknown, and the linear or additive forms on F can be restrictive. Besides, it is highly nontrivial to couple the basis function-based solutions with the interactions. We give an example in Section 2.1, where a commonly used enzyme network ODE system involves both nonlinear functionals and two-way interactions. Such examples are often the rules rather than the exceptions, motivating us to consider a more flexible form of ODE. Moreover, the existing ODE methods have primarily focused on sparse estimation, but few tackled the problem of statistical inference, which is challenging due to the complicated correlation structure of ODE.

In this article, we propose a novel approach of kernel ordinary differential equation (KODE) for estimation and inference of the ODE system in (1) given noisy observations from (2). We adopt the general formulation of (3), but we do not assume the functional forms of F are known, or restrict them to be linear or additive, and we allow pairwise interactions. As such, we consider a more general ODE system that encompasses (4)–(6) as special cases. We further introduce sparsity regularization to achieve selection of individual functionals in (3), which yields a sparse recovery of the regulatory relations among F, and improves the model interpretability. Moreover, we derive the confidence interval for the estimated signal trajectory $x_j(t)$. We establish the estimation optimality and selection consistency of kernel ODE, under both low-dimensional and high-dimensional settings, where the number of unknown functionals p can be smaller or larger than the number of time points n, and we study the regime-switching phenomenon. These differences clearly separate our proposal from the existing ODE solutions in the literature.

Our proposal is built upon the smoothing spline analysis of variance (SS-ANOVA) framework that was first introduced by Wahba et al. (1995), then further developed in regression and functional data analysis settings by Huang (1998), Lin and Zhang (2006), and Zhu, Yao, and Zhang (2014). We adopt a similar component selection and smoothing operator (COSSO) type penalty of Lin and Zhang (2006) for regularization, and conceptually, our work extends COSSO to the ODE setting. However, our proposal considerably differs from COSSO and the existing SS-ANOVA methods, in multiple ways. First, unlike the standard SS-ANOVA models, the regressors of kernel ODE are not directly observed and need to be estimated from the data with error. This extra layer of randomness and estimation error introduces additional difficulty to SS-ANOVA. Second, we employ the integral of the estimated trajectories in the loss function to improve the estimation properties (Dattner and Klaassen 2015). The use of the integral and the inclusion of the interaction terms pose some identifiability question that we tackle explicitly. Third, we establish the estimation optimality and selection consistency in the RKHS framework, which is utterly different from Zhu, Yao, and Zhang (2014), and requires new technical tools. Moreover, our theoretical analysis extends that of Chen, Shojaie, and Witten (2017) from the finite bases setting of cubic splines to the infinite bases setting of RKHS. Finally, for statistical inference, we derive the confidence bands to provide uncertainty quantification for the penalized estimators of the signal trajectories in the ODE model. Our solution builds on the confidence intervals idea of Wahba (1983). But unlike the classical methods focusing on the fixed dimensionality p (Wahba 1983; Opsomer and Ruppert 1997), we allow a diverging p that can far exceed the sample size n. In summary, our proposal tackles several crucial problems that are not yet fully addressed in the existing SS-ANOVA framework, and it is far from a straightforward extension. We believe the proposed kernel ODE method not only makes a useful addition to the toolbox of ODE modeling, but also extends the scope of SS-ANOVA-based kernel learning.

The rest of the article is organized as follows. We propose kernel ODE in Section 2, and develop the estimation algorithm and inference procedure in Section 3. We derive the consistency and optimality of the proposed method in Section 4. We investigate the numerical performance in Section 5, and illustrate with a real data example in Section 6. We conclude the paper with a discussion in Section 7, and relegate all proofs and some additional numerical results to the Supplementary Appendix.

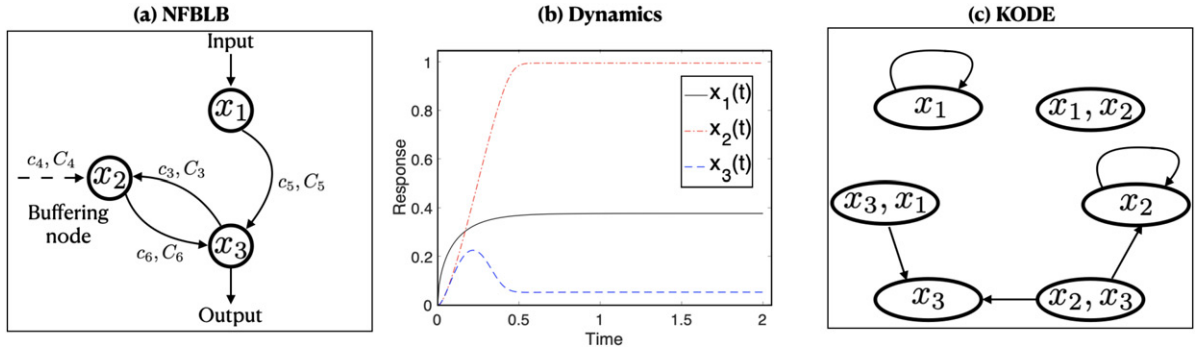

Figure 1. (a) Diagram of the NFBLB regulatory network following (7). (b) Phase dynamics for the three nodes x_1, x_2, x_3 over time [0, 1], with a random input x_0 uniformly drawn from [0.5, 1.5]. (c) Illustration of the NFBLB network in terms of the interactions in KODE.

2. Kernel Ordinary Differential Equations

2.1. Motivating example

We consider an enzymatic regulatory network as an example to demonstrate that nonlinear functionals as well as interactions are common in the system of ODEs. Ma et al. (2009) found that all circuits of three-node enzyme network topologies that perform biochemical adaptation can be well approximated by two architectural classes: a negative feedback loop with a buffering node, and an incoherent feedforward loop with a proportioner node. The mechanism of the first class follows the Michaelis–Menten kinetic equations (Tzafriri 2003),

$$\begin{aligned}
\frac{dx_1(t)}{dt} &= c_1 \frac{x_0\{1-x_1(t)\}}{\{1-x_1(t)\}+C_1} - \tilde{c}_1 c_2 \frac{x_1(t)}{x_1(t)+C_2},\\
\frac{dx_2(t)}{dt} &= c_3 \frac{\{1-x_2(t)\}x_3(t)}{\{1-x_2(t)\}+C_3} - \tilde{c}_2 c_4 \frac{x_2(t)}{x_2(t)+C_4},\\
\frac{dx_3(t)}{dt} &= c_5 \frac{x_1(t)\{1-x_3(t)\}}{\{1-x_3(t)\}+C_5} - c_6 \frac{x_2(t)x_3(t)}{x_3(t)+C_6},
\end{aligned} \tag{7}$$

where $x_1(t), x_2(t), x_3(t)$ are three interacting nodes, such that $x_1(t)$ receives the input, $x_2(t)$ plays the diverse regulatory role, and $x_3(t)$ transmits the output, x_0 is the initial input stimulus, and $c_1, \ldots, c_6, C_1, \ldots, C_6, \tilde{c}_1, \tilde{c}_2$ denote the catalytic rate parameters, the Michaelis–Menten constants, and the concentration parameters, respectively. See Figure 1(a) for a graphical illustration of this ODE system. In this model, the functionals F_1, F_2, F_3 are all nonlinear, and both F_2 and F_3 involve two-way interactions. It is of great interest to estimate F_j's given the observed data, to verify model (7), and to carry out statistical inference of the unknown parameters. This example, along with many other ODE systems with nonlinear functionals and interaction terms motivate us to consider a general ODE system as in (3).

2.2. Two-Step Collocation Estimation

Before presenting our method, we first briefly review the two-step collocation estimation method, which is commonly used for parameter estimation in ODE, and is also useful in our setting. The method was first proposed by Varah (1982), then extended to various ODE models. In the first step, it fits a smoothing estimate,

$$\widehat{x}_j(t) = \arg\min_{z_j \in \mathcal{F}} \left\{ \frac{1}{n}\sum_{i=1}^{n} \{y_{ij} - z_j(t_i)\}^2 + \lambda_{nj} J_1(z_j) \right\}, \quad j = 1, \ldots, p,$$

where $J_1(\cdot)$ is a smoothness penalty in the function space $\mathcal{F}$, and z_j is a function in $\mathcal{F}$ that we minimize over. In the second step, it solves an optimization problem to estimate the model parameters $\theta_{j0} \in \mathbb{R}$ and $\theta_j = (\theta_{j1}, \ldots, \theta_{jp})^\top \in \mathbb{R}^p$, for $j = 1, \ldots, p$. Particularly, Varah (1982) considered the derivative $d\widehat{x}_j(t)/dt$ and the following minimization,

$$\min_{\theta_{j0}, \theta_j} \int_0^1 \left(\frac{d\widehat{x}_j(t)}{dt} - \theta_{j0} - \sum_{k=1}^{p} \theta_{jk} \widehat{x}_k(t) \right)^2 dt, \quad j = 1, \ldots, p.$$

Wu et al. (2014) developed a similar two-step collocation method for their additive ODE model (6), and estimated the model parameters $\theta_{j0} \in \mathbb{R}$ and $\theta_{jk} = (\theta_{jk1}, \ldots, \theta_{jkd})^\top \in \mathbb{R}^d$, for $j, k = 1, \ldots, p$, with a standardized group ℓ_1-penalty,

$$\min_{\theta_{j0}, \theta_{jk}} \int_0^1 \left\| \frac{d\widehat{x}_j(t)}{dt} - \theta_{j0} - \sum_{k=1}^{p} \theta_{jk}^\top \psi(\widehat{x}_k(t)) \right\|_2^2 dt + \tau_{nj} \sum_{k=1}^{p} \left[\int_0^1 \left\{ \theta_{jk}^\top \psi(\widehat{x}_k(t)) \right\}^2 dt \right]^{1/2}.$$

They further discussed adaptive group ℓ_1 and regular ℓ_1-penalties. Meanwhile, Henderson and Michailidis (2014) considered an extra ℓ_2-penalty.

Alternatively, in the second step, Dattner and Klaassen (2015) proposed to focus on the integral $\int_0^t g_j(\widehat{x}(u))du$, rather than the derivative $d\widehat{x}_j(t)/dt$, and they estimated the model parameters $\theta_{j0} \in \mathbb{R}$ and $\theta_j = (\theta_{j1}, \ldots, \theta_{jd})^\top \in \mathbb{R}^d$, for $j = 1, \ldots, p$, in (5) by,

$$\min_{\theta_{j0}, \theta_j} \sum_{j=1}^{p} \int_0^1 \left\{ \widehat{x}_j(t) - \theta_{j0} - \theta_j^\top \int_0^t \psi_j(\widehat{x}(u))du \right\}^2 dt.$$

They found that this modification from the derivative to integral leads to a more robust estimate and also an easier derivation of the asymptotic properties. Chen, Shojaie, and Witten (2017) adopted this idea for their additive ODE model (6),

and estimated the parameters $\theta_{j0} \in \mathbb{R}$, $\tilde{\theta}_j \in \mathbb{R}$, and $\theta_{jk} = (\theta_{jk1}, \ldots, \theta_{jkd})^\top \in \mathbb{R}^d$, for $j, k = 1, \ldots, p$, by

$$\min_{\theta_{j0}, \tilde{\theta}_j, \theta_{jk}} \frac{1}{2n} \sum_{i=1}^{n} \left\{ y_{ij} - \theta_{j0} - b_j t_i - \sum_{k=1}^{p} \theta_{jk}^\top \int_0^{t_i} \psi(\widehat{x}_k(u)) du \right\}^2 + \tau_{nj} \sum_{k=1}^{p} \left[\frac{1}{n} \sum_{i=1}^{n} \left\{ \theta_{jk}^\top \int_0^{t_i} \psi(\widehat{x}_k(u)) dt \right\}^2 \right]^{1/2}.$$

2.3. Kernel ODE

We build the proposed kernel ODE within the smoothing spline ANOVA framework; see Wahba et al. (1995) and Gu (2013) for more background on SS-ANOVA. Specifically, let $\mathcal{H}_k$ denote a space of functions of $x_k(t) \in \mathcal{X}$ with zero marginal integral, where $\mathcal{X} \subset \mathbb{R}$ is a compact set. Let $\{1\}$ denote the space of constant functions. We construct the tensor product space as

$$\mathcal{H} = \{1\} \oplus \sum_{k=1}^{p} \mathcal{H}_k \oplus \sum_{k=1, k\neq l}^{p} \sum_{l=1}^{p} (\mathcal{H}_k \otimes \mathcal{H}_l). \tag{8}$$

We assume the functionals F_j, $j = 1, \ldots, p$, in the ODE model (3) are located in the space of $\mathcal{H}$. The identifiability of the terms in (3) is assured by the conditions specified through the averaging operators: $\int_{\mathcal{T}} F_{jk}(x_k(t)) dt = 0$ for $k = 1, \ldots, p$. Let $\|\cdot\|_{\mathcal{H}}$ denote the norm of $\mathcal{H}$, and $\mathcal{P}^k F_j$ and $\mathcal{P}^{kl} F_j$ denote the orthogonal projection of F_j onto $\mathcal{H}_k$ and $\mathcal{H}_k \otimes \mathcal{H}_l$, respectively. We consider a two-step collocation estimation method, by first obtaining a smoothing spline estimate $\widehat{x}(t) = (\widehat{x}_1(t), \ldots, \widehat{x}_p(t))^\top$, where

$$\widehat{x}_j(t) = \arg\min_{z_j \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \{ y_{ij} - z_j(t_i) \}^2 + \lambda_{nj} \|z_j(t)\|_{\mathcal{F}}^2 \right\}, \quad j = 1, \ldots, p, \tag{9}$$

then estimating $F_j \in \mathcal{H}$ and $\theta_{j0} \in \mathbb{R}$ by the following penalized optimization,

$$\min_{\theta_{j0}, F_j} \frac{1}{n} \sum_{i=1}^{n} \left\{ y_{ij} - \theta_{j0} - \int_0^{t_i} F_j(\widehat{x}(t)) dt \right\}^2 + \tau_{nj} \left(\sum_{k=1}^{p} \|\mathcal{P}^k F_j\|_{\mathcal{H}} + \sum_{k\neq l, k=1}^{p} \sum_{l=1}^{p} \|\mathcal{P}^{kl} F_j\|_{\mathcal{H}} \right). \tag{10}$$

Our proposal deals with the integral $\int_0^{t_i} F_j(\widehat{x}(u)) du$, rather than the derivative $d\widehat{x}_j(t)/dt$, which is in a similar spirit as Dattner and Klaassen (2015). Besides, it involves two penalty functions, $J_1 \equiv \|\cdot\|_{\mathcal{F}}^2$ in (9), and $J_2(F_j) \equiv \sum_{k=1}^{p} \|\mathcal{P}^k F_j\|_{\mathcal{H}} + \sum_{k=1}^{p} \sum_{l=1}^{p} \|\mathcal{P}^{kl} F_j\|_{\mathcal{H}}$ in (10), with λ_{nj} and τ_{nj} as two tuning parameters. We next make some remarks about this proposal.

For the functionals, the formulation in (10) is highly flexible, nonlinear, and incorporates two-way interactions. Meanwhile, it naturally covers the linear ODE in (4) and (5), and the additive ODE in (6) as special cases. In particular, if $\mathcal{H}$ is the linear functional space, $\mathcal{H} = \{1\} \oplus \sum_{k=1}^{p} \{x_k - 1/2\} \oplus \sum_{k\neq l} [\{x_k - 1/2\} \otimes \{x_l - 1/2\}]$ with the input space $\mathcal{X} = [0, 1]^p$, then any F of the form in (4) belongs to $\mathcal{H}$. If $\mathcal{H}$ is spanned by some known generalized functions, $\mathcal{H} = \psi_{j1}(x) \oplus \cdots \oplus \psi_{jp}(x)$, then any F in (5) belongs to $\mathcal{H}$. If $\mathcal{H}$ is the additive functional space, $\mathcal{H} = \{1\} \oplus \sum_{k=1}^{p} \mathcal{H}_k$ with the ℓ_2-norm, then for $F_{jk}(x_k(t)) = \psi(x_k(t))^\top \theta_{jk}$, the penalty on the main effects becomes $\sum_{k=1}^{p} \|\mathcal{P}^k F_j\|_{\mathcal{H}} = \sum_{k=1}^{p} [\int_0^1 \{\psi(x_k(t))^\top \theta_{jk}\}^2 dt]^{1/2}$, which is exactly the same as the ODE model of Chen, Shojaie, and Witten (2017).

For the penalties, the first penalty function J_1 is the squared RKHS norm corresponding to the RKHS $\{\mathcal{F}, \|\cdot\|_{\mathcal{F}}\}$. It is for estimating $\widehat{x}_j$, and $\mathcal{F}$ does not have to be the same as $\mathcal{H}$. The second penalty function J_2 is a sum of RKHS norms on the main effects and pairwise interactions. This penalty is similar as the COSSO penalty of Lin and Zhang (2006). But as we outline in Section 1, our extension is far from trivial. We also note that, we do not impose a hierarchical structure for the main effects and interactions, in that if an interaction term is selected, the corresponding main effect term does not have to be selected (Wang et al. 2009). This is motivated by the observation that, for example, in the enzymatic regulatory network example in Section 2.1, the interaction terms $x_1(t)x_3(t)$ and $x_2(t)x_3(t)$ both appear in the ODE regulating $x_3(t)$, but the main effect terms $x_1(t)$ and $x_2(t)$ are not present.

Theorem 1. Assume that the RKHS $\mathcal{H}$ can be decomposed as in (8). Then there exists a minimizer of (10) in $\mathcal{H}$ for any tuning parameter $\tau_{nj} \geq 0$. Moreover, the minimizer is in a finite-dimensional space.

Theorem 1 is a generalization of the well-known representer theorem (Wahba 1990). The difference is that, unlike the smoothing splines model as studied in Wahba (1990), the minimization of (10) involves an integral in the loss function, and the penalty is not a norm in $\mathcal{H}$ but a convex pseudo-norm. A direct implication of Theorem 1 is that, although the minimization with respect to F_j is taken over an infinite-dimensional space in (10), the solution to (10) can actually be found in a finite-dimensional space. We next develop an estimation algorithm to solve (10).

3. Estimation and Inference

3.1. Estimation Procedure

The estimation of the proposed kernel ODE system consists of two major steps. The first step is the smoothing spline estimation in (9), which is standard and the tuning of the smoothness parameter λ_{nj} is often done through generalized cross-validation (see, e.g., Gu 2013). The second step is to solve (10). Toward that end, we first propose an optimization problem that is equivalent to (10), but is computationally easier to tackle. We then develop an estimation algorithm to solve this new equivalent problem.

Specifically, we consider the following optimization problem, for $j = 1, \ldots, p$,

$$\min_{\theta_{j0}, \theta_j, F_j} \frac{1}{n} \sum_{i=1}^{n} \left\{ y_{ij} - \theta_{j0} - \int_0^{t_i} F_j(\widehat{x}(t)) dt \right\}^2 + \eta_{nj} \left(\sum_{k=1}^{p} \theta_{jk}^{-1} \|\mathcal{P}^k F_j\|_{\mathcal{H}}^2 + \theta_{jkl}^{-1} \sum_{k=1, k\neq l}^{p} \sum_{l=1}^{p} \|\mathcal{P}^{kl} F_j\|_{\mathcal{H}}^2 \right)$$

$$+\kappa_{nj}\left(\sum_{k=1}^{p}\theta_{jk}+\sum_{k=1,k\neq l}^{p}\sum_{l=1}^{p}\theta_{jkl}\right), \tag{11}$$

subject to $\theta_k \geq 0, \theta_{kl} \geq 0, k, l = 1, \ldots, p, k \neq l$, where $\theta_j = (\theta_{j1}, \ldots, \theta_{jp}, \theta_{j12}, \ldots, \theta_{j1p}, \ldots, \theta_{jp1}, \ldots, \theta_{jp(p-1)})^\top \in \mathbb{R}^{p^2}$ collects the parameters to estimate, and $\eta_{nj}, \kappa_{nj\geq 0}$ are the tuning parameters, $j = 1, \ldots, p$. Comparing (11) to (10), we introduce the parameters θ_{jk} and θ_{jkl} to control the sparsity of the main effect and interaction terms in F_j. This is similar to Lin and Zhang (2006). The two optimization problems (10) and (11) are equivalent, in the following sense. Let $\kappa_{nj} = \tau_{nj}^2/(4\eta_{nj})$. Then we have,

$$\eta_{nj}\theta_{jk}^{-1}\|\mathcal{P}^k F_j\|_{\mathcal{H}}^2 + \kappa_{nj}\theta_{jk} \geq 2\eta_{nj}^{1/2}\kappa_{nj}^{1/2}\|\mathcal{P}^k F_j\|_{\mathcal{H}} = \tau_{nj}\|\mathcal{P}^k F_j\|_{\mathcal{H}},$$

where the equality holds if $\theta_{jk} = \eta_{nj}^{1/2}\kappa_{nj}^{-1/2}\|P^k F_j\|_{\mathcal{H}}$. A similar result holds for $\theta_{jkl} = \eta_{nj}^{1/2}\kappa_{nj}^{-1/2}\|P^{kl} F_j\|_{\mathcal{H}}$. In other words, if $(\widehat{\theta}_{j0}, \widehat{F}_j)$ minimizes (10), then $(\widehat{\theta}_{j0}, \widehat{\theta}_j, \widehat{F}_j)$ minimizes (11), with $\widehat{\theta}_{jk} = \eta_{nj}^{1/2}\kappa_{nj}^{-1/2}\|\mathcal{P}^k \widehat{F}_j\|_{\mathcal{H}}$, and $\theta_{jkl} = \eta_{nj}^{1/2}\kappa_{nj}^{-1/2}\|\mathcal{P}^{kl} F_j\|_{\mathcal{H}}$, for any $k, l = 1, \ldots, p, k \neq l$. Meanwhile, if $(\widehat{\theta}_{j0}, \widehat{\theta}_j, \widehat{F}_j)$ minimizes (11), then $(\widehat{\theta}_{j0}, \widehat{F}_j)$ minimizes (10).

Next, we devise an iterative alternating optimization approach to solve (11). That is, we first estimate θ_{j0} given fixed F_j and θ_j, then estimate the functional F_j given fixed θ_{j0} and θ_j, and finally estimate θ_j given fixed θ_{j0} and F_j.

For given $\widehat{F}_j$ and $\widehat{\theta}_j$, we have that,

$$\widehat{\theta}_{j0} = \bar{y}_j - \int_{\mathcal{T}} \bar{T}(t)\widehat{F}_j(\widehat{x}(t))dt,$$

where $T_i(t) = 1\{0 \leq t \leq t_i\}$, $\bar{T}(t) = \frac{1}{n}\sum_{i=1}^n T_i(t)$, and $\bar{y}_j = n^{-1}\sum_{i=1}^n y_{ij}$.

For given $\widehat{\theta}_{j0}$ and $\widehat{\theta}_j$, the optimization problem (11) becomes,

$$\min_{F_j}\left\{\frac{1}{n}\sum_{i=1}^{n}\left[(y_{ij}-\bar{y}_j) - \int_{\mathcal{T}}\left\{T_i(t)-\bar{T}(t)\right\}F_j(\widehat{x}(t))dt\right]^2 + \eta_{nj}\left(\sum_{k=1}^{p}\widehat{\theta}_{jk}^{-1}\|\mathcal{P}^k F_j\|_{\mathcal{H}}^2 + \widehat{\theta}_{jkl}^{-1}\sum_{k=1,k\neq l}^{p}\sum_{l=1}^{p}\|\mathcal{P}^{kl} F_j\|_{\mathcal{H}}^2\right)\right\}. \tag{12}$$

Let $K_j(\cdot,\cdot) : \mathcal{X}\times\mathcal{X} \mapsto \mathbb{R}$ denote the Mercer kernel generating the RKHS $\mathcal{H}_j, j = 1, \ldots, p$. Then $K_{kl} \equiv K_k K_l$ is the reproducing kernel of the RKHS $\mathcal{H}_k \otimes \mathcal{H}_l$. Let $K_{\theta_j} = \sum_{k=1}^p \widehat{\theta}_{jk}K_k + \sum_{k\neq l}\widehat{\theta}_{jkl}K_{kl}$. By the representer theorem (Wahba 1990), the solution $\widehat{F}_j$ to (12) is of the form,

$$\widehat{F}_j(\widehat{x}(t)) = b_j + \sum_{i=1}^{n} c_{ij}\int_{\mathcal{T}} K_{\theta_j}(\widehat{x}(t),\widehat{x}(s))\left\{T_i(s)-\bar{T}(s)\right\}ds \tag{13}$$

for some $b_j \in \mathbb{R}$ and $c_j = (c_{1j}, \ldots, c_{nj}) \in \mathbb{R}^n$. Write $y_j = (y_{1j}, \ldots, y_{nj})^\top \in \mathbb{R}^n$ and $\bar{y}_j = (\bar{y}_j, \ldots, \bar{y}_j)^\top \in \mathbb{R}^n$. Let B be an $n \times 1$ vector whose ith entry is $B_i = \int_{\mathcal{T}}\{T_i(t) - \bar{T}(t)\}dt$, $i = 1, \ldots, n$. Let Σ be an $n \times n$ matrix whose (i, i')th entry is $\Sigma_{ii'} = \int_{\mathcal{T}}\int_{\mathcal{T}}\{T_i(s) - \bar{T}(s)\}K_{\theta_j}(\widehat{x}(t),\widehat{x}(s))\{T_{i'}(t) - \bar{T}(t)\}dsdt$, $i, i' = 1, \ldots, n$. Plugging (13) into (12), we obtain the following quadratic minimization problem in terms of $\{b_j, c_j\}$,

$$\min_{b_j,c_j}\frac{1}{n}\|(y_j-\bar{y}_j) - (Bb_j + \Sigma c_j)\|_2^2 + \eta_{nj}c_j^\top \Sigma c_j,$$

which has a closed-form solution. Consider the QR decomposition $B = [Q_1\ Q_2][R\ 0]^\top$, where $Q_1 \in \mathbb{R}^{n\times 1}$, $Q_2 \in \mathbb{R}^{n\times(n-1)}$, and $[Q_1\ Q_2]$ is orthogonal such that $B^\top Q_2 = 0_{1\times(n-1)}$. Write $W_j = \Sigma + n\eta_{nj}I_n$, where I_n is the $n \times n$ identity matrix. Then the minimizers are,

$$c_j = Q_2(Q_2^\top W_j Q_2)^{-1}Q_2^\top(y_j - \bar{y}_j),$$
$$b_j = R^{-1}Q_1^\top(y_j - \bar{y}_j - W_j c_j).$$

Following the usual smoothing splines literature, we tune the parameter η_{nj} in (12) by minimizing the generalized cross-validation criterion (GCV, Wahba et al. 1995),

$$\text{GCV} = \frac{\|A_j(\eta_{nj})(y_j-\bar{y}_j) - (y_j-\bar{y}_j)\|^2}{[n^{-1}\text{tr}\{I_n - A_j(\eta_{nj})\}]^2},$$

where the smoothing matrix $A_j(\eta_{nj}) \in \mathbb{R}^{n\times n}$ is of the form,

$$A_j(\eta_{nj}) = I_n - n\eta_{nj}Q_2(Q_2^\top W_j Q_2)^{-1}Q_2^\top. \tag{14}$$

For given $\widehat{\theta}_{j0}$ and $\widehat{F}_j$, θ_j is the solution to a usual ℓ_1-penalized regression problem,

$$\min_{\theta_j}\left\{(z_j - G\theta_j)^\top(z_j - G\theta_j) + n\kappa_{nj}\left(\sum_{k=1}^{p}\theta_{jk} + \sum_{k\neq l,k=1}^{p}\sum_{l=1}^{p}\theta_{jkl}\right)\right\}, \tag{15}$$

subject to $\theta_k \geq 0, \theta_{kl} \geq 0, k, l = 1, \ldots, p, k \neq l$, where the "response" is $z_j = (y_j - \bar{y}_j) - (1/2)n\eta_{nj}c_j - Bb_j$, the "predictor" is $G \in \mathbb{R}^{n\times p^2}$, whose first p columns are $\Sigma^k c_j$ with $k = 1, \ldots, p$, and the last $p(p-1)$ columns are $\Sigma^{kl}c_j$ with $k, l = 1, \ldots, p, k \neq l$, and $\Sigma^k = (\Sigma_{ii'}^k)$, $\Sigma^{kl} = (\Sigma_{ii'}^{kl})$ are both $n \times n$ matrices whose (i, i')th entries are $\Sigma_{ii'}^k = \int_{\mathcal{T}}\int_{\mathcal{T}}\{T_i(s) - \bar{T}(s)\}K_k(\widehat{x}(t),\widehat{x}(s))\{T_{i'}(t) - \bar{T}(t)\}dsdt$, and $\Sigma_{ii'}^{kl} = \int_{\mathcal{T}}\int_{\mathcal{T}}\{T_i(s) - \bar{T}(s)\}K_{kl}(\widehat{x}(t),\widehat{x}(s))\{T_{i'}(t) - \bar{T}(t)\}dsdt$, respectively, where $i, i' = 1, \ldots, n, j = 1, \ldots, p$. We employ Lasso for (15) in our implementation, and tune the parameter κ_{nj} using 10-fold cross-validation, following the usual Lasso literature.

We repeat the above optimization steps iteratively until some stopping criterion is met; that is, when the estimates in two consecutive iterations are close enough, or when the number of iterations reaches some maximum number. In our simulations, we have found that the algorithm converges quickly, usually within 10 iterations. Another issue is the identifiability of $\mathcal{P}^k F_j$'s and $\mathcal{P}^{kl}F_j$'s in (11) in the sense of unique solutions. We introduce the collinearity indices $\mathcal{C}_{jk}$ and $\mathcal{C}_{jkl}$ to reflect the identifiability. Specifically, let $\mathcal{W}$ denote a $p^2 \times p^2$ matrix, whose entries are $\cos(\mathcal{P}^k F_j, \mathcal{P}^{k'}F_j)$, $\cos(\mathcal{P}^k F_j, \mathcal{P}^{k'l'}F_j)$, $\cos(\mathcal{P}^{kl}F_j, \mathcal{P}^{k'}F_j)$, $\cos(\mathcal{P}^{kl}F_j, \mathcal{P}^{k'l'}F_j)$, $j, k, l = 1, \ldots, p$. Then $\mathcal{C}_{jk}^2$ and $\mathcal{C}_{jkl}^2$ are defined by the diagonals of $\mathcal{W}^{-1}$. When some $\mathcal{C}_{jk}$ and $\mathcal{C}_{jkl}$ are much larger than one, then the identifiability issue occurs (Gu 2013). This is often due to insufficient amount of data relative to the complexity of the model we fit. In this case, we find that increasing η_{nj} and κ_{nj} in (11) often helps with the identifiability issue, as it helps reduce the model complexity.

We summarize the above estimation procedure in Algorithm 1.

Algorithm 1 Iterative optimization algorithm for kernel ODE.

1: Initialization: the initial values for $\theta_{jk} = \theta_{jkl} = 1, j, k, l = 1, \ldots, p, k \neq l$, and the tuning parameters: (η_{nj}, κ_{nj}).
2: Fit smoothing spline model (9), and obtain $\widehat{x}_j(t)$, $j = 1, \ldots, p$.
3: **repeat**
4: Solve $\widehat{\theta}_{j0}$ given $\widehat{F}_j$ and $\widehat{\theta}_j$, $j = 1, \ldots, p$.
5: Solve $\widehat{F}_j$ in (12) given $\widehat{\theta}_{j0}$ and $\widehat{\theta}_j$, $j = 1, \ldots, p$.
6: Solve $\widehat{\theta}_j$ in (15) given $\widehat{\theta}_{j0}$ and $\widehat{F}_j$, $j = 1, \ldots, p$.
7: **until** the stopping criterion is met.

3.2. Confidence Intervals

Next, we derive the confidence intervals for the estimated trajectory $\widehat{x}_j(t_i)$. This is related to post-selection inference, as the actual coverage probability of the confidence interval ignoring the preceding sparse estimation uncertainty can be dramatically smaller than the nominal level. Our result extends the recent work of Berk et al. (2013) and Bachoc, Leeb, and Pötscher (2019) from linear regression models to nonparametric ODE models, while our setting is more challenging, as it involves infinite-dimensional functional objects.

Let $\widehat{\theta}_j$ denote the estimator of θ_j obtained from Algorithm 1. Denote $\mathcal{M} \equiv \{1, \ldots, p, (1,2), \ldots, (1,p), \ldots, (p,1), \ldots, (p,p-1)\}$, and denote $M_j \subseteq \mathcal{M}$ as the index set of the nonzero entries of the sparse estimator $\widehat{\theta}_j$. Note that M_j is allowed to be an empty set. Let $\widehat{\theta}_{M_j}$ be the least squares estimate with M_j as the support that minimizes the unpenalized objective function in (15), that is, $(z_j - G\theta_j)^\top (z_j - G\theta_j)$. Plugging this estimate $\widehat{\theta}_{M_j}$ into (13) gets the corresponding estimate of the functional F_j as,

$$\widehat{F}_{j,\widehat{\theta}_{M_j}}(\widehat{x}(t)) = b_j + \sum_{i=1}^{n} c_{ij} \int_{\mathcal{T}} K_{\widehat{\theta}_{M_j}}(\widehat{x}, \widehat{x}(s)) \left\{ T_i(s) - \bar{T}(s) \right\} ds.$$

For a nominal level $\alpha \in (0,1)$ and $i = 1, \ldots, n$, define $c_0(\widehat{x}_j(t_i))$ as the smallest constant satisfying that,

$$\mathbb{P}_{n,F_j,\sigma_j} \left[\max_{M_j \subseteq \mathcal{M}} \sigma_j^{-1} \left| \{\widetilde{A}_{M_j}\}_{i\cdot} (y_j - \bar{y}_j) \right| \leq c_0(\widehat{x}_j(t_i)) \right] \geq 1 - \alpha, \tag{16}$$

where $\{\widetilde{A}_{M_j}\}_{i\cdot} = \{A_{M_j}\}_{i\cdot} / \|\{A_{M_j}\}_{i\cdot}\|_{l_2}$, $\{A_{M_j}\}_{i\cdot}$ is the ith row of A_{M_j}, A_{M_j} is the smoothing matrix as defined in (14) with the corresponding $\widehat{\theta}_{M_j}$, and σ_j^2 is the variance of the error term ϵ_{ij} in (2). We then construct the confidence interval $\mathrm{CI}(\widehat{x}_j(t))$ for the prediction of true trajectory $x_j(t)$ following model selection as,

$$\mathrm{CI}(\widehat{x}_j(t_i)) = \int_{\mathcal{T}} \left\{ T_i(t) - \bar{T}(t) \right\} \widehat{F}_{j,\widehat{\theta}_{M_j}}(\widehat{x}(t)) dt \pm c_0(\widehat{x}_j(t_i)) \sigma_j \|\{A_{M_j}\}_{i\cdot}\|, \tag{17}$$

for any $i = 1, \ldots, n$ and $j = 1, \ldots, p$.

Next, we show that the confidence interval in (17) has the desired coverage probability. Later we develop a procedure to estimate the cutoff value $c_0(\widehat{x}_j)$ in (16) given the data.

Theorem 2. Let $M_j \subseteq \mathcal{M}$ be the index set of the nonzero entries of the sparse estimator $\widehat{\theta}_j$. Then the choice of $c_0(\widehat{x}_j(t_i))$ in (16) does not depend on F_j, and $\mathrm{CI}(\widehat{x}_j(t_i))$ in (17) satisfies the coverage property, for any $i = 1, \ldots, n$ and $j = 1, \ldots, p$, in that,

$$\inf_{F_j \in \mathcal{H}, \sigma_j > 0} \mathbb{P} \left\{ \int_{\mathcal{T}} \left\{ T_i(t) - \bar{T}(t) \right\} \mathbb{E} \left[\widehat{F}_{j,\widehat{\theta}_{M_j}}(\widehat{x}(t)) \right] dt \in \mathrm{CI}(\widehat{x}_j(t_i)) \right\} \geq 1 - \alpha.$$

A few remarks are in order. First, the coverage in Theorem 2 is guaranteed for all sparse estimation and selection procedures. As such, $\mathrm{CI}(\widehat{x}_j)$ in (17), following the terminology of Berk et al. (2013), is a universally valid post-selection confidence interval. Second, if we replace $c_0(\widehat{x}_j(t_i))$ in (17) by $z_{\alpha/2}$, that is, the $\alpha/2$ cutoff value of a standard normal distribution, then $\mathrm{CI}(\widehat{x}_j(t_i))$ reduces to the "naive" confidence interval. It is constructed as if M_j were fixed a priori, and it ignores any uncertainty or error of the sparse estimation step. This naive confidence interval, however, does not have the coverage property as in Theorem 2, and thus is not a truly valid confidence interval. Finally, data splitting is a commonly used alternative strategy for post-selection inference. But it is not directly applicable in our ODE setting, because it is difficult to split the time series data into independent parts.

Next, we devise a procedure to compute the cutoff value $c_0(\widehat{x}_j(t_i))$.

Proposition 1. The value $c_0(\widehat{x}_j(t_i))$ in (16) is the same as the solution of $t \geq 0$ satisfying,

$$\mathbb{E}_U \mathbb{P} \left(\max_{M_j \subseteq \mathcal{M}} \left| \{\widetilde{A}_{M_j}\}_{i\cdot} V \right| \leq t/U \,\middle|\, U \right) = 1 - \alpha,$$

where V is uniformly distributed on the unit sphere in $\mathbb{R}^n$, and U is a nonnegative random variable such that U^2 follows a chi-squared distribution $\chi^2(n)$.

Following Proposition 1, we compute $c_0(\widehat{x}_j(t_i))$ as follows. We first generate N iid copies of random vectors $V_1, \ldots, V_N$ uniformly distributed on the unit sphere in $\mathbb{R}^n$. We then calculate the quantity, $c_\nu = \max_{M_j \subseteq \mathcal{M}} |\{\widetilde{A}_{M_j}\}_{i\cdot} V_\nu|$ for $\nu = 1, \ldots, N$. Let D_U denote the cumulative distribution function of U, and D_{χ^2} denote the cumulative distribution function of a $\chi^2(n)$ distribution. Then $D_U(t) = D_{\chi^2}(t^2)$. We next obtain $c_0(\widehat{x}_j(t_i))$ by searching for c that solves $N^{-1} \sum_{i=1}^{N} D_U(c/c_i) = 1 - \alpha$, using, for example, a bisection searching method.

Finally, we estimate the error variance σ_j^2 in (17) using the usual noise estimator in the context of RKHS (Wahba 1990); that is, $\widehat{\sigma}_j^2 = \|A_{M_j}(y_j - \bar{y}_j) - (y_j - \bar{y}_j)\|^2 / \mathrm{tr}(I - A_{M_j})$.

We also remark that, the inference on the prediction of the trajectory $x_j(t)$ following model selection as described in Theorem 2 amounts to the inference on the estimation of the integration $\int_0^t F_j(x(s))ds$. This type of inference is of great importance in dynamic systems (Izhikevich 2007; Chou and Voit 2009; Ma et al. 2009). Our solution takes the selected model as an approximation to the truth, but does not require that the true data generation model has to be among the candidates of model selection. We note that, it is also possible to do inference on the individual components of F_j directly; for example, one could construct the confidence interval for F_{jk} in (3). But this is achieved at the cost of imposing additional assumptions, including the requirement that the true data generation model

is among the class of pairwise interaction model as in (3), and the orthogonality property as in Chernozhukov, Hansen, and Spindler (2015), or its equivalent characterization as in Zhang and Zhang (2014); Javanmard and Montanari (2014). For nonparametric kernel estimators, the orthogonality property is shown to hold if the covariates x_j's are assumed to be weakly dependent (Lu, Kolar, and Liu 2020). It is interesting to further investigate if such a property holds in the context of kernel ODE model under a similar condition of weakly dependent covariates. We leave this as our future research.

4. Theoretical Properties

We next establish the estimation optimality and selection consistency of kernel ODE. These theoretical results hold for both the low-dimensional and high-dimensional settings, where the number of functionals p can be smaller or larger than the sample size n. We first introduce two assumptions.

Assumption 1. The number of nonzero functional components is bounded, that is, $\text{card}\big(\{k : F_{jk} \neq 0\} \cup \{1 \leq l \neq k \leq p : F_{jkl} \neq 0\}\big)$ is bounded for any $j = 1, \ldots, p$.

Assumption 2. For any $F_j \in \mathcal{H}$, there exists a random variable B, with $\mathbb{E}(B) < \infty$, and

$$\left|\frac{\partial F_j(x)}{\partial x_k}\right| \leq B\|F_j\|_{L_2}, \text{ almost surely.}$$

Assumption 1 concerns the complexity of the functionals. Similar assumptions have been adopted in the sparse additive model over RKHS when $F_{jkl} = 0$ (see, e.g., Koltchinskii and Yuan 2010; Raskutti, Wainwright, and Yu 2011). Assumption 2 is an inverse Poincaré inequality type condition, which places regularization on the fluctuation in F_j relative to the ℓ_2-norm. The same assumption was also used in additive models in RKHS (Zhu, Yao, and Zhang 2014).

We begin with the error bound for the estimated trajectory $\widehat{x}(t)$ uniformly for $j = 1, \ldots, p$. This is a relatively standard result, which is needed for both analyzing the error of the functional estimators in kernel ODE, and establishing the selection consistency later.

Theorem 3 (Optimal estimation of the trajectory). Suppose that $x_j(t) \in \mathcal{F}, j = 1, \ldots, p$, and the RKHS $\mathcal{F}$ is embedded to a β_1th order Sobolev space, $\beta_1 > 1/2$. Then the smoothing spline estimate from (9) satisfies that, for any $j = 1, \ldots, p$,

$$\min_{\lambda_{nj} \geq 0} \int_{\mathcal{T}} \left\{\widehat{x}_j(t) - x_j(t)\right\}^2 dt = O_p\left(n^{-\frac{2\beta_1}{2\beta_1+1}}\right),$$

which achieves the minimax optimal rate.

Next, we derive the convergence rate for the estimated functional F_j. Because the trajectory $\widehat{x}$ is estimated, to establish the optimal rate of convergence, it requires extra theoretical attention, which is related to recent work on errors in variables for lasso-type regressions (Loh and Wainwright 2012; Zhu, Yao, and Zhang 2014). The proof involves several tools for the Rademacher processes (van der Vaart and Wellner 1996), and the concentration inequalities for empirical processes (Talagrand 1996; Yuan and Zhou 2016).

Theorem 4 (Optimal estimation of the functional). Suppose that $F_j \in \mathcal{H}, j = 1, \ldots, p$, where $\mathcal{H}$ satisfies (8), and the RKHS $\mathcal{H}_j$ is embedded to a β_2th order Sobolev space, $\beta_2 > 1$. Suppose Assumptions 1 and 2 hold. Then, as long as F_j is not a constant function, the KODE estimate $\widehat{F}_j$ from (10) satisfies that, for any $j = 1, \ldots, p$,

$$\min_{\tau_{nj} \geq 0} \int_{\mathcal{T}} \left\{\widehat{F}_j(x(t)) - F_j(x(t))\right\}^2 dt = O_p\left(\left(\frac{n}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}} + \frac{\log p}{n} + n^{-\frac{2\beta_1}{2\beta_1+1}}\right),$$

which achieves the minimax optimal rate.

This theorem is one of our key results, and we make a few remarks. First, there are three error terms in Theorem 4, which are attributed to the estimation of the interactions, the Lasso estimation, and the measurement errors in variables, respectively. Particularly, the error term $O_p\left(n^{-2\beta_1/(2\beta_1+1)}\right)$ arises due to the unobserved $x(t)$, which is instead measured at discrete time points and is subject to measurement errors. Since this error term achieves the optimal rate, it fully characterizes the influence of the estimated $\widehat{x}(t)$ on the resulting estimator $\widehat{F}_j$. Moreover, β_1 and β_2 measure the orders of smoothness for estimating x_j and F_j, respectively. They can be different, which makes it flexible when choosing kernels for the estimation procedure. For instance, if there is prior knowledge that $x(t)$ is smooth, we may then choose $\beta_1 > \beta_2$, and the resulting estimator $\widehat{F}_j$ achieves a convergence rate of $O_p\left((n/\log n)^{-2\beta_2/(2\beta_2+1)} + \log p/n\right)$. It is interesting to note that this rate is the same as the rate as if $x(t)$ were directly observed and there were no integral involved in the loss function, for example, in the setting of Lin and Zhang (2006).

Second, there exists a regime-switching phenomenon, depending on the dimensionality p with respect to the sample size n. On one hand, if it is an ultrahigh-dimensional setting, that is, $p > \exp\left[\left\{n(\log n)^{2\beta_2}\right\}^{\frac{1}{2\beta_2+1}}\right]$, then the minimax optimal rate in Theorem 4 becomes $O_p\left(\log p/n + n^{-2\beta_1/(2\beta_1+1)}\right)$. Here, the first rate $O_p(\log p/n)$ matches with the minimax optimal rate for estimating a p-dimensional linear regression when the vector of regression coefficients has a bounded number of nonzero entries (Raskutti, Wainwright, and Yu 2011). Hence, we pay no extra price in terms of the rate of convergence for adopting a nonparametric modeling of F_j in (3), when compared with the more restrictive linear ODE model in (4) (Zhang et al. 2015). On the other hand, if it is a low-dimensional setting, that is, $p \leq \exp\left[\left\{n(\log n)^{2\beta_2}\right\}^{\frac{1}{2\beta_2+1}}\right]$, then the optimal rate becomes $O_p\left((n/\log n)^{-2\beta_2/(2\beta_2+1)} + n^{-2\beta_1/(2\beta_1+1)}\right)$. Here, the first rate $O_p\left((n/\log n)^{-2\beta_2/(2\beta_2+1)}\right)$ is the same as the optimal rate of estimating F_j as if we knew a priori that F_j comes from a two-dimensional tensor product functional space, rather than the p-variate functional space $\mathcal{H}$ in (8); see also Lin (2000) for a similar observation.

Third, the optimal rate in Theorem 4 is immune to the "curse of dimensionality", in the following sense. We introduce $p(p-1)$ pairwise interaction components to $\mathcal{H}$ in (8), and henceforth, for each $x_j(t)$, $j = 1, \ldots, p$, it requires to estimate a total of p^2 functions. A direct application of an existing basis expansion approach, for instance, Brunton, Proctor, and Kutz (2016), leads to a rate of $O_p\left(n^{-O(1/p^2)}\right)$. This rate degrades fast when p increases. By contrast, we proceed in a different way, where we simultaneously aim for the flexibility of a nonparametric ODE model by letting $\mathcal{H}$ obey a tensor product structure as in (8), while exploiting the interaction structure of the system. As a result, our optimal error bound $O_p\left((n/\log n)^{-2\beta_2/(2\beta_2+1)}\right)$ does not depend on the dimensionality p.

Finally, the incorporation of the integral, $\int_0^{t_i} F_j(\widehat{x}(t))dt$, in the loss function in (10) makes the estimation error of $\widehat{F}_j$ depend on the convergence of $\mathbb{E}\int_{\mathcal{T}}\{\widehat{x}_j(t) - x_j(t)\}^2 dt$. As a comparison, if we use the derivative instead of the integration, then the estimation error would depend on the convergence of the derivative, $\mathbb{E}\int_{\mathcal{T}}\{d\widehat{x}_j(t)/dt - dx_j(t)/dt\}^2 dt$ (Wu et al. 2014). However, it is known that the derivative estimation in the reproducing kernel Hilbert space has a slower convergence rate than the function estimation (Cox 1983). That is, $\mathbb{E}\int_{\mathcal{T}}\{d\widehat{x}_j(t)/dt - dx_j(t)/dt\}^2 dt$ converges at a slower rate than $\mathbb{E}\int_{\mathcal{T}}\{\widehat{x}_j(t) - x_j(t)\}^2 dt$. This demonstrates the advantage of working with the integral in our KODE formulation, and our result echoes the observation for the additive ODE model (Chen, Shojaie, and Witten 2017).

Next, we establish the selection consistency of KODE. Putting all the functionals $\{F_1, \ldots, F_p\}$ together forms a network of regulatory relations among the p variables $\{x_1(t), \ldots, x_p(t)\}$. Recall that, we say x_k is a regulator of x_j, if in (3) F_{jk} is nonzero, or if F_{jkl} is nonzero for some $l \neq k$. Denote the set of the true regulators and the estimated regulators of $x_j(t)$ by

$$S_j^0 = \big\{1 \le k \le p : F_{jk} \neq 0, \text{ or } F_{jkl} \neq 0 \text{ for some } 1 \le l \neq k \le p\big\},$$

$$\widehat{S}_j = \big\{1 \le k \le p : \|\widehat{F}_{jk}\|_{\mathcal{H}} \neq 0, \text{ or } \|\widehat{F}_{jkl}\|_{\mathcal{H}} \neq 0 \text{ for some } 1 \le l \neq k \le p\big\},$$

respectively, $j = 1, \ldots, p$. We need some extra regularity conditions on the minimum regulatory effect and the design matrix, which are commonly adopted in the literature of Lasso regression (Zhao and Yu 2006; Ravikumar, Wainwright, and Lafferty 2010). In the interest of space, we defer those conditions to Section S.1.6.2 of the Appendix. The next theorem establishes that KODE is able to recover the true regulatory network asymptotically.

Theorem 5 (Recovery of the regulatory network). Suppose that $F_j \in \mathcal{H}, j = 1, \ldots, p$, where $\mathcal{H}$ satisfies (8), and the RKHS $\mathcal{H}_j$ is embedded to a β_2th order Sobolev space, with $\beta_2 > 1$. Suppose Assumption 1, and Assumptions 3, 4, 5 in the Appendix hold. Then, the KODE correctly recovers the true regulatory network, in that, for all $j = 1, \ldots, p$,

$$\mathbb{P}\left(\widehat{S}_j = S_j^0\right) \to 1, \quad \text{as } n \to \infty.$$

5. Simulation Studies

5.1. Setup

We study the empirical performance of the proposed KODE using two ODE examples, the enzyme regulatory network in Section 5.2, and the Lotka–Volterra equations in Section 5.3. For a given system of ODEs and the initial condition, we obtain the numerical solutions of the ODEs using the Euler method with step size 0.01. The data observations are drawn from the solutions at an evenly spaced time grid, with measurement errors. To implement KODE, we fit the smoothing spline to estimate $x_j(t)$ in (9) using a Matérn kernel, $K_{\mathcal{F}}(x, x') = (1 + \sqrt{3}\|x - x'\|/\nu)\exp(-\sqrt{3}\|x - x'\|/\nu)$, where the smoothing parameter λ_{nj} is chosen by GCV, and the bandwidth ν is chosen by 10-fold cross-validation. We compute the integral $\int_0^{t_i} F_j(\widehat{x}(t))dt$ in (10) numerically with independent sets of 1000 Monte Carlo points. We compare KODE with linear ODE with interactions in (4) (Zhang et al. 2015), and additive ODE in (6) (Chen, Shojaie, and Witten 2017). Due to the lack of available code online, we implement the two competing methods in the framework of Algorithm 1, using a linear kernel for (6), and using an additive Matérn kernel for (6). We evaluate the performance using the prediction error, plus the false discovery proportion and power for edge selection of the corresponding regulatory network. Furthermore, we compare with the family of ODE solutions assuming a known F (Zhang, Cao, and Carroll 2015; Mikkelsen and Hansen 2017) in Section S2.1 of the Appendix. We also carry out a sensitivity analysis in Section S2.2 of the Appendix to study the robustness of the choice of kernel function and initial parameters.

5.2. Enzymatic Regulatory Network

The first example is a three-node enzyme regulatory network of a negative feedback loop with a buffering node (Ma et al. 2009, NFBLB). The ODE system is given in (7) in Section 2.1. Figure 1(a) shows the NFBLB network diagram consisting of the three interacting nodes: x_1 receives the input, x_3 transmits the output, and x_2 plays a regulatory role, leading a negative regulatory link to x_3. We note that, although biological circuits can have more than three nodes, many of those circuits can be reduced to a three-node framework, given that multiple molecules often function as a single virtual node. Moreover, despite the diversity of possible network topologies, NFBLB is one of the two core three-node topologies that could perform adaption in the sense that the system resets itself after responding to a stimulus; see Ma et al. (2009) for more discussion of NFBLB. For the ODE system in (7), we set the catalytic rate parameters of the enzymes as $c_1 = c_2 = c_3 = c_5 = c_6 = 10, c_4 = 1$, the Michaelis–Menten constants as $C_1 = \cdots = C_6 = 0.1$, and the concentration parameters of enzymes as $\tilde{c}_1 = 1, \tilde{c}_2 = 0.2$. These parameters achieve the adaption as shown in Figure 1(b). The output node x_3 shows a strong initial response to the stimulus, and also exhibits strong adaption, since its post-stimulus steady state $x_3 = 0.052$ is close to the prestimulus state $x_3 = 0$. The input $x_0 \in \mathbb{R}^3$ is drawn uniformly from $[0.5, 1.5]$, with the initial value $x(0) = 0$, and the measurement errors are iid normal with mean zero and variance σ_j^2. The time points are evenly

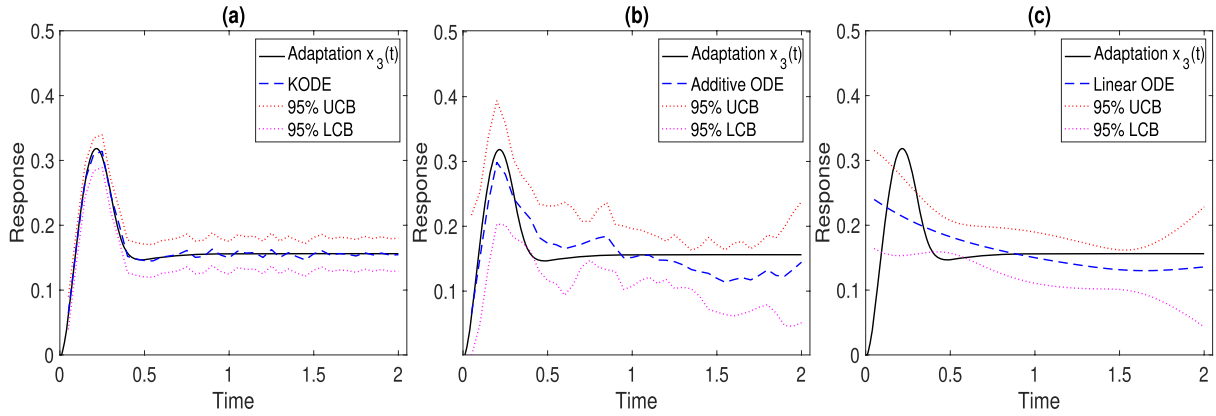

Figure 2. The true (black solid line) and the estimated (blue dashed line) trajectory of $x_3(t)$, with the 95% upper and lower confidence bounds (red dotted lines). The results are averaged over 500 data replications. (a) KODE; (b) additive ODE; (c) linear ODE.

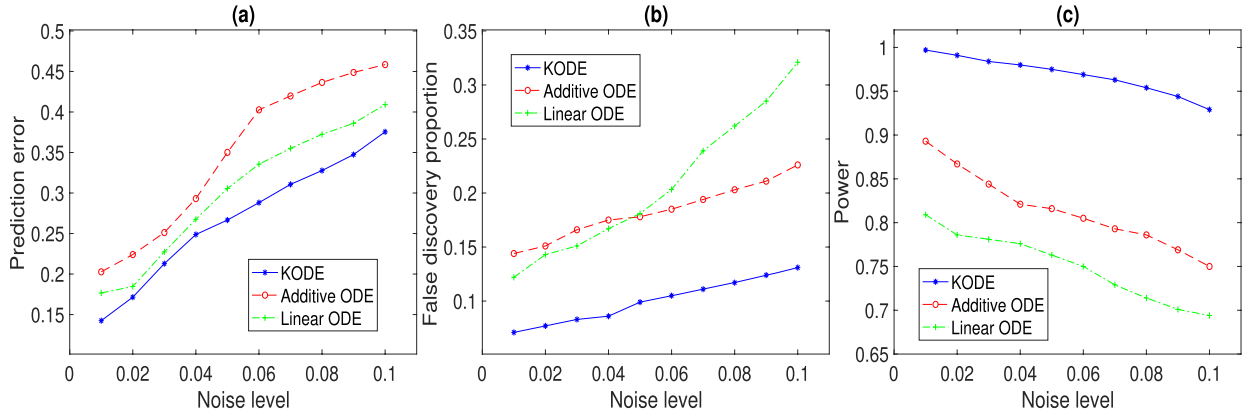

Figure 3. The prediction and selection performance of three ODE methods with varying noise level. The results are averaged over 500 data replications. (a) Prediction error; (b) false discovery proportion; (c) empirical power.

distributed, $t_i = (i-1)/20, i = 1, \ldots, n$. In this example, $p = 3$, and for each function $x_j(t), j = 1, 2, 3$, there are $p^2 = 9$ functions to estimate, and in total there are 27 functions to estimate under the sample size $n = 40$.

Figure 2 reports the true and estimated trajectory of $x_3(t)$, with 95% upper and lower confidence bounds, of the three ODE methods, where we use the tensor product Matérn kernel for KODE in (10). The noise level is set as $\sigma_j = 0.1, j = 1, 2, 3$, and the results are averaged over 500 data replications. It is seen that the KODE estimate has a smaller variance than the additive and linear ODE estimates. Moreover, the confidence interval of KODE achieves the desired coverage for the true trajectory. In contrast, the confidence intervals of additive and linear ODE models mostly fail to include the truth. This is because there is a discrepancy between the additive and linear ODE model specifications and the true ODE model in (7), and this discrepancy accumulates as the course of the ODE evolves.

Figure 3 reports the prediction and selection performance of the three ODE methods, with varying noise level $\sigma_j \in \{0.01, 0.02, \ldots, 0.1\}, j = 1, 2, 3$. The results are averaged over 500 data replications. The prediction error is defined as the squared root of the sum of predictive mean squared errors for $x_1(t), x_2(t), x_3(t)$ at the unseen "future" time point $t = 2$. The false discovery proportion is defined as the proportion of falsely selected edges in the regulatory network out of the total number of edges. The empirical power is defined as the proportion of selected true edges in the network. It is seen that KODE clearly outperforms the two alternative solutions in both prediction and selection accuracy. Moreover, we report graphically the sparse recovery of this regulatory network in Section S2.3 of the Appendix.

5.3. Lotka–Volterra Equations

The second example is the high-dimensional Lotka–Volterra equations, which are pairs of first-order nonlinear differential equations describing the dynamics of biological systems in which predators and prey interact (Volterra 1928). We consider a 10-node system,

$$\begin{aligned} \frac{dx_{2j-1}(t)}{dt} &= \alpha_{1,j}x_{2j-1}(t) - \alpha_{2,j}x_{2j-1}(t)x_{2j}(t), \\ \frac{dx_{2j}(t)}{dt} &= \alpha_{3,j}x_{2j-1}(t)x_{2j}(t) - \alpha_{4,j}x_{2j}(t), \end{aligned} \tag{18}$$

where $\alpha_{1,j} = 1.1 + 0.2(j-1), \alpha_{2,j} = 0.4 + 0.2(j-1), \alpha_{3,j} = 0.1 + 0.2(j-1)$, and $\alpha_{4,j} = 0.4 + 0.2(j-1), j = 1, \ldots, 5$. The parameters $\alpha_{2,j}$ and $\alpha_{3,j}$ define the interaction between the two populations such that $dx_{2j-1}(t)/dt$ and $dx_{2j}(t)/dt$ are nonadditive functions of x_{2j-1} and x_{2j}, where x_{2j-1} is the prey and x_{2j} is the predator. Figure 4(a) shows an illustration of the interaction between $x_1(t)$ and $x_2(t)$. The input $x_0 \in \mathbb{R}^{10}$ is drawn uniformly from $[5, 15]^{10}$, with the initial value $x_{2j-1}(0) =$

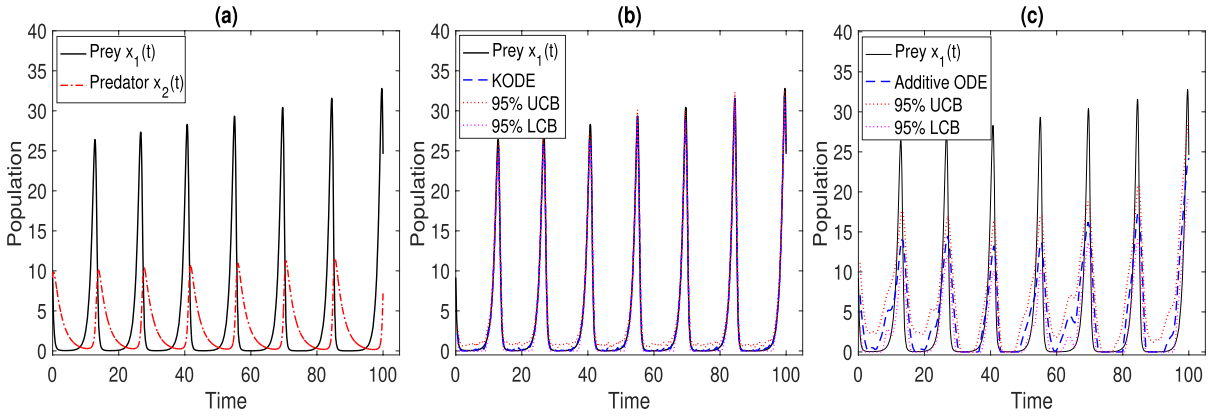

Figure 4. (a) The true trajectories of the prey $x_1(t)$ and the predator $x_2(t)$. (b) The estimated trajectory $\widehat{x}_1(t)$ (blue dashed line), with the 95% upper and lower confidence bounds (red dotted lines), by KODE. (c) By additive ODE. The results are averaged over 500 data replications.

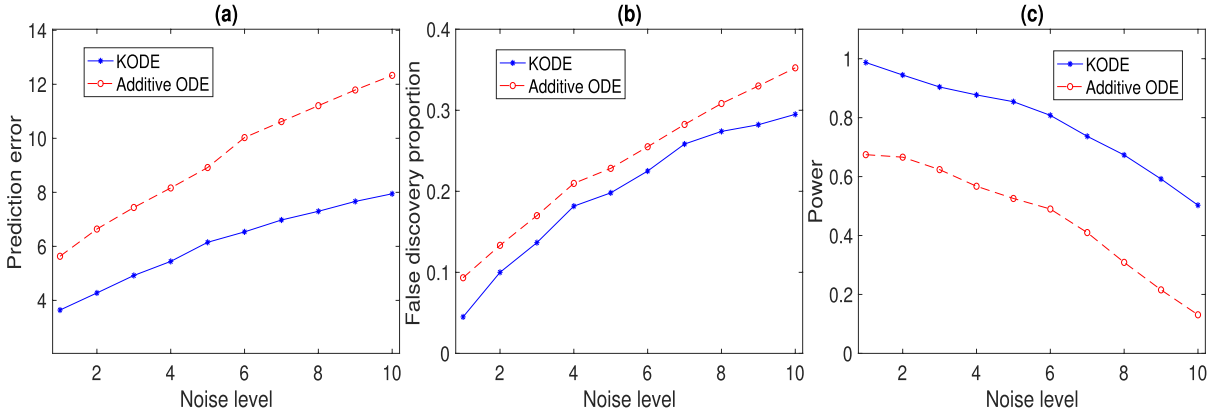

Figure 5. The prediction and selection performance of two ODE methods with varying noise level. The results are averaged over 500 data replications. (a) Prediction error; (b) false discovery proportion; (c) empirical power.

$x_{2j}(0)$, and the measurement errors are iid normal $N(0,\sigma_j^2)$, where σ_j again reflects the noise level. The time points are evenly distributed in $[0,100]$ with $n = 200$. In this example, $p = 10$, and for each function $x_j(t)$, $j = 1,\ldots,10$, there are $p^2 = 100$ functions to estimate, and in total there are 1000 functions to estimate under the sample size $n = 200$.

Figure 4(b) and (c) report the estimated trajectory of $x_1(t)$, with 95% upper and lower confidence bounds, of KODE and additive ODE, where the noise level is set as $\sigma_j = 1, j = 1,\ldots,10$. The confidence interval of KODE achieves a better empirical coverage for the true trajectory compared to that of additive ODE. For this example, we use the linear kernel for KODE in (10), since the functional forms in (18) are known to be linear. For this reason, we only compare KODE with the additive ODE method. Moreover, in the implementation, the estimates $\widehat{F}_j(\widehat{x}(t))$ are thresholded to be nonnegative to ensure the physical constraint that the number of population cannot be negative. Figure 5 reports the prediction and selection performance of the two ODE methods, with varying noise level $\sigma_j \in \{1,2,\ldots,10\}, j = 1,\ldots,10$. All the results are averaged over 500 data replications. It is seen that the KODE estimate achieves a smaller prediction error, and a higher selection accuracy, since KODE allows flexible nonadditive structures, which results in significantly smaller bias and variance in functional estimation as compared to the additive modeling.

6. Application to Gene Regulatory Network

We illustrate KODE with a gene regulatory network application. Schaffter, Marbach, and Floreano (2011) developed an open-source platform called GeneNetWeaver (GNW) that generates in silico benchmark gene expression data using dynamical models of gene regulations and nonlinear ODEs. The generated data have been used for evaluating the performance of network inference methods in the DREAM3 competition (Marbach et al. 2009), and were also analyzed by Henderson and Michailidis (2014); Chen, Shojaie, and Witten (2017) in additive ODE modeling. GNW extracts two regulatory networks of *E.coli* (*E.coli1*, *E.coli2*), and three regulatory networks of yeast (yeast1, yeast2, yeast 3), each of which has two dimensions, $p = 10$ nodes and $p = 100$ nodes. This yields totally 10 combinations of network structures. Figures 6(a) and (b) show an example of the 10-node and the 100-node *E.coli1* networks, respectively. The systems of ODEs for each extracted network are based on a thermodynamic approach, which leads to a nonadditive and nonlinear ODE structure (Marbach et al. 2010). Besides, the network structures are sparse; for example, for the 10-node *E.coli1* network, there are 11 edges out of 90 possible pairwise edges, and for the 100-node *E.coli1* network, there are 125 edges out of 9900 possible pairwise edges. Moreover, for the 10-node network, GNW provides $R = 4$ perturbation

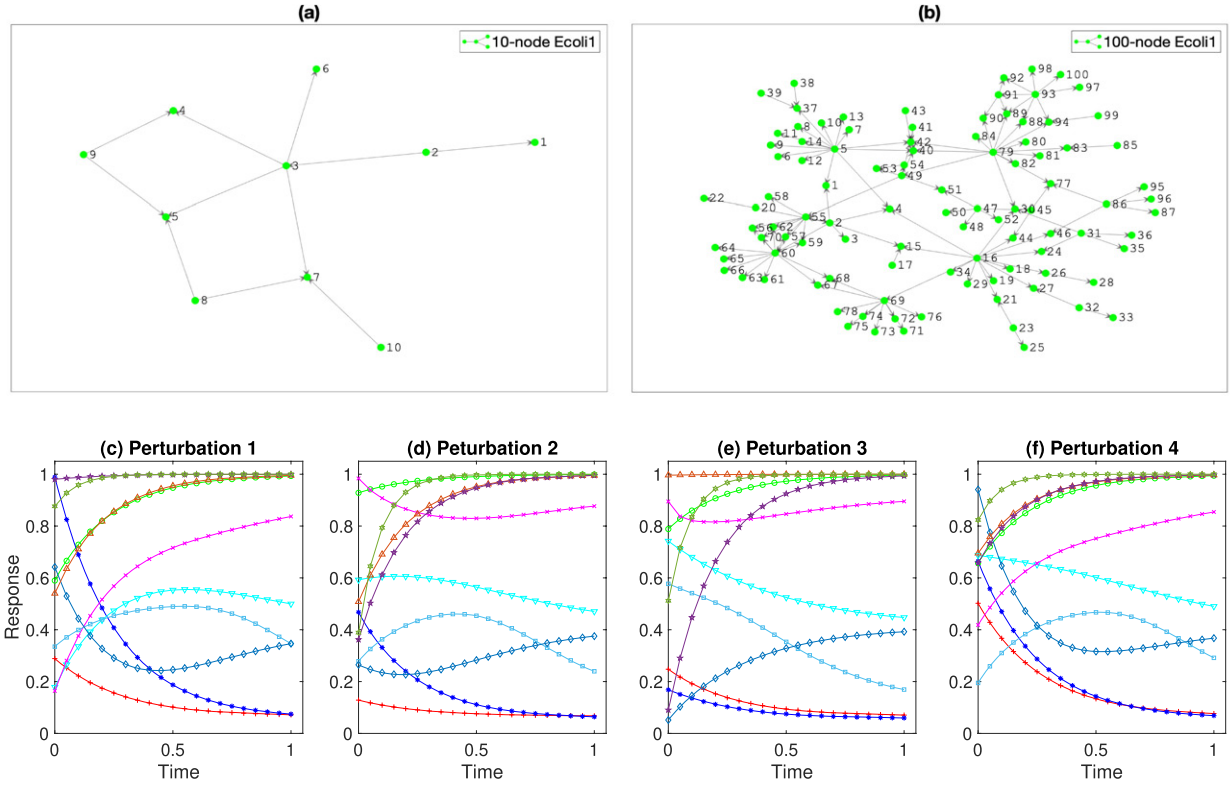

Figure 6. (a) The 10-node *E.coli1* network. (b) The 100-node *E.coli1* network. (c–f) Four perturbation experiments for the 10-node *E.coli1* network, where each experiment corresponds to a different initial condition of the ODE system.

experiments, and for the 100-node network, GNW provides $R = 46$ experiments. In each experiment, GNW generates the time-course data with different initial conditions of the ODE system to emulate the diversity of gene expression trajectories (Marbach et al. 2009). Figures 6(c)–(f) show the time-course data under $R = 4$ experiments for the 10-node *E.coli1* network. All the trajectories are measured at $n = 21$ evenly spaced time points in $[0, 1]$. We add independent measurement errors from a normal distribution with mean zero and standard deviation 0.025, which is the same as the DREAM3 competition and the data analysis done in Henderson and Michailidis (2014) and Chen, Shojaie, and Witten (2017).

The kernel ODE model we have developed focuses on a single experiment data, but it can be easily generalized to incorporate multiple experiments. Specifically, let $\{y_{ij}^{(r)}; i = 1, \ldots, n, j = 1, \ldots, p, r = 1, \ldots, R\}$ denote the observed data from n subjects for p variables under R experiments, with unknown initial conditions $x^{(r)}(0) \in \mathbb{R}^p, r = 1, \ldots, R$. Then we modify the KODE method in (9) and (10), by seeking $F_j \in \mathcal{H}$ and $\theta_{j0} \in \mathbb{R}$ that minimize

$$\frac{1}{Rn}\sum_{r=1}^{R}\sum_{i=1}^{n}\left\{y_{ij}^{(r)} - \theta_{j0} - \int_0^{t_i} F_j(\widehat{x}^{(r)}(t))dt\right\}^2 + \tau_{nj}\left(\sum_{k=1}^{p}\|\mathcal{P}^k F_j\|_{\mathcal{H}} + \sum_{k\neq l,k=1}^{p}\sum_{l=1}^{p}\|\mathcal{P}^{kl}F_j\|_{\mathcal{H}}\right), \quad (19)$$

where $\widehat{x}^{(r)}(t) = (\widehat{x}_1^{(r)}(t), \ldots, \widehat{x}_p^{(r)}(t))^\top$ is the smoothing spline estimate obtained by,

$$\widehat{x}_j^{(r)}(t) = \arg\min_{z_j \in \mathcal{F}}\left\{\frac{1}{n}\sum_{i=1}^{n}(y_{ij}^{(r)} - z_j(t_i))^2 + \lambda_{nj}\|z_j(t)\|_{\mathcal{F}}^2\right\}, \quad j = 1, \ldots, p,\ r = 1, \ldots, R.$$

Algorithm 1 can be modified accordingly to work with multiple experiments.

We again compare KODE with the additive ODE (Chen, Shojaie, and Witten 2017) and the linear ODE (Zhang et al. 2015), adopting the same implementation as in the simulations. Since we know the true edges of the generated gene regulatory networks, we use the area under the ROC curve (AUC) as the evaluation criterion. Table 1 reports the results averaged over 100 data realizations for all ten combinations of network structures. It is clearly seen that KODE outperforms both alternative methods in all cases. We further report graphically the sparse recovery of the 10-node *E.coli1* network in Section S2.4 of the Appendix. This example shows that our proposed KODE is a competitive and useful tool for ODE modeling. In addition, it also shows that the proposed method can scale up and work with reasonably large networks. For instance, for the network with $p = 100$ nodes, there are $p^2 = 10{,}000$ functions to estimate, and the sample size is $n = 21$ with $R = 46$ perturbations.

7. Conclusion and Discussion

In this article, we have developed a new reproducing kernel-based approach for a general family of ODE models that learn a

Table 1. The area under the ROC curve, and the 95% confidence interval, for 10 combinations of network structures from GNW.

	$p = 10$			$p = 100$		
	KODE	Additive ODE	Linear ODE	KODE	Additive ODE	Linear ODE
E.coli1	**0.582** (0.577, 0.587)	0.541 (0.535, 0.547)	0.460 (0.453, 0.467)	**0.711** (0.708, 0.714)	0.677 (0.672, 0.682)	0.640 (0.637, 0.643)
E.coli2	**0.662** (0.658, 0.666)	0.632 (0.625, 0.639)	0.562 (0.555, 0.569)	**0.685** (0.681, 0.689)	0.659 (0.652, 0.666)	0.533 (0.527, 0.539)
Yeast1	**0.603** (0.599, 0.607)	0.541 (0.536, 0.546)	0.436 (0.430, 0.442)	**0.619** (0.616, 0.622)	0.589 (0.581, 0.597)	0.569 (0.562, 0.576)
Yeast2	**0.599** (0.595, 0.603)	0.562 (0.555, 0.570)	0.536 (0.530, 0.542)	**0.606** (0.603, 0.609)	0.588 (0.582, 0.594)	0.541 (0.536, 0.546)
Yeast3	**0.612** (0.608, 0.616)	0.569 (0.564, 0.573)	0.487 (0.481, 0.493)	**0.621** (0.617, 0.625)	0.613 (0.607, 0.619)	0.609 (0.605, 0.613)

NOTE: The results are averaged over 100 data replications. Boldface indicates the method with larger AUC.

dynamic system from noisy time-course data. We employ sparsity regularization to select individual functionals and uncover the underlying regulatory network, and we derive the post-selection confidence interval for the estimated signal trajectory. Our proposal is built upon but also extends the smoothing spline analysis of variance framework. We establish the theoretical properties of the method, while allowing the number of functionals to be either smaller or larger than the number of time points.

In numerous scientific applications, ODE is often employed to understand the regulatory effects and causal mechanisms within a dynamic system under interventions. Our proposed KODE method can be applied for this very purpose. There are different formulations of causal modeling for dynamic systems in the literature. We next consider and illustrate with two relatively common scenarios, one regarding dynamic causal modeling (DCM) under experimental stimuli (Friston, Harrison, and Penny 2003), and the other about kinetic modeling that is invariant across heterogeneous experiments (Pfister, Bauer, and Peters 2019).

The first scenario concerns DCM that infers the regulatory effects within a dynamic system under experimental stimuli (Friston, Harrison, and Penny 2003). Specifically, the DCM characterizes the variations of the state variables $x(t) = (x_1(t), \ldots, x_p(t))^\top \in \mathbb{R}^p$ under the stimulus inputs $u(t) = (u_1(t), \ldots, u_q(t))^\top \in \mathbb{R}^q$ via a set of ODEs, $dx(t)/dt = F(x(t), u(t))$, where the functional F is modeled by a bilinear form,

$$F_j(x(t), u(t)) = \theta_{j0} + \sum_{k=1}^{p} \theta_{jk}^{(1)} x_k(t) + \sum_{l=1}^{q} \theta_{jl}^{(2)} u_l(t) + \sum_{k=1}^{p}\sum_{l=1}^{q} \theta_{jkl}^{(1,2)} x_k(t) u_l(t), \quad j = 1, \ldots, p. \tag{20}$$

In this model, $\theta_{jk}^{(1)} \in \mathbb{R}$ reflects the strength of intrinsic connection from $x_k(t)$ to $x_j(t)$, $\theta_{jl}^{(2)} \in \mathbb{R}$ reflects the effect of the lth input stimulus $u_l(t)$ on $x_j(t)$, and $\theta_{jkl}^{(1,2)} \in \mathbb{R}$ reflects the influence of $u_l(t)$ on the directional connection between $x_k(t)$ and $x_j(t)$, $j, k = 1, \ldots, p, l = 1, \ldots, q$. Note that $\theta_{jk}^{(1)}$ and $\theta_{kj}^{(1)}$ can be different, and thus the effect from $x_k(t)$ to $x_j(t)$ and that from $x_j(t)$ to $x_k(t)$ can be different. Similarly, $\theta_{jkl}^{(1,2)}$ and $\theta_{kjl}^{(1,2)}$ can be different. As such, model (20) encodes a directional network, and under certain conditions, a causal network. DCM has been widely used in biology and neuroscience (see, e.g., Friston, Harrison, and Penny 2003; Zhang et al. 2015, 2017; Cao, Sandstede, and Luo 2019).

We can combine the proposed KODE with the DCM model (20) straightforwardly. Such a combination allows us to estimate and infer the causal regulatory effects under experimental stimuli without specifying the forms of the functionals F. This is appealing, as there have been evidences suggesting that the regulatory effects can be nonlinear (Buxton et al. 2004; Friston et al. 2019). More specifically, we model F such that,

$$F_j(x(t), u(t)) = \theta_{j0} + \sum_{k=1}^{p} F_{jk}^{(1)}(x_k(t)) + \sum_{l=1}^{q} F_{jl}^{(2)}(u_l(t)) + \sum_{k=1}^{p}\sum_{l=1}^{q} F_{jkl}^{(1,2)}(x_k(t), u_l(t)), \ j = 1, \ldots, p. \tag{21}$$

Similar as the tensor product space defined in (8), let $\mathcal{H}_k^{(1)}$ and $\mathcal{H}_l^{(2)}$ denote the space of functions of $x_k(t)$ and $u_l(t)$ with zero marginal integral, respectively. We impose that the functionals $F_j, j = 1, \ldots, p$ in (21) are located in the following space,

$$\mathcal{H} = \{1\} \oplus \sum_{k=1}^{p} \mathcal{H}_k^{(1)} \oplus \sum_{l=1}^{q} \mathcal{H}_l^{(2)} \oplus \sum_{k=1}^{p}\sum_{l=1}^{q} \left(\mathcal{H}_k^{(1)} \otimes \mathcal{H}_l^{(2)}\right).$$

Parallel to (20), the functions $F_{jk}^{(1)}$, $F_{jl}^{(2)}$, and $F_{jkl}^{(1,2)}$ in (21) capture the regulatory effects, and together, they encode a directional network. Moreover, Algorithm 1 of KODE is directly applicable to estimate $F_{jk}^{(1)}, F_{jl}^{(2)}$, and $F_{jkl}^{(1,2)}$. As we have shown in our simulations, the DCM model (21) based on KODE should outperform (20) that is based on linear ODE.

The second scenario concerns learning the causal structure of kinetic systems by identifying a stable model from noisy observations generated from heterogeneous experiments. Pfister, Bauer, and Peters (2019) proposed the CausalKinetiX method, where the main idea is to optimize a noninvariance score to identify a causal ODE model that is invariant across heterogeneous experiments. Again, we can combine the proposed KODE with CausalKinetiX to learn the causal structure, while balancing between predictability and causality of the ODE model, and extending from a linear ODE model to a more flexible ODE model. We refer to this integrated method as KODE-CKX.

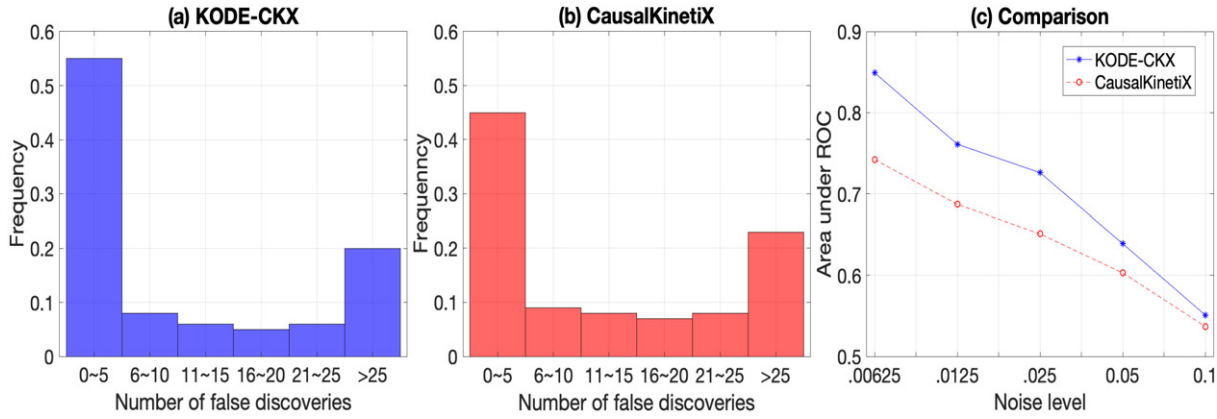

Figure 7. The selection performance of KODE-CKX and CausalKinetiX. The results are averaged over 100 data replications. (a) Number of false discoveries in the estimated model based on KODE-CKX; (b) number of false discoveries in the estimated model based on CausalKinetiX; (c) area under ROC under different noise levels.

More specifically, consider R heterogeneous experiments, which stem from interventions such as manipulations of initial or environmental conditions. Following Algorithm 1 of KODE, we obtain $\widehat{\theta}_j^{(r)}$ for each experiment $r \in \{1,\ldots,R\}$, and $j = 1,\ldots,p$. Let $M_j^{(r)} \subseteq \mathcal{M}$ denote the index set of the nonzero entries of the sparse estimator $\widehat{\theta}_j^{(r)}$. We propose the following four-step procedure to score each model $M_j^{(r)}$. In the first step, we obtain the smoothing spline estimate $\widehat{x}_j^{(r)}(t)$ by (9) using the data from the rth experiment. In the second step, we apply Algorithm 1 to compute $\widehat{F}_j^{(r)}$, by setting $\kappa_{nj} = 0$, restricting $\theta_j \in M_j^{(r)}$, and using the data from all other experiments except for the rth experiment. Here leaving out the rth experiment is to ensure a good generalization capability. In the third step, we estimate the signal trajectory under the derivative constraint,

$$\widetilde{x}_j^{(r)}(t) = \underset{z_j \in \mathcal{F}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left\{ y_{ij} - z_j(t_i) \right\}^2 + \lambda_{nj} \|z_j(t)\|_{\mathcal{F}}^2 \right\},$$

$$\text{such that } \widetilde{x}_j^{(r)}(t_i) = \widehat{F}_j^{(r)}(\widehat{x}_j^{(r)}(t_i)), \tag{22}$$

for $i = 1,\ldots,n, j = 1,\ldots,p$. In the last step, similar as CausalKinetiX, we obtain for each model $M_j^{(r)} \subseteq \mathcal{M}$ the noninvariance score,

$$\text{NS}\left(M_j^{(r)}\right) \equiv \frac{1}{R} \sum_{r=1}^{R} \frac{\text{RSS}_B^{(r)} - \text{RSS}_A^{(r)}}{\text{RSS}_A^{(r)}},$$

where $\text{RSS}_A^{(r)} = n^{-1} \sum_{i=1}^{n} \left\{ y_{ij}^{(r)} - \widehat{x}_j^{(r)}(t_{ij}) \right\}^2$, and $\text{RSS}_B^{(r)} = n^{-1} \sum_{i=1}^{n} \left\{ y_{ij}^{(r)} - \widetilde{x}_j^{(r)}(t_{ij}) \right\}^2$ are the residual sums of squares based on $\widehat{x}_j^{(r)}(t)$ and $\widetilde{x}_j^{(r)}(t)$, respectively. Due to the additional constraint in (22), $\text{RSS}_B^{(r)}$ is always larger than $\text{RSS}_A^{(r)}$. Following Pfister, Bauer, and Peters (2019), the model $M_j^{(r)} \subseteq \mathcal{M}$ with a small score $\text{NS}(M_j^{(r)})$ is predictive and invariant. Such an invariant ODE model allows researchers to predict the behavior of the dynamic system under interventions, and it is closely related to the causal mechanism of the underlying dynamic system from the structural casual model and modularity perspective (Rubenstein et al. 2018; Pfister, Bauer, and Peters 2019). Compared to CausalKinetiX, our proposed KODE-CKX further extends the linear ODE to a general class of nonlinear and nonadditive ODE.

To verify the empirical performance of KODE-CKX and to compare with CausalKinetiX, we consider the 100-node *E.coli1* gene regulatory network example in Section 6. Figure 7 compares the models with the smallest noninvariance score from KODE-CKX and CausalKinetiX, respectively, based on 100 data replications. Comparing Figures 7(a) and (b), it is seen that in the majority of cases, KODE-CKX is able to recover the causal parents, and it outperforms CausalKinetiX by achieving a smaller number of false discoveries. Here, the measurement errors were drawn from a normal distribution with mean zero and standard deviation 0.025, the same setup as in Section 6. We next further evaluate the performance of the two methods when we vary the standard deviation of the measurement errors. Figure 7(c) reports the AUC averaged over 100 data replications. It is seen again that, for all noise levels, KODE-CKX performs better than CausalKinetiX.

In summary, our proposed KODE is readily applicable to numerous scenarios to facilitate the understanding of the regulatory causal mechanisms within a dynamic system from noisy data under interventions.

Supplementary Materials

The supplementary material contains proofs and additional numerical results for the main article.

Acknowledgments

We thank the Editor, the Associate Editor, and two referees for their constructive comments and suggestions.

Funding

Xiaowu Dai's research was partially supported by CDAR, Department of Economics, the University of California, Berkeley, and this work was done while Xiaowu Dai was visiting the Simons Institute for the Theory of Computing. Lexin Li's research was partially supported by NSF grant DMS-1613137, and NIH grants R01AG061303, R01AG062542, and R01AG034570.

References

Bachoc, F., Leeb, H., and Pötscher, B. M. (2019), "Valid Confidence Intervals for Post-Model-Selection Predictors," *The Annals of Statistics*, 47, 1475–1504. [1716]

Berk, R., Brown, L., Buja, A., Zhang, K., and Zhao, L. (2013), "Valid Post-Selection Inference," *The Annals of Statistics*, 41, 802–837. [1716]

Brunton, S. L., Proctor, J. L., and Kutz, J. N. (2016), "Discovering Governing Equations From Data by Sparse Identification of Nonlinear Dynamical Systems," *Proceedings of the National Academy of Sciences of the United States of America*, 113, 3932–3937. [1718]

Buxton, R. B., Uludağ, K., Dubowitz, D. J., and Liu, T. T. (2004), "Modeling the Hemodynamic Response to Brain Activation," *Neuroimage*, 23, S220–S233. [1722]

Cao, J., and Zhao, H. (2008), "Estimating Dynamic Models for Gene Regulation Networks," *Bioinformatics*, 24, 1619–1624. [1711]

Cao, X., Sandstede, B., and Luo, X. (2019), "A Functional Data Method for Causal Dynamic Network Modeling of Task-Related fMRI," *Frontiers in Neuroscience*, 13, 127. [1711,1722]

Chen, S., Shojaie, A., and Witten, D. M. (2017), "Network Reconstruction From High-Dimensional Ordinary Differential Equations," *Journal of the American Statistical Association*, 112, 1697–1707. [1712,1713,1714,1718,1720,1721]

Chernozhukov, V., Hansen, C., and Spindler, M. (2015), "Valid Post-Selection and Post-Regularization Inference: An Elementary, General Approach," *Annual Review of Economics*, 7, 649–688. [1717]

Chou, I.-C., and Voit, E. O. (2009), "Recent Developments in Parameter Estimation and Structure Identification of Biochemical and Genomic Systems," *Mathematical Biosciences*, 219, 57–83. [1711,1716]

Cox, D. D. (1983), "Asymptotics for M-Type Smoothing Splines," *The Annals of Statistics*, 11, 530–551. [1718]

Dattner, I., and Klaassen, C. A. J. (2015), "Optimal Rate of Direct Estimators in Systems of Ordinary Differential Equations Linear in Functions of the Parameters," *Electronic Journal of Statistics*, 9, 1939–1973. [1712,1713,1714]

Friston, K. J., Harrison, L., and Penny, W. (2003), "Dynamic Causal Modelling," *Neuroimage*, 19, 1273–1302. [1722]

Friston, K. J., Preller, K. H., Mathys, C., Cagnan, H., Heinzle, J., Razi, A., and Zeidman, P. (2019), "Dynamic Causal Modelling Revisited," *Neuroimage*, 199, 730–744. [1722]

González, J., Vujačić, I., and Wit, E. (2014), "Reproducing Kernel Hilbert Space Based Estimation of Systems of Ordinary Differential Equations," *Pattern Recognition Letters*, 45, 26–32. [1712]

Gu, C. (2013), *Smoothing Spline ANOVA Models*, New York: Springer-Verlag. [1714,1715]

Henderson, J., and Michailidis, G. (2014), "Network Reconstruction Using Nonparametric Additive ODE Models," *PLOS ONE*, 9, 1–15. [1711,1712,1713,1720,1721]

Huang, J. Z. (1998), "Projection Estimation in Multiple Regression With Application to Functional ANOVA Models," *The Annals of Statistics*, 26, 242–272. [1712]

Izhikevich, E. (2007), *Dynamical Systems in Neuroscience*, Cambridge, MA: MIT Press. [1711,1716]

Javanmard, A., and Montanari, A. (2014), "Confidence Intervals and Hypothesis Testing for High-Dimensional Regression," *Journal of Machine Learning Research*, 15, 2869–2909. [1717]

Koltchinskii, V., and Yuan, M. (2010), "Sparsity in Multiple Kernel Learning," *The Annals of Statistics*, 38, 3660–3695. [1717]

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. [1711]

Lin, Y. (2000), "Tensor Product Space ANOVA Models," *The Annals of Statistics*, 28, 734–755. [1717]

Lin, Y., and Zhang, H. H. (2006), "Component Selection and Smoothing in Multivariate Nonparametric Regression," *The Annals of Statistics*, 34, 2272–2297. [1712,1714,1715,1717]

Loh, P.-L., and Wainwright, M. J. (2012), "High-Dimensional Regression With Noisy and Missing Data: Provable Guarantees With Nonconvexity," *The Annals of Statistics*, 40, 1637–1664. [1717]

Lu, J., Kolar, M., and Liu, H. (2020), "Kernel Meets Sieve: Post-Regularization Confidence Bands for Sparse Additive Model," *Journal of the American Statistical Association*, 115, 2084–2099. [1717]

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. [1711,1712]

Ma, W., Trusina, A., El-Samad, H., Lim, W. A., and Tang, C. (2009), "Defining Network Topologies That Can Achieve Biochemical Adaptation," *Cell*, 138, 760–773. [1711,1713,1716,1718]

Marbach, D., Prill, R. J., Schaffter, T., Mattiussi, C., Floreano, D., and Stolovitzky, G. (2010), "Revealing Strengths and Weaknesses of Methods for Gene Network Inference," *Proceedings of the National Academy of Sciences of the United States of America*, 107, 6286–6291. [1720]

Marbach, D., Schaffter, T., Mattiussi, C., and Floreano, D. (2009), "Generating Realistic in Silico Gene Networks for Performance Assessment of Reverse Engineering Methods," *Journal of Computational Biology*, 16, 229–239. [1720,1721]

Mikkelsen, F. V., and Hansen, N. R. (2017), "Learning Large Scale Ordinary Differential Equation Systems," arXiv no. 1710.09308. [1712,1718]

Opsomer, J. D., and Ruppert, D. (1997), "Fitting a Bivariate Additive Model by Local Polynomial Regression," *The Annals of Statistics*, 25, 186–211. [1712]

Pfister, N., Bauer, S., and Peters, J. (2019), "Learning Stable and Predictive Structures in Kinetic Systems," *Proceedings of the National Academy of Sciences of the United States of America*, 116, 25405–25411. [1722,1723]

Raskutti, G., Wainwright, M. J., and Yu, B. (2011), "Minimax Rates of Estimation for High-Dimensional Linear Regression Over ℓ_q-Balls," *IEEE Transactions on Information Theory*, 57, 6976–6994. [1717]

Ravikumar, P., Wainwright, M. J., and Lafferty, J. (2010), "High-Dimensional Ising Model Selection Using l_1-Regularized Logistic Regression," *The Annals of Statistics*, 38, 1287–1319. [1718]

Rubenstein, P. K., Bongers, S., Schölkopf, B., and Mooij, J. M. (2018), "From Deterministic ODEs to Dynamic Structural Causal Models," in *Proceedings of the 34th Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI)*. [1723]

Schaffter, T., Marbach, D., and Floreano, D. (2011), "GeneNetWeaver: In Silico Benchmark Generation and Performance Profiling of Network Inference Methods," *Bioinformatics*, 27, 2263–2270. [1720]

Talagrand, M. (1996), "New Concentration Inequalities in Product Spaces," *Inventiones Mathematicae*, 126, 505–563. [1717]

Tzafriri, A. R. (2003), "Michaelis–Menten Kinetics at High Enzyme Concentrations," *Bulletin of Mathematical Biology*, 65, 1111–1129. [1713]

van der Vaart, A. W., and Wellner, J. A. (1996), *Weak Convergence and Empirical Processes*, New York: Springer-Verlag. [1717]

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. [1713]

Volterra, V. (1928), "Variations and Fluctuations of the Number of Individuals in Animal Species Living Together," *ICES Journal of Marine Science*, 3, 3–51. [1719]

Wahba, G. (1983), "Bayesian 'Confidence Intervals' for the Cross-Validated Smoothing Spline," *Journal of the Royal Statistical Society*, Series B, 45, 133–150. [1712]

——— (1990), *Spline Models for Observational Data*, Philadelphia, PA: SIAM. [1714,1715,1716]

Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1995), "Smoothing Spline ANOVA for Exponential Families, With Application to the Wisconsin Epidemiological Study of Diabetic Retinopathy," *The Annals of Statistics*, 23, 1865–1895. [1712,1714,1715]

Wang, S., Nan, B., Zhu, N., and Zhu, J. (2009), "Hierarchically Penalized Cox Regression With Grouped Variables," *Biometrika*, 96, 307–322. [1714]

Wu, H., Lu, T., Xue, H., and Liang, H. (2014), "Sparse Additive Ordinary Differential Equations for Dynamic Gene Regulatory Network Modeling," *Journal of the American Statistical Association*, 109, 700–716. [1711,1712,1713,1718]

Yuan, M., and Zhou, D.-X. (2016), "Minimax Optimal Rates of Estimation in High Dimensional Additive Models," *The Annals of Statistics*, 44, 2564–2593. [1717]

Zhang, C.-H., and Zhang, S. S. (2014), "Confidence Intervals for Low Dimensional Parameters in High Dimensional Linear Models," *Journal of the Royal Statistical Society*, Series B, 76, 217–242. [1717]

Zhang, T., Wu, J., Li, F., Caffo, B., and Boatman-Reich, D. (2015), "A Dynamic Directional Model for Effective Brain Connectivity Using Electrocorticographic (ECoG) Time Series," *Journal of the American Statistical Association*, 110, 93–106. [1711,1717,1718,1721,1722]

Zhang, T., Yin, Q., Caffo, B., Sun, Y., and Boatman-Reich, D. (2017), "Bayesian Inference of High-Dimensional, Cluster-Structured Ordinary Differential Equation Models With Applications to Brain Connectivity Studies," *The Annals of Applied Statistics*, 11, 868–897. [1711,1722]

Zhang, X., Cao, J., and Carroll, R. J. (2015), "On the Selection of Ordinary Differential Equation Models With Application to Predator-Prey Dynamical Models," *Biometrics*, 71, 131–138. [1712,1718]

Zhao, P., and Yu, B. (2006), "On Model Selection Consistency of Lasso," *Journal of Machine Learning Research*, 7, 2541–2563. [1718]

Zhu, H., Yao, F., and Zhang, H. H. (2014), "Structured Functional Additive Regression in Reproducing Kernel Hilbert Spaces," *Journal of the Royal Statistical Society*, Series B, 76, 581–603. [1712,1717]