Journal of Machine Learning Research 25 (2024) 1-51 Submitted 5/22; Revised 1/24; Published 2/24

Post-Regularization Confidence Bands for Ordinary Differential Equations

Xiaowu Dai DAI@STAT.UCLA.EDU
Department of Statistics and Data Science and Department of Biostatistics
University of California, Los Angeles, CA 90095-1554, USA

Lexin Li LEXINLI@BERKELEY.EDU
Department of Biostatistics and Epidemiology
University of California, Berkeley, CA 94720-1776, USA

Editor: Mladen Kolar

Abstract

Ordinary differential equation (ODE) is an important tool to study a system of biological and physical processes. A central question in ODE modeling is to infer the significance of individual regulatory effect of one signal variable on another. However, building confidence band for ODE with unknown regulatory relations is challenging, and it remains largely an open question. In this article, we construct the post-regularization confidence band for the individual regulatory function in ODE with unknown functionals and noisy data observations. Our proposal is the first of its kind, and is built on two novel ingredients. The first is a new localized kernel learning approach that combines reproducing kernel learning with local Taylor approximation, and the second is a new de-biasing method that tackles infinite-dimensional functionals and additional measurement errors. We show that the constructed confidence band has the desired asymptotic coverage probability, and the recovered regulatory network approaches the truth with probability tending to one. We establish the theoretical properties when the number of variables in the system can be either smaller or larger than the number of sampling time points, and we study the regime-switching phenomenon. We demonstrate the efficacy of the proposed method through both simulations and illustrations with two data applications.

Keywords: De-biasing; Local polynomial approximation; Ordinary differential equations; Reproducing kernel Hilbert space; Smoothing spline analysis of variance; Time series.

1. Introduction

Characterizing the dynamics of biological and physical processes is of fundamental interest in a large variety of scientific fields, and ordinary differential equation (ODE) is a frequently used tool to address such type of questions. Examples include infectious disease (Liang and Wu, 2008), genomics (Cao and Zhao, 2008; Ma et al., 2009; Wu et al., 2014), neuroscience (Izhikevich, 2007; Zhang et al., 2015a, 2017; Cao et al., 2019), economics (Dai and He, 2023), among many others. An ODE system models the changes of a set of variables, quantified by their derivatives with respect to time, as functions of all other variables in the system. Typically, the system is observed on discrete time points with some additive measurement errors. In recent years, there have been an increased number of proposals for ODE modeling. One family of ODE models adopt linear function forms in the ODE system;

for instance, Lu et al. (2011) proposed a set of linear ODEs, Zhang et al. (2015a) extended to include two-way interactions, and Dattner and Klaassen (2015) further extended to a generalized linear form using a finite set of known basis functions. Another family of ODE models consider additive functionals; for instance, Henderson and Michailidis (2014); Wu et al. (2014); Chen et al. (2017) proposed the generalized additive models with a set of common basis functions plus an unknown residual function. The third family studies the ODE system when the functional forms are completely known (González et al., 2014; Li et al., 2015; Zhang et al., 2015b; Mikkelsen and Hansen, 2017). Recently, Dai and Li (2022) proposed reproducing kernel-based ODEs to flexibly model the unknown functionals of both main effects and two-way interactions.

A central question in ODE modeling is about *inference* of significance of individual regulatory relations among the variables in the system (Ma et al., 2009). The majority of existing ODE solutions, however, have been focusing on regularized sparse *estimation* of the ODEs. Even though sparse estimation can in effect identify such relations, it does not produce a quantification of statistical significance, nor explicitly controls the false discovery rate (FDR). By contrast, inference provides both an explicit uncertainty quantification and an explicit FDR control. As such, inference is generally a different problem from sparse estimation. When the functional forms are completely *known* in an ODE system, confidence intervals for the ODE parameters have been studied and been mostly built upon asymptotic normality of finite-dimensional parameters (Ramsay et al., 2007; Qi and Zhao, 2010; Xue et al., 2010; Miao et al., 2014; Zhang et al., 2015b; Wu et al., 2019). When the functional forms are *unknown*, however, there is *no* existing solution for individual ODE parameter inference. Dai and Li (2022) derived the confidence interval of the entire signal trajectory in kernel ODEs with unknown functionals. Nevertheless, their method could *not* infer the *individual* regulatory effect of one variable on another, but instead only the *sum* of all individual effects. Building confidence intervals for individual ODE parameters with unknown regulatory relations is particularly challenging, as it involves infinite-dimensional functionals. There is a clear gap in the current literature on ODE inference.

In this article, we tackle the ODE inference problem with unknown functionals and noisy data observations. Our goal is to directly construct the confidence band for any individual regulatory functional that measures the effect of one signal variable on another. The constructed confidence band provides both an uncertainty quantification for the individual regulatory relation, and also a sparse recovery of the entire regulatory system when coupled with a proper false discovery rate (FDR) control. We establish the confidence band for both low-dimensional and high-dimensional settings, where the number of variables in the system can be smaller or larger than the number of sampling time points, and we study the regime-switching phenomenon. We show that the constructed confidence band has the desired asymptotic coverage probability, and the recovered regulatory network approaches the truth with probability tending to one. Toward our goal, we propose and develop two novel methodological ingredients: a new *localized kernel learning* approach that combines reproducing kernel learning with local Taylor approximation, and a new de-biasing method that tackles infinite-dimensional functionals and additional measurement errors. Consequently, our proposal makes useful contributions on multiple fronts, including ODE inference, nonparametric modeling, as well as high-dimensional inference with measurement errors.

The first component is a new localized kernel learning approach that in effect fuses two widely used nonparametric modeling techniques, reproducing kernel learning methods (Wahba, 1990), and local polynomial methods (Fan and Gijbels, 1996). More specifically, we adopt the kernel ODE model of Dai and Li (2022), which is built upon the learning framework of reproducing kernel Hilbert space (RKHS, Aronszajn, 1950; Wahba, 1990) and smoothing spline analysis of variance (SS-ANOVA, Wahba et al., 1995; Huang, 1998; Lin and Zhang, 2006). This model allows highly flexible and unknown forms for the functions in the ODE system as well as interactions. Meanwhile, we employ the Taylor expansion and local approximation idea, which is frequently employed in local polynomial nonparametric regressions (Fan and Gijbels, 1996; Opsomer and Ruppert, 1997). Such a fused method essentially characterizes the regulatory effect of one variable on another in the ODE system through a scalar quantity, which in turn allows us to derive the corresponding confidence band. Moreover, this localized kernel learning approach is potentially useful for other nonparametric modeling problems beyond ODEs.

We remark that our method is related to but also substantially different from the kernel ODE method of Dai and Li (2022), and the kernel-sieve hybrid method of Lu et al. (2020). Compared to Dai and Li (2022), although we adopt the same ODE model framework, our estimator is utterly different after introducing the local approximation. Actually, our localized kernel estimator has a slower convergence rate compared to the minimax rate of the estimator of Dai and Li (2022) in a low-dimensional setting; see Section 5.1. On the other hand, this slower rate is sufficient for constructing an asymptotically valid confidence band. More importantly, the method of Dai and Li (2022) can only obtain the confidence interval for the entire signal trajectory, which is the *sum* of *all* individual effects, but not for a *single* individual effect of one variable on another. Directly applying the estimator of Dai and Li (2022) *cannot* obtain the individual confidence band we target. Later, we also numerically demonstrate that the proposed method outperforms a naive modification of Dai and Li (2022) that aggregates the point-wise confidence intervals at grid points with Bonferroni correction. Compared to Lu et al. (2020) who tackled the inference problem of a nonparametric additive model, our ODE inference method differs in multiple ways, including that the signals are not directly observed but need to be estimated from the data with error, there are pairwise interaction terms, and the ODE estimation involves integrals. These differences have introduced a whole new set of challenges than Lu et al. (2020) and the classical sieve method of Shen and Wong (1994), and thus a different solution.

We also remark that, by choosing a proper linear kernel or additive kernel, the kernel ODE model includes the linear ODE (Zhang et al., 2015a) and the additive ODE (Chen et al., 2017) as special cases. Consequently, our inference solution is applicable to a range of different ODE models. Our work thus addresses a scientific question that is crucial but currently still remains open, and makes a useful addition to the ODE toolbox.

The second component is a new de-biasing method for the localized kernel ODE estimator. We observe that the individual regulatory effects in ODE models are nonparametric functionals, which are estimated through proper regularization in terms of the RKHS norms. However, the regularization introduces bias, and essentially offers a trade-off between bias and overfitting. Consequently, it is crucial to perform de-biasing in post-regularization high-dimensional statistical inference (Zhang and Zhang, 2014; van de Geer et al., 2014; Ning and Liu, 2017; Zhang and Cheng, 2017). On the other hand, there are some extra layers of

complications for de-biasing in ODEs. First of all, the signal variables themselves are not directly observed, but only their noisy counterparts, and they need to be estimated given the noisy data. Besides, the objects of inference are infinite-dimensional functionals, and their estimation involves integrals. To overcome those difficulties, we introduce a new bias correction score with integral of the estimated functional. We also generalize Chernozhukov et al. (2014) and perform a new approximation analysis for the Gaussian multiplier bootstrap within the RKHS framework. Our method is the first de-biasing solution for ODE models, and thus also contributes to the literature on de-biasing. In addition, it is potentially useful for high-dimensional inference of other statistical models that involve latent variables and measurement errors (Wansbeek and Meijer, 2001).

The rest of the article is organized as follows. Section 2 introduces the kernel ODE model, then develops the localized kernel learning approach. Section 3 presents the parameter estimation, and Section 4 derives the confidence band formula. Section 5 establishes the convergence rate and coverage property of the proposed method. Section 6 investigates the finite-sample performance, and Section 7 illustrates with two real data examples. Section 8 concludes the paper with a discussion, and the Appendix collects all technical proofs.

2. Localized Kernel Learning for ODE

In this section, we first present our kernel ODE model system, which consists of models (1), (2) and (3). We then propose the localized kernel learning method, which fuses reproducing kernel learning and local polynomial approximation, and is crucial for ODE inference.

2.1 Kernel ODE Model

Let $x(t) = (x_1(t), \ldots, x_p(t))^\top \in \mathbb{R}^p$ denote the set of p variables of interest, and t index the time in a standardized interval $\mathcal{T} = [0, 1]$. We consider the ODE system,

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

where $F = \{F_1, \ldots, F_p\}$ denotes the set of unknown functionals that characterize the regulatory relations among $x(t)$. Typically, the system (1) is observed on a set of n discrete time points $\{t_1, \ldots, t_n\}$, with additional measurement errors,

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

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

In a biological and physical system, given the observed noisy time-course data $\{y_i\}_{i=1}^n$, a central question of interest is to uncover the structure of the system of ODEs in terms of which variables regulate which other variables. We say that $x_k(t)$ regulates $x_j(t)$, if F_j is

a non-zero functional of $x_k(t)$. That is, $x_k(t)$ affects $x_j(t)$ through the functional F_j on its derivative $dx_j(t)/dt$, for $j, k = 1, \ldots, p$. We consider the following model for F_j,

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

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

We next build model (3) within the smoothing spline ANOVA framework (Wahba et al., 1995; Gu, 2013; Lin and Zhang, 2006). Specifically, let $\mathcal{H}_k$ denote a space of functions of $x_k(t)$ with zero marginal integral, which is specified through the averaging operator, $\int_{\mathcal{T}} F_{jk}(x_k(t))dt = 0$ for any $k, j = 1, \ldots, p$. The zero marginal integrals of functions in $\mathcal{H}_k$ ensure that the decomposition in (3) is well defined over its domain, and that the terms θ_{j0}, F_{jk}, and F_{jkl} in (3) are identifiable and can be estimated uniquely. More discussion of the averaging operator is given in Section C.1 of the Appendix. Let $x_k(t) \in \mathcal{X}$, where $\mathcal{X}$ is a compact set in $\mathbb{R}$. Let $\{1\}$ denote the space of constant functions. Construct the tensor product space as

$$\mathcal{H} = \{1\} \oplus \sum_{k=1}^{p} \mathcal{H}_k \oplus \sum_{k=1, k\neq l}^{p} \sum_{l=1}^{p} (\mathcal{H}_k \otimes \mathcal{H}_l), \tag{4}$$

where $\oplus$ and $\otimes$ denote the direct sum operator and the tensor product operator, respectively. The space $\mathcal{H}_k \otimes \mathcal{H}_l$ includes the tensor products of the functions in $\mathcal{H}_k$ and $\mathcal{H}_l$. We assume the functionals F_j, $j = 1, \ldots, p$, in the ODE model (3) are located in the space of $\mathcal{H}$.

We note that our kernel ODE model is the same as the model used in Dai and Li (2022). Nevertheless, one *cannot* achieve the goal of inferring individual regulatory functional using the approach of Dai and Li (2022), which estimates the collective functional F_j, rather than the individual functional F_{jk}. Instead, we propose a completely new localized kernel learning approach for inference, which is a key novelty of this article.

2.2 Localized Kernel Learning

We next introduce the Taylor expansion and local approximation idea into our kernel ODE model framework. In effect, we fuse two popular nonparametric modeling techniques, reproducing kernel learning (Wahba, 1990) and local polynomial learning (Fan and Gijbels, 1996). Our primary goal is to infer the individual regulatory effect of $x_k(t)$ on $x_j(t)$, for any given pair of $j, k = 1, \ldots, p$, in the ODE system (1). Toward that goal, we observe that such a regulatory effect is encoded in two parts: the main effect term, $F_{jk}(x_k(t))$, $j, k = 1, \ldots, p$, and the two-way interaction terms, $F_{jkl}(x_k(t), x_l(t))$, $l = 1, \ldots, p, l \neq k$. We next study the two parts separately.

First, for the main effect term $F_{jk}(x_k(t))$, we consider the Taylor expansion with Lagrange remainder at a fixed time point $t = t_0$ (Section 7.7, Apostol, 1967). Specifically, letting $\tilde{t}$ be a point locating between t_0 and t, by the chain rule, we have

$$F_{jk}(x_k(t)) = F_{jk}(x_k(t_0)) + \frac{dF_{jk}(x_k(\tilde{t}))}{dx_k} \frac{dx_k(\tilde{t})}{dt}(t - t_0). \tag{5}$$

Since $F_{jk}(x_k(t_0))$ is a constant at the fixed t_0 and does not vary with t, we denote $F_{jk}(x_k(t_0)) \equiv \alpha_{jk,t_0}$. Besides, let $\mathcal{F}$ be the function space where $x_k(t)$ reside in, and $\mathcal{F}$ does not have to be the same as $\mathcal{H}_k$. Suppose the functions in $\mathcal{H}_k$ and $\mathcal{F}$ have continuous first derivatives, which is true for most of commonly used RKHS, including those generated by the Gaussian, Laplace, or Matérn kernel (Scholkopf and Smola, 2018). As such, both $dF_{jk}(x_k(t))/dx_k$ and $dx_k(t)/dt$ are continuous functions in t, and when $t \to t_0$, the second term in (5) goes to zero. Henceforth, we can approximate $F_{jk}(x_k(t))$ by α_{jk,t_0}. We remark that, we consider the first-order Taylor expansion, instead of the zeroth-order, in (5). Nevertheless, due to the smoothness property of RKHS, we approximate $F_{jk}(x_k(t))$ by α_{jk,t_0}, which is a constant at a fixed time point t_0 while it changes as t_0 varies. Moreover, when inferring the effect of $x_k(t)$ on $x_j(t)$, we focus on the main effect term of interest $F_{jk}(x_k(t))$, while treating the rest of the main effect terms $F_{jl}(x_l(t))$, $l = 1, \ldots, p, l \neq k$, as nuisance parameters.

Next, for the interaction terms $F_{jkl}(x_k(t), x_l(t))$, $l = 1, \ldots, p, l \neq k$, since $\{1\} \otimes \mathcal{H}_l = \mathcal{H}_l$ (Wahba, 1990), $F_{jkl}(x_k(t), x_l(t))$ is absorbed into the main effect term $F_{jl}(x_l(t))$, when $F_{jk}(x_k(t))$ is approximated by the constant α_{jk,t_0}. In other words, when the effect of $x_k(t)$ on $x_j(t)$ holds constant, the change of the joint effect of $(x_k(t), x_l(t))$ on $x_j(t)$ only depends on the change of the effect of $x_l(t)$ on $x_j(t)$. Another way to see this is that, since $F_{jkl} \in \mathcal{H}_k \otimes \mathcal{H}_l$ in (4), there exists a finite integer s and functions $F_{jk_\nu} \in \mathcal{H}_k, F_{jl_\nu} \in \mathcal{H}_l$ for $\nu = 1, \ldots, s$, such that

$$F_{jkl} = \sum_{\nu=1}^{s} F_{jk_\nu} F_{jl_\nu} = \sum_{\nu=1}^{s} \alpha_{jk_\nu,t_0} F_{jl_\nu} \in \mathcal{H}_l, \tag{6}$$

where the second equality is due to (5) with $F_{jk_\nu}(x_k(t)) = \alpha_{jk_\nu,t_0}$ as $t \to t_0$, and the last step is due to the fact that the linear combination of RKHS functions is still in the RKHS (Wahba, 1990).

Combining the above results, and following the Taylor approximation that $F_{jk}(x_k(t)) = \alpha_{jk,t_0}$ as $t \to t_0$, the regulatory effect of $x_k(t)$ on $x_j(t)$ is now captured by the scalar α_{jk,t_0}. Consequently, we estimate α_{jk,t_0} at any given time point $t_0 \in \mathcal{T}$, along with other nuisance terms in F_j, then build a confidence band based on α_{jk,t_0} to infer the effect of $x_k(t)$ on $x_j(t)$. Specifically, write $F_j = \theta_{j0} + \tilde{F}_j$, where θ_{j0} is the global mean, and $\tilde{F}_j$ is the centralized functional, $j = 1, \ldots, p$. For any given $k = 1, \ldots, p$, if t is within a local neighborhood of t_0 in that $|t - t_0| < \varepsilon$ for some small $\varepsilon > 0$, we write

$$\begin{aligned} \tilde{F}_j(x(t)) &= \alpha_{jk,t_0} + \sum_{k'=1,k'\neq k}^{p} F_{jk'}(x_{k'}(t)) + \sum_{k'=1,k'\neq k,l}^{p} \sum_{l=1,l\neq k}^{p} F_{jk'l}(x_{k'}(t), x_l(t)) \\ &\equiv \alpha_{jk,t_0} + H_{jk}(x(t)), \end{aligned} \tag{7}$$

where H_{jk} collects all the nuisance terms when evaluating the effect of $x_k(t)$ on $x_j(t)$. We next develop a procedure to estimate the global mean θ_{j0} and the functional $\tilde{F}_j$ in (7).

3. ODE Estimation

In this section, we develop an estimation procedure, where we adopt a two-step collocation approach (Varah, 1982). We first estimate model (3) under the constraint that $F_j \in \mathcal{H}$ in

(4), then incorporate the local approximation in (7). Next, we derive the corresponding optimization algorithm to estimate the unknown parameters.

3.1 Two-Step Collocation Estimation

We adopt the two-step collocation method that is commonly used in ODE estimation.

The first step is to obtain a smoothing estimate $\widehat{x}(t) = (\widehat{x}_1(t), \ldots, \widehat{x}_p(t))^\top$,

$$\widehat{x}_j(t) = \underset{z_j \in \mathcal{F}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} [y_{ij} - z_j(t_i)]^2 + \lambda_{nj} \|z_j(t)\|_{\mathcal{F}}^2 \right\}, \quad j = 1, \ldots, p, \tag{8}$$

where $\|\cdot\|_{\mathcal{F}}$ is the norm of the RKHS $\mathcal{F}$, z_j is a function in $\mathcal{F}$ we minimize over, and $\lambda_{nj} \geq 0$ is the smoothness parameter often tuned by generalized cross-validation (Wahba, 1990).

The second step is to estimate F_j in (3). We first follow Dai and Li (2022) and obtain an estimator of the global mean θ_{j0} in F_j as,

$$\widehat{\theta}_{j0} = \bar{y}_j - \int_{\mathcal{T}} \bar{T}(t) \tilde{F}_j(\widehat{x}(t)) dt, \tag{9}$$

where $\bar{y}_j = n^{-1} \sum_{i=1}^{n} y_{ij}$, $\bar{T}(t) = n^{-1} \sum_{i=1}^{n} T_i(t)$, $T_i(t) = \mathbb{I}\{0 \leq t \leq t_i\}$, and $\mathbb{I}(\cdot)$ is the indicator function. We then plug in this global mean estimator and estimate the centralized component $\tilde{F}_j$ in F_j by solving the penalized optimization,

$$\begin{aligned} \min_{\tilde{F}_j \in \mathcal{H}} \Bigg\{ & \frac{1}{n} \sum_{i=1}^{n} \left[(y_{ij} - \bar{y}_j) - \int_{\mathcal{T}} \left(T_i(t) - \bar{T}(t) \right) \tilde{F}_j(\widehat{x}(t)) dt \right]^2 \\ & + \tau_{nj} \left(\sum_{k=1}^{p} \|\mathcal{P}^k \tilde{F}_j\|_{\mathcal{H}} + \sum_{k=1, k \neq l}^{p} \sum_{l=1}^{p} \|\mathcal{P}^{kl} \tilde{F}_j\|_{\mathcal{H}} \right) \Bigg\}, \end{aligned} \tag{10}$$

where $\|\cdot\|_{\mathcal{H}}$ is the norm of $\mathcal{H}$, $\mathcal{P}^k \tilde{F}_j$ and $\mathcal{P}^{kl} \tilde{F}_j$ are the orthogonal projections of $\tilde{F}_j$, or equivalently F_j, onto $\mathcal{H}_k$ and $\mathcal{H}_k \otimes \mathcal{H}_l$, respectively, and τ_{nj} is the penalty parameter. Note that the optimization problem (10) deals with the integral $\int_0^{t_i} \tilde{F}_j(\widehat{x}(u)) du$, rather than the derivative $d\widehat{x}_j(t)/dt$. This follows a similar spirit as Dattner and Klaassen (2015), and is to produce a more robust estimate. Moreover, the penalty function in (10) is a sum of RKHS norms on the main effects and pairwise interactions, which is similar in spirit as the component selection and smoothing operator penalty of Lin and Zhang (2006).

Next, we introduce the localization as specified in (7). Let $R : \mathcal{T} \mapsto \mathbb{R}$ be a symmetric density function with bounded support. Denote $R_h(\cdot) = h^{-1} R(\cdot/h)$, where $h > 0$ is the bandwidth. Following (7) and (10), we estimate $\alpha_{jk,t_0} \in \mathbb{R}$ and $H_{jk} \in \mathcal{H}$ through the localized and penalized optimization,

$$\begin{aligned} \min_{\alpha_{jk,t_0}, H_{jk}} \Bigg\{ & \frac{1}{n} \sum_{i=1}^{n} R_h(t_i - t_0) \left[(y_{ij} - \bar{y}_j) - \alpha_{jk,t_0} \bar{t}_i - \int_{\mathcal{T}} \left(T_i(t) - \bar{T}(t) \right) H_{jk}(\widehat{x}(t)) dt \right]^2 \\ & + \tau_{nj} \left(\sum_{k'=1, k' \neq k}^{p} \|\mathcal{P}^{k'} H_{jk}\|_{\mathcal{H}} + \sum_{k'=1, k' \neq k, l}^{p} \sum_{l=1, l \neq k}^{p} \|\mathcal{P}^{k'l} H_{jk}\|_{\mathcal{H}} \right) \Bigg\}, \end{aligned} \tag{11}$$

where $\bar{t}_i = t_i - n^{-1}\sum_{i=1}^n t_i$. The estimate $\widehat{\alpha}_{jk,t_0}$ obtained from (11) captures the individual regulatory effect of $x_k(t)$ on $x_j(t)$ in a local neighborhood of $t_0 \in \mathcal{T}$. The local weight R_h introduced in (11) is to facilitate both the estimation and subsequent inference. It places more weight to the data observations close to t_0 and less weight to those far away from t_0, in the same spirit as the local polynomial method (Fan and Gijbels, 1996). Besides, it allows us to later construct confidence bands using tools such as the extreme value theory as well as the Gaussian multiplier bootstrap procedure.

Now, given the estimates $\widehat{x}_j(t)$ from (8), $\widehat{\theta}_{j0}$ from (9), and $\widehat{\alpha}_{jk,t_0}, \widehat{H}_{jk}$ from (11), we estimate the p-dimensional functional $F_j(x(t_0)) = \theta_{j0} + F_{jk}(x_k(t_0)) + \sum_{k'=1,k'\neq k}^p F_{jk'}(x_k'(t_0)) + \sum_{k'=1,k'\neq l}^p \sum_{l=1}^p F_{jk'l}(x_k'(t_0), x_l(t_0))$ at any time point $t_0 \in \mathcal{T}$ by

$$\widehat{F}_j(\widehat{x}(t_0)) = \widehat{\theta}_{j0} + \widehat{\alpha}_{jk,t_0} + \widehat{H}_{jk}(\widehat{x}(t_0)), \quad j = 1, \ldots, p. \tag{12}$$

We comment that, due to the Taylor approximation and localization, our localized kernel ODE estimator in (12) is different from the kernel ODE estimator of Dai and Li (2022) obtained from (10). In Section 5.1, we study its convergence rate, and compare it to the minimax optimal rate of the kernel ODE estimator of Dai and Li (2022).

3.2 Optimization Algorithm

We next develop an optimization algorithm to solve (11). Toward that end, we first propose an optimization problem that is equivalent to (11) but is computationally easier to tackle. We then develop an iterative algorithm to solve this equivalent optimization problem. We also remark that, the new algorithm differs from that of Dai and Li (2022), in that a local weight R_h is introduced, and the estimations of the parameter of interest α_{jk,t_0} and the rest of nuisance parameters are separated.

Specifically, we consider the following optimization problem that is equivalent to (11),

$$\begin{aligned}\min_{\theta_j, \alpha_{jk,t_0}, H_{jk}} \Bigg\{ & \frac{1}{n}\sum_{i=1}^n R_h(t_i - t_0)\left[(y_{ij} - \bar{y}_j) - \alpha_{jk,t_0}\bar{t}_i - \int_{\mathcal{T}} \left(T_i(t) - \bar{T}(t)\right) H_{jk}(\widehat{x}(t))dt\right]^2 \\ & + \eta_{nj}\left(\sum_{k'=1,k'\neq k}^p \theta_{jk'}^{-1}\|\mathcal{P}^{k'}H_{jk}\|_{\mathcal{H}}^2 + \sum_{k'=1,k'\neq k}^p \sum_{l=1,l\neq k',k}^p \theta_{jk'l}^{-1}\|\mathcal{P}^{k'l}H_{jk}\|_{\mathcal{H}}^2\right) \\ & + \kappa_{nj}\left(\sum_{k'=1,k'\neq k}^p \theta_{jk'} + \sum_{k'=1,k'\neq k}^p \sum_{l=1,l\neq k',k}^p \theta_{jk'l}\right)\Bigg\}, \end{aligned} \tag{13}$$

subject to $\theta_j \geq 0$, where $\theta_j = \{\theta_{jk'}\}_{k'\neq k} \cup \{\theta_{jk'l}\}_{k'\neq k; l\neq k',k} \in \mathbb{R}^{(p-1)^2}$ collects all the parameters to estimate, and $\eta_{nj}, \kappa_{nj} \geq 0$ are the tuning parameter, $j = 1, \ldots, p$. The optimization problem (13) utilizes the parameter $\theta_{jk'}$ to control the sparsity of each component $\mathcal{P}^{k'}H_{jk}$, and $\theta_{jk'l}$ to control the sparsity of $\mathcal{P}^{k'l}H_{jk}$, $k' \neq k$ and $l \neq k', k$. The shrinkage of θ_j gives rise to zero function components in the final estimate. The two optimizations (11) and (13) are equivalent in the sense that, if $(\widehat{\alpha}_{jk,t_0}, \widehat{H}_{jk})$ minimizes (11), then $(\widehat{\theta}_j, \widehat{\alpha}_{jk,t_0}, \widehat{H}_{jk})$ minimizes (13), with $\widehat{\theta}_{jk'} = \eta_{nj}^{1/2}\kappa_{nj}^{-1/2}\|\mathcal{P}^{k'}\widehat{H}_{jk}\|_{\mathcal{H}}$, and $\widehat{\theta}_{jk'l} = \eta_{nj}^{1/2}\kappa_{nj}^{-1/2}\|\mathcal{P}^{k'l}\widehat{H}_{jk}\|_{\mathcal{H}}$, $k', l =$

$1, \ldots, p, k', l \neq k, k' \neq l$. Meanwhile, if $(\widehat{\theta}_j, \widehat{\alpha}_{jk,t_0}, \widehat{H}_{jk})$ minimizes (13), then $(\widehat{\alpha}_{jk,t_0}, \widehat{H}_{jk})$ minimizes (11). The reason of introducing θ_j is to benefit the computation. The optimization in (11) is challenging. By contrast, we develop an iterative procedure to solve the equivalent optimization problem (13), where each iteration either has a closed-form solution, or becomes a standard Lasso regression. As such, the computation is greatly simplified. That is, we employ the representer theorem to obtain a closed-form estimate for $(\alpha_{jk,t_0}, H_{jk})$ given θ_j. We then employ the Lasso method to obtain a sparse estimate of θ_j given $(\alpha_{jk,t_0}, H_{jk})$.

Specifically, for a given estimate $\widehat{\theta}_j$, the optimization problem (13) becomes,

$$\min_{\alpha_{jk,t_0}, H_{jk}} \left\{ \frac{1}{n} \sum_{i=1}^{n} R_h\left(t_i - t_0\right) \left[(y_{ij} - \bar{y}_j) - \alpha_{jk,t_0} \bar{t}_i - \int_{\mathcal{T}} \left(T_i(t) - \bar{T}(t)\right) H_{jk}(\widehat{x}(t)) dt \right]^2 \right. \\ \left. + \eta_{nj} \left(\sum_{k'=1, k' \neq k}^{p} \widehat{\theta}_{jk'}^{-1} \|\mathcal{P}^{k'} H_{jk}\|_{\mathcal{H}}^2 + \sum_{k'=1, k' \neq k}^{p} \sum_{l=1, l \neq k', k}^{p} \widehat{\theta}_{jk'l}^{-1} \|\mathcal{P}^{k'l} H_{jk}\|_{\mathcal{H}}^2 \right) \right\}. \tag{14}$$

Let $K'_k(\cdot, \cdot) : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ denote the kernel generating the RKHS $\mathcal{H}'_k$, $k' \neq k$. Then $K_{k'l} = K'_k K_l$ is the reproducing kernel of the RKHS $\mathcal{H}'_k \otimes \mathcal{H}_l$ (Aronszajn, 1950). Let $K_{\theta_j} = \sum_{k'=1, k' \neq k}^{p} \widehat{\theta}_{jk'} K'_k + \sum_{k'=1, k' \neq k}^{p} \sum_{l=1, l \neq k', k}^{p} \widehat{\theta}_{jk'l} K_{k'l}$; see also Wahba et al. (1995, Section 2) for a discussion on the weighted kernel. Employing the representer theorem of Wahba (1990, Theorem 1.3.1), the solution $\widehat{H}_{jk}$ to (14) is of the form,

$$\widehat{H}_{jk}(\widehat{x}(t)) = \sum_{i=1}^{n} c_{ij} \int_{\mathcal{T}} K_{\theta_j}\left(\widehat{x}(t), \widehat{x}(s)\right) \left(T_i(s) - \bar{T}(s)\right) ds \tag{15}$$

for some $c_j = (c_{1j}, \ldots, c_{nj})^\top \in \mathbb{R}^n$. Define two $n \times n$ matrices,

$$\Sigma = (\Sigma_{ii'}) \in \mathbb{R}^{n \times n}, \Sigma_{ii'} = \int_{\mathcal{T}} \int_{\mathcal{T}} \{T_i(s) - \bar{T}(s)\} K_{\theta_j}(\widehat{x}(t), \widehat{x}(s)) \{T_{i'}(t) - \bar{T}(t)\} ds dt, \\ R_{t_0} = \mathrm{diag}\{R_h(t_1 - t_0), \ldots, R_h(t_n - t_0)\} \in \mathbb{R}^{n \times n}. \tag{16}$$

Write $y_j = (y_{1j}, \ldots, y_{nj})^\top \in \mathbb{R}^n$. Plugging (15) into (14), we reach the following weighted quadratic minimization problem in terms of $\{\alpha_{jk,t_0}, c_j\}$,

$$\min_{\alpha_{jk,t_0}, c_j} \left\{ \frac{1}{n} [(y_j - \bar{y}_j) - \alpha_{jk,t_0} \bar{t} - \Sigma c_j]^\top R_{t_0} [(y_j - \bar{y}_j) - \alpha_j \bar{t} - \Sigma c_j] + \eta_{nj} c_j^\top \Sigma c_j \right\}, \tag{17}$$

where $\bar{t} = (\bar{t}_1, \ldots, \bar{t}_n)^\top \in \mathbb{R}^n$. The optimization problem (17) has a closed-form solution; see also Dai and Li (2022). We tune the parameter $\eta_{nj} \geq 0$ using generalized cross-validation, following Wahba (1990).

Next, for a given estimate $(\widehat{\alpha}_{jk,t_0}, \widehat{H}_{jk})$, the optimization problem (13) becomes,

$$\min_{\theta_j} \left\{ \frac{1}{n} (z_j - G\theta_j)^\top R_{t_0} (z_j - G\theta_j) + \kappa_{nj} \left(\sum_{k'=1, k' \neq k}^{p} \theta_{jk'} + \sum_{k'=1, k' \neq k}^{p} \sum_{l=1, l \neq k', k}^{p} \theta_{jk'l} \right) \right\}, \tag{18}$$

Algorithm 1 Estimation and inference procedure for a given pair $(j,k) \in \{1,\ldots,p\}$.

1: Initialization with the values for $\theta_{jk'} = \theta_{jk'l} = 1$, for $k', l = 1,\ldots,p, k' \neq k, l \neq k', k$.
2: Obtain the smoothing spline estimate $\widehat{x}_j(t)$ from (8).
3: **repeat**
4: Solve $(\widehat{\alpha}_{jk,t_0}, \widehat{H}_{jk})$ in (14) given $\widehat{\theta}_j$ through (15) and (17).
5: Solve $\widehat{\theta}_j$ in (18) given $(\widehat{\alpha}_{jk,t_0}, \widehat{H}_{jk})$ through the Lasso penalized regression (18).
6: **until** the stopping criterion is met.
7: Construct the confidence band by Gaussian multiplier bootstrap from (20).

subject to $\theta_{jk'} \geq 0, \theta_{jk'l} \geq 0$, where the "response" is $z_j = (y_j - \bar{y}_j) - \widehat{\alpha}_{jk,t_0}\bar{t} - (1/2)n\eta_{nj}c_j$, the "predictor" is $G \in \mathbb{R}^{n\times(p-1)^2}$, whose first $(p-1)$ columns are $\Sigma^{k'}c_j$ with $k' = 1,\ldots,k-1,k+1,\ldots,p$, and the last $(p-1)(p-2)$ columns are $\Sigma^{lr}c_j$ with $k', l = 1,\ldots,k-1,k+1,\ldots,p, k' \neq l$, and $\Sigma^{k'} = (\Sigma^{k'}_{ii'}), \Sigma^{k'l} = (\Sigma^{k'l}_{ii'})$ are both $n \times n$ matrices whose (i,i')th entries are $\Sigma^{k'}_{ii'} = \int_{\mathcal{T}}\int_{\mathcal{T}}\{T_i(s) - \bar{T}(s)\}K_{k'}(\widehat{x}(t),\widehat{x}(s))\{T_{i'}(t) - \bar{T}(t)\}dsdt$, and $\Sigma^{k'l}_{ii'} = \int_{\mathcal{T}}\int_{\mathcal{T}}\{T_i(s) - \bar{T}(s)\}K_{k'l}(\widehat{x}(t),\widehat{x}(s))\{T_{i'}(t) - \bar{T}(t)\}dsdt$, respectively, where $i,i' = 1,\ldots,n, j = 1,\ldots,p$. We employ the standard Lasso for (18) in our implementation, and tune the parameter κ_{nj} using tenfold cross-validation, following the usual Lasso literature.

We repeat the above optimization steps iteratively until some stopping criterion is met, e.g., when the estimates in two consecutive iterations are close enough, or when the number of iterations reaches some maximum number. We summarize the above iterative procedure, along with the confidence band derived in the next section, in Algorithm 1.

We remark on the computational complexity of the proposed estimation method. Specifically, the computational cost is $O(n^3 + np^4)$, where $O(n^3)$ is due to the reproducing kernel-type regression in (14), and $O(np^4)$ is due to the Lasso-type regression in (18) with $O(p^2)$ parameters. We note that this computational complexity is comparable to the ones of existing ODE estimation methods such as Dai and Li (2022) and Chen et al. (2017). Moreover, the computational complexity of the inference method we develop in (20) is $O(n^2)$, which is substantially smaller than that for the estimation method when $n \to \infty$.

4. ODE Inference

In this section, we construct the confidence band for the regulatory effect $F_{jk}(x_k(t))$ of $x_k(t)$ on $x_j(t)$ for any given pair $(j,k) \in \{1,\ldots,p\}$. Toward that end, we first construct a de-biased estimator for $\widehat{\alpha}_{jk,t}$, since $\widehat{\alpha}_{jk,t}$ is obtained through the regularization in (11), which has an ℓ_1-type penalty with the sum of RKHS norms and inevitably introduces bias (Zhang and Zhang, 2014; Javanmard and Montanari, 2014; van de Geer et al., 2014). Then based on our de-biased estimator, we employ the Gaussian multiplier process to construct a valid confidence band. We also discuss hypothesis testing for a single pair of variables, then multiple testing for all pairs of variables to reconstruct the entire regulatory system. Finally, we briefly discuss the extension from a single experiment to multiple experiments.

4.1 Confidence Band

As the first step, we propose the following de-biased estimator, given that the functional F_{jk} is approximated by $\widehat{\alpha}_{jk,t_0}$ in our localized kernel learning, to reduce the bias in the estimate $\widehat{\alpha}_{jk,t_0}$,

$$\widehat{F}_{jk}(x_k(t_0)) = \widehat{\alpha}_{jk,t_0} + \frac{1}{n}\sum_{i=1}^{n} \Sigma_{i\cdot} R_{t_0} \left[(y_j - \bar{y}_j) - \int_{\mathcal{T}} \left(T_i(t) - \bar{T}(t) \right) \widehat{H}_{jk}(\widehat{x}(t)) dt \right], \qquad (19)$$

where $\Sigma_{i\cdot}, i = 1, \ldots, n$, is the kth row of the $n \times n$ matrix Σ defined in (16), and $\widehat{H}_{jk}(\widehat{x}(t))$ is obtained from (15). We make a few remarks. First of all, we employ the integral of the infinite-dimensional functional $\widehat{H}_{jk}(\widehat{x}(t))$ in (19) to correct the bias in $\widehat{\alpha}_{jk,t_0}$. As a result, the inference of the de-biased estimator $\widehat{F}_{jk}(x_k(t_0))$ relies on analyzing the distribution of the integral, $\int_{\mathcal{T}} \{T_i(t) - \bar{T}(t)\} \widehat{H}_{jk}(\widehat{x}(t)) dt$, and the measurement error introduced by the estimated trajectory $\widehat{x}(t)$. These features clearly differentiate our de-biasing solution from the existing ones. Second, we note that a similar approach to constructing a de-biased solution involving an infinite dimensional object has been studied by Lu et al. (2020) in a different context. Third, we briefly examine an alternative de-biased estimator that uses the derivative instead of the integral in (19). The convergence of this alternative de-biased estimator hinges on the estimation error of the derivative term, $\mathbb{E} \int_{\mathcal{T}} \{d\widehat{x}_j(t)/dt - dx_j(t)/dt\}^2 dt$, which has a slower convergence rate than its integral counterpart (Chen et al., 2017; Dai and Li, 2022). As such, it is to have inferior inference properties, and we choose to use the integral instead of the derivative in our de-biased estimator (19).

Next, we obtain the critical value for the confidence band based on the de-biased estimator (19) using Gaussian multiplier bootstrap. Specifically, we consider the distribution of the supremum of the empirical process, $\sup_{t_0 \in \mathcal{T}} \mathbb{H}_n(t_0)$, where

$$\mathbb{H}_n(t_0) \equiv \sqrt{nh} \left[\widehat{F}_{jk}(x_k(t_0)) - F_{jk}(x_k(t_0)) \right], \quad \text{for any } t_0 \in \mathcal{T}.$$

Recognizing that the finite sample distribution of $\mathbb{H}_n(t_0)$ is unknown, we approximate the distribution of $\mathbb{H}_n(t_0)$ by the Gaussian multiplier process (Chernozhukov et al., 2014),

$$\widehat{\mathbb{H}}_n(t_0) \equiv \frac{1}{\sqrt{nh^{-1}}} \sum_{i=1}^{n} \xi_i \cdot \frac{\widehat{\sigma}_j R_h(t_i - t_0) R_{t_0,i\cdot}^{\top} \Sigma_{k\cdot}}{\widehat{\sigma}_n(t_0)}, \quad \text{for any } t_0 \in \mathcal{T},$$

where $\xi_1, \ldots, \xi_n$ are independent standard normal random variables, the error variance estimator $\widehat{\sigma}_j^2 = \|A_j(y_j - \bar{y}_j) - (y_j - \bar{y}_j)\|^2 / \mathrm{trace}(I_{n\times n} - A_j)$, with $A_j = I_{n\times n} - n\eta_{nj} M_{t_0}^{-1} \big[I_{n\times n} - \bar{t} \left(\bar{t}^{\top} M_{t_0}^{-1} \bar{t} \right)^{-1} \bar{t}^{\top} M_{t_0}^{-1} \big] \in \mathbb{R}^{n\times n}$ being the smoothing matrix, and $M_{t_0} = \Sigma R_{t_0}^{-1} + n\eta_{nj} R_{t_0}^{-2} \in \mathbb{R}^{n\times n}$, and $\widehat{\sigma}_n^2(t_0) = n^{-1} \Sigma_{k\cdot} R_{t_0}^2 \Sigma_{k\cdot}$, and $R_{t_0,i\cdot}, i = 1, \ldots, n$, being the ith row of the $n \times n$ matrix R_{t_0} defined in (16). We then compute the critical value $\widehat{c}_n(\alpha)$ as the $(1-\alpha)$-quantile of the supremum of the empirical process $\sup_{t_0 \in \mathcal{T}} \widehat{\mathbb{H}}_n(t_0)$ given the observed data (Giné and Zinn, 1990; Chernozhukov et al., 2014).

Finally, we construct the $100 \times (1-\alpha)\%$ confidence band for $F_{jk}(x_k(t))$ based on the de-biased estimator $\widehat{F}_{jk}(x_k(t_0))$ in (19). The de-biasing step essentially removes the impact of regularization bias on the estimation of the individual regulatory effect, which in turn

guarantees the validity of the confidence band. We also remark that our proposed de-biasing is built upon but also considerably extends the existing de-biasing methods in high-dimensional inference such as Zhang and Zhang (2014); van de Geer et al. (2014), because our setting is more challenging, and it involves the dynamic ODE system with an infinite-dimensional functional object as well as additional measurement error. Specifically, we construct the $100 \times (1-\alpha)\%$ confidence band as,

$$\mathcal{C}_{n,\alpha} = \left\{ \left[\widehat{F}_{jk}(x_k(t_0)) - \frac{\widehat{c}_n(\alpha)\widehat{\sigma}_n(t_0)}{\sqrt{nh}},\ \ \widehat{F}_{jk}(x_k(t_0)) + \frac{\widehat{c}_n(\alpha)\widehat{\sigma}_n(t_0)}{\sqrt{nh}} \right] \ \middle|\ t_0 \in \mathcal{T} \right\}, \tag{20}$$

Given the confidence band in (20), we can perform hypothesis testing for any given pair $(j,k) \in \{1,\ldots,p\}$ and any $t_0 \in \mathcal{T}$ that,

$$H_{0,jk} : x_k(t_0) \text{ has no regulatory effect on } x_j(t_0)$$
$$H_{1,jk} : x_k(t_0) \text{ has nonzero regulatory effect on } x_j(t_0).$$

We use the standardized regulatory effect as the test statistic,

$$z_{jk}(t_0) = \frac{\widehat{F}_{jk}(x_k(t_0))\sqrt{nh}}{\widehat{\sigma}_n(t_0)}, \quad \text{for any } t_0 \in \mathcal{T},$$

which follows the asymptotic distribution $\mathbb{H}_n(t_0)$ under the null hypothesis. In practice, we apply this test to a set of grid points $t_0 \in \mathcal{T}$, and we reject the null that $x_k(t)$ has no regulatory effect on $x_j(t)$, if any single test rejects the null at a given t_0.

Moreover, the confidence band (20) is for the inference of the regulatory effect of $x_k(t)$ on $x_j(t)$ for a given pair (j,k). We can easily couple it with existing multiple testing procedure for *all* pairs of (j,k) to recover the entire regulatory system, e.g., the Benjamini–Hochberg procedure, while controlling the FDR (Benjamini and Hochberg, 1995).

4.2 Extension to Multiple Experiments

The localized kernel learning method we have developed so far focuses on a single experiment. Meanwhile, it can be easily generalized to incorporate multiple experiments. Specifically, let $\left\{y_{ij}^{(s)}; i = 1,\ldots,n, j = 1,\ldots,p, s = 1,\ldots,S\right\}$ denote the observed data from n subjects for p variables under S experiments, with unknown initial conditions $x^{(s)}(0) \in \mathbb{R}^p, s = 1,\ldots,S$. Then we modify the localized kernel learning in (11), by seeking $\widehat{H}_{jk} \in \mathcal{H}$ and $\widehat{\alpha}_{jk,t_0} \in \mathbb{R}$ that minimize

$$\frac{1}{Sn}\sum_{s=1}^{S}\sum_{i=1}^{n} R_h(t_i - t_0)\left[\left(y_{ij}^{(s)} - \bar{y}_j^{(s)}\right) - \alpha_{jk,t_0}\bar{t}_i - \int_{\mathcal{T}}\left(T_i(t) - \bar{T}(t)\right)H_{jk}(\widehat{x}^{(s)}(t))dt\right]^2$$
$$+ \tau_{nj}\left(\sum_{k'=1,k'\neq k}^{p}\|\mathcal{P}^l H_{jk}\|_{\mathcal{H}} + \sum_{k'=1,k'\neq k,l}^{p}\sum_{l=1,l\neq k}^{p}\|\mathcal{P}^{k'l}H_{jk}\|_{\mathcal{H}}\right),$$

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

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

The de-biased estimator in (19) becomes,

$$\widehat{F}_{jk}(x_k(t_0)) = \widehat{\alpha}_{jk,t_0} + \frac{1}{Sn}\sum_{s=1}^{S}\sum_{i=1}^{n}\Sigma_{i\cdot}^{(s)} R_{t_0}\left[\left(y_j^{(s)} - \bar{y}_j^{(s)}\right) - \int_{\mathcal{T}}\left\{T_i(t) - \bar{T}(t)\right\}\widehat{H}_{jk}(\widehat{x}^{(s)}(t))dt\right].$$

The rest of Algorithm 1 for estimation and inference remains largely the same.

5. Theoretical Properties

In this section, we first establish the convergence rate for the localized kernel ODE estimator, which characterizes its estimation accuracy and is needed for establishing the properties of the subsequent inference. We then establish the asymptotic validity in terms of the coverage property for both the constructed confidence band and the recovered regulatory system. Our theoretical results hold for both the low-dimensional setting and the high-dimensional setting, where the number of variables p can be smaller or larger than the sample size n. We also study the regime-switching phenomenon.

5.1 Statistical Convergence Rate

We begin with three regularity conditions.

Assumption 1 *The kernel density $R(t)$ is a continuous function that has a bounded support, and satisfies that $\int_{\mathcal{T}} R(t)dt = 1$ and $\int_{\mathcal{T}} tR(t)dt = 0$.*

Assumption 2 *The number of nonzero functional components, card$\big(\{k' : F_{jk'} \neq 0\} \cup \{1 \leq k' \neq l \leq p : F_{jk'l} \neq 0\}\big)$, is bounded, for any $j = 1, \ldots, p$.*

Assumption 3 *For any $F_j \in \mathcal{H}$, there exists a random variable B, with $\mathbb{E}(B) < \infty$, and $|\partial F_j(x)/\partial x_k| \leq B\|F_j\|_{L_2}$ almost surely.*

Assumption 1 is standard in the local polynomial regression literature (Fan and Gijbels, 1996). Assumption 2 regards the complexity of the functionals. Similar assumptions have been adopted in the sparse additive model over RKHS without interactions (Koltchinskii and Yuan, 2010; Raskutti et al., 2011). Assumption 3 is an inverse Poincaré inequality type condition, which places a regularization on the fluctuation in F_j relative to the L_2-norm. The same assumption has also been used in additive models in RKHS (Zhu et al., 2014; Dai and Li, 2022).

Recall $F_j(x(t))$ represents the true functional in (3), and $\widehat{F}_j(\widehat{x}(t))$ the proposed localized kernel estimator in (12). The next theorem obtains the rate of convergence of $\widehat{F}_j(\widehat{x}(t))$.

Theorem 1 *Suppose Assumptions 1 to 3 hold. Suppose $x_j(t) \in \mathcal{F}$, and the RKHS $\mathcal{F}$ is embedded to a β_1th-order Sobolev space with $\beta_1 > 1/2$. Suppose $F_j \in \mathcal{H}$, where $\mathcal{H}$ satisfies (4), and the RKHS $\mathcal{H}_k$ in (4) is embedded to a β_2th-order Sobolev space with $\beta_2 > 1$, $j, k = 1, \ldots, p$. Then, the localized kernel ODE estimator $\widehat{F}_j(\widehat{x}(t))$ in (12) satisfies that,*

$$\min_{\lambda_{nj},\tau_{nj}\geq 0}\int_{\mathcal{T}}\left[\widehat{F}_j(\widehat{x}(t)) - F_j(x(t))\right]^2 dt = O_p\left(\left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta 2+1}} + h^{2\beta_2} + \frac{\log p}{n} + n^{-\frac{2\beta_1}{2\beta_1+1}}\right). \tag{21}$$

Furthermore, if $h = O\left((n/\log n)^{-1/(2\beta_2+2)}\right)$*, then the localized kernel ODE estimator in* (12) *satisfies that,*

$$\min_{\lambda_{nj},\tau_{nj}\geq 0}\int_{\mathcal{T}}\left[\widehat{F}_j(\widehat{x}(t))-F_j(x(t))\right]^2 dt = O_p\left(\left(\frac{n}{\log n}\right)^{-\frac{\beta_2}{\beta_2+1}}+\frac{\log p}{n}+n^{-\frac{2\beta_1}{2\beta_1+1}}\right). \quad (22)$$

We first note that the convergence rate in (21) is established for the estimator $\widehat{F}_j(\widehat{x}(t))$ of the entire regulatory effect $F_j(x(t))$. We can further obtain the convergence rate for the estimators of the individual effect and nuisance parameter $(\alpha_{jk,t}, H_{jk})$, which turns to be the same as the rate in (21). Specifically, for the estimators $(\widehat{\alpha}_{jk,t}, \widehat{H}_{jk})$ from (11), we have,

$$\min_{\lambda_{nj},\tau_{nj}\geq 0}\int_{\mathcal{T}}\left[\widehat{\alpha}_{jk,t}-F_{jk}(x(t))\right]^2 dt = O_p\left(\left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta 2+1}}+h^{2\beta_2}+\frac{\log p}{n}+n^{-\frac{2\beta_1}{2\beta_1+1}}\right),$$

$$\min_{\lambda_{nj},\tau_{nj}\geq 0}\int_{\mathcal{T}}\left[\widehat{H}_{jk}(\widehat{x}(t))-H_{jk}(x(t))\right]^2 dt = O_p\left(\left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta 2+1}}+h^{2\beta_2}+\frac{\log p}{n}+n^{-\frac{2\beta_1}{2\beta_1+1}}\right).$$

We next examine the sources of the estimation error in (21). There are totally four sources. Specifically, the first source of error $O_p((nh/\log n)^{-2\beta_2/(2\beta_2+1)})$ comes from the error in estimating the interaction terms in the true functional F_j. The second term $O_p(h^{2\beta_2})$ comes from the localized estimation. The third term $O_p(\log p/n)$ comes from the bias introduced by the Lasso estimation. The last term $O_p(n^{-2\beta_1/(2\beta_1+1)})$ comes from the measurement errors in $x(t)$.

We also observe some interesting regime-switching phenomenon in (22). That is, when p is ultrahigh-dimensional, in that $p > \exp[\{n(\log n)^{\beta_2}\}^{1/(\beta_2+1)}]$, then the convergence rate in (22) becomes $O_p(\log p/n + n^{-2\beta_1/(2\beta_1+1)})$. In this case, it matches with the minimax optimal rate for estimating the functional F_j in (3) obtained in Dai and Li (2022). Henceforth, we pay no extra price in terms of the rate of convergence for adopting the localized estimation in this ultrahigh-dimensional setting. On the other hand, when p is low-dimensional, in that $p < \exp[\{n(\log n)^{\beta_2}\}^{1/(\beta_2+1)}]$, then the convergence rate in (22) becomes $O_p((n/\log n)^{-\beta_2/(\beta_2+1)} + n^{-2\beta_1/(2\beta_1+1)})$. Here, the first term $O_p((n/\log n)^{-\beta_2/(\beta_2+1)})$ matches, up to some logarithmic factors, with the rate as if we knew a priori that F_j is an additive model with pairwise interactions. Moreover, the rate $O_p((n/\log n)^{-\beta_2/(\beta_2+1)})$ in our proposed method is slower than the optimal rate $O_p((n/\log n)^{-2\beta_2/(2\beta_2+1)})$ of the kernel ODE estimator of Dai and Li (2022). Such a slower rate is due to the weight matrix R_h utilized in the localized estimation in (11), which increases the variance of estimating F_j by $O_p(h^{-1})$, when compared to the non-localized estimation method in Dai and Li (2022). Nevertheless, as we show next, this slower rate is sufficient to establish an asymptotically valid confidence band.

5.2 Coverage Property

A confidence band $\mathcal{C}_{n,\alpha}$ is said to be asymptotically valid with level $100\times(1-\alpha)\%$ for F_{jk}, if it satisfies that, for some constants $c, C > 0$,

$$\mathbb{P}\left(F_{jk}(x_k(t_0)) \in \mathcal{C}_{n,\alpha},\ \forall t_0 \in \mathcal{T}\right) \geq 1 - \alpha - Cn^{-c}. \tag{23}$$

The condition (23) implies that the confidence band $\mathcal{C}_{n,\alpha}$ has an asymptotic coverage probability of at least $1-\alpha$ for a given data generating process.

The next theorem establishes the theoretical property of our proposed confidence band $\mathcal{C}_{n,\alpha}$ in (20). We introduce another regularity condition. For any $t \in \mathcal{T}$, let p_j denote the marginal density of $x_j(t)$, p_{jk} denote the bivariate density of $(x_j(t), x_k(t))$, and p_{jkl} denote the joint density of $(x_j(t), x_k(t), x_l(t))$, for $j, k, l = 1, \ldots, p$.

Assumption 4 *The density function of $x(t)$ satisfies that,*

$$\sum_{j=1, j\neq k}^{p} \|p_{jk} - p_j p_k\|_2 \leq \frac{\rho_{\min}}{2B}, \quad \textit{and} \quad \sup_{l\neq k} \sum_{j,k=1, j\neq k}^{p} \|p_{j,k,l} - p_j p_k p_l\|_2 \leq \frac{\rho_{\min}}{2B},$$

for some constant $\rho_{\min} > 0$, and B is as defined in Assumption 3.

Assumption 4 quantifies how weak the dependency between the signal variables can be. Similar conditions have also been used in the inference of the high-dimensional linear regression (Zhang and Zhang, 2014; van de Geer et al., 2014), and the nonparametric additive regression (Lu et al., 2020). We show this condition is also sufficient for establishing the validity of the confidence band in (20) in a complex ODE system.

Theorem 2 *Suppose Assumptions 1 to 4 hold. If $h = O(n^{-r})$, for $r \in (1/5, 3/13)$, then there exist constants $c, C_1 > 0$, such that, for any $\alpha \in (0,1)$, the confidence band $\mathcal{C}_{n,\alpha}$ in (20) is asymptotically valid.*

Putting all the functionals $\{F_1, \ldots, F_p\}$ together forms a network of regulatory relations among the p variables $\{x_1(t), \ldots, x_p(t)\}$. The next theorem shows that the estimated system based on our localized kernel ODE approaches the truth with probability tending to one. Denote the set of the true and the estimated regulators of $x_j(t)$ by

$$S_j^* = \left\{1 \leq k \leq p : F_{jk} \neq 0, \text{ or } F_{jkl} \neq 0 \text{ for some } 1 \leq l \neq k \leq p\right\},$$
$$\widehat{S}_j = \left\{1 \leq k \leq p : \widehat{\alpha}_{jk,t_0} \neq 0 \text{ for any } t_0 \in \mathcal{T}\right\}.$$

We also need two additional regularity conditions. Let $s_j = \mathrm{card}(S_j^*)$. Recall the definitions of $R_{t_0} \in \mathbb{R}^{n\times n}$ in (16), $G \in \mathbb{R}^{n\times(p-1)^2}$ in (18), and the tuning parameter κ_{nj} in (13). Let $G_{S_j^*} \in \mathbb{R}^{n\times s_j}$ denote the sub-matrix of G with the column indices in the set S_j^*.

Assumption 5 *Suppose there exists a constant $C_{\min} > 0$, such that the minimal eigenvalue of the matrix $G_{S_j^*}^\top R_{t_0} G_{S_j^*}$ is no smaller than $C_{\min}/2$ as $n \to \infty$. In addition, suppose there exists a constant $0 \leq \xi_G < 1$, such that $\max_{(k,l)\notin S_j^*} \left\| G_{kl}^\top R_{t_0} G_{S_j^*} (G_{S_j^*}^\top R_{t_0} G_{S_j^*})^{-1} \right\|_{\ell_2} \leq \xi_G$.*

Assumption 6 *Let $\theta_{\min} = \min_{(k,l)\in S_j} \|\theta_{jkl}\|_{L_2}$, and $\eta_{\mathcal{R}}$ is given in Lemma 6 in Section B.3.2 of the Appendix. Suppose the following inequalities hold:*

$$\frac{\eta_{\mathcal{R}}\sqrt{s_j}}{C_{\min}} + n\kappa_{nj}\frac{\sqrt{s_j}}{C_{\min}} \leq \frac{2}{3}\theta_{\min}, \quad \textit{and} \quad \frac{(\xi_G+1)\sqrt{s_j}}{n\kappa_{nj}}\eta_{\mathcal{R}} + \xi_G\sqrt{s_j} < 1.$$

Assumption 5 ensures the sub-matrix $G_{S_j^*}$ is not degenerated, and the irrelevant variables would not exert too strong an effect on the relevant variables. It is similar to Assumptions 3 and 4 in Dai and Li (2022) except for the additional term R_{t_0}. It is interesting to note that the bandwidth h in the localization and R_{t_0} does not affect the validity of this assumption, as long as $h \to 0$ when $n \to \infty$. More discussion on this assumption is given in Section C.2 of the Appendix. Assumption 6 imposes some regularity on the minimum effect $\theta_{\min}$, and also characterizes the relationship among ξ_G, κ_{nj}, and s_j. Both assumptions are mild, and similar conditions as Assumptions 5 and 6 have been commonly imposed in the literature (see, e.g., Zhao and Yu, 2006; Meinshausen and Bühlmann, 2006; Ravikumar et al., 2010; Chen et al., 2017; Dai and Li, 2022).

Theorem 3 *Suppose Assumptions 1, 2, 5 and 6 hold. Then, the localized kernel ODE estimator correctly recovers the true regulatory system, in that,*

$$\mathbb{P}\left(\widehat{S}_j = S_j^*\right) \to 1, \quad as \;\; n \to \infty, \quad for\ all\ j = 1, \ldots, p.$$

6. Simulation Studies

In this section, we study the finite-sample performance of the confidence band as well as the localized kernel ODE estimator using two well-known ODE systems. In the first example, we focus on the coverage of the confidence band for some *given pairs*, while in the second example, we study the performance of recovery of the *entire* regulatory system.

6.1 Enzymatic Regulation Equations

The first example is a three-node enzyme regulatory system of a negative feedback loop with a buffering node (Ma et al., 2009, NFBLB). The ODE system is given by,

$$\begin{aligned}
\frac{dx_1(t)}{dt} &= 10\frac{x_0\{1-x_1(t)\}}{\{1-x_1(t)\}+0.1} - 10\frac{x_1(t)}{x_1(t)+0.1}, \\
\frac{dx_2(t)}{dt} &= 10\frac{\{1-x_2(t)\}x_3(t)}{\{1-x_2(t)\}+0.1} - 0.2\frac{x_2(t)}{x_2(t)+0.1}, \\
\frac{dx_3(t)}{dt} &= 10\frac{x_1(t)\{1-x_3(t)\}}{\{1-x_3(t)\}+0.1} - 10\frac{x_2(t)x_3(t)}{x_3(t)+0.1}.
\end{aligned} \tag{24}$$

The coefficient $x_0 \in \mathbb{R}$ is drawn uniformly from $[0.5, 1.5]$. The initial values are chosen as $(x_1(0), x_2(0), x_3(0)) = (0, 0, 0)$. The errors are drawn independently from Normal$(0, \sigma_j^2)$, with three noise levels, $\sigma_j \in \{0.1, 0.3, 0.5\}$. The time points are evenly distributed, $t_i = (i-1)/20, i = 1, \ldots, n$, with the sample size fixed at $n = 40$. In this example, $p = 3$, and there are $p^2 = 9$ functions in (24) to estimate for each $j = 1, 2, 3$, and in total there are 27 unknown functions.

In this example, we focus on the performance of the confidence band for some given node pairs. Specifically, we examine a nonzero effect $F_{23}(x_3(t))$ that captures the regulatory effect of $x_3(t)$ on $x_2(t)$, and a zero effect $F_{12}(x_2(t))$ of $x_2(t)$ on $x_1(t)$. Correspondingly, we construct the confidence band for $F_{23}(x_3(t))$ and $F_{12}(x_2(t))$, with the 95% significance level, on 500 evenly distributed grid points on $[0, 1]$. In our implementation, we use a first-order

Matérn kernel for both steps in (8) and (10) of the collocation method, where $K_{\mathcal{H}_1}(x, x') = (1 + \sqrt{3}\|x - x'\|/\nu)\exp(-\sqrt{3}\|x - x'\|/\nu)$, and ν is chosen by tenfold cross-validation. We have found the inference results are not overly sensitive to the choice of kernel functions here. Moreover, we use the quadratic density $R_h(t) = (15/16)\cdot(1-t^2/h^2)^2\mathbf{1}(|t| < h)$ for the local weight function, where the bandwidth h is chosen by tenfold cross-validation. We have carried out a sensitivity analysis in Section D.2 of the Appendix, and show that the inference results are again not sensitive to the choice of the weight function or the bandwidth. We compute the quantile $\widehat{c}_n(\alpha)$ in (20) by bootstrap with 500 repetitions.

We note that there is *no* direct competitor to our confidence band solution in the literature. Alternatively, we compare to three commonly used ODE solutions, the linear ODE with interactions (Zhang et al., 2015a), the additive ODE (Chen et al., 2017), and the kernel ODE (Dai and Li, 2022), and couple them with a confidence band that aggregates the point-wise confidence intervals at 500 grid points on $[0, 1]$. For a fair comparison, we adjust the significance level at each of these 500 time points with the Bonferroni correction (Holm, 1979), i.e., $(1 - \alpha/500)\%$, where $\alpha = 0.05$.

We consider two evaluation criteria. One is the empirical coverage probability of the confidence band, and the other is the area of the confidence band, defined as

$$\int_{t_0\in\mathcal{T}} 2\widehat{c}_n(\alpha)(nh)^{-1/2}\widehat{\sigma}_n(t_0)dt_0,$$

where the integration is computed by discretizing the interval into 1000 grids. A larger coverage probability and a smaller area indicates a better performance.

Table 1 reports the results based on 500 data replications. We see that the proposed confidence band achieves the desired coverage. By contrast, the confidence bands of the additive and linear ODEs 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 (24), and this discrepancy accumulates as the course of the ODE evolves. Meanwhile, the kernel ODE has a much larger confidence band compared to our method. This is because the Bonferroni correction makes the confidence band of kernel ODE overly conservative. We report some additional results graphically in Section D.1 of the Appendix.

6.2 Lotka-Volterra Equations

The second example is the classical Lotka-Volterra system, which consists of pairs of first-order nonlinear differential equations describing the dynamics of biological system in which predators and prey interact (Volterra, 1928). The ODE is given by,

$$\begin{aligned}\frac{dx_{2j-1}(t)}{dt} &= 0.1(2j+11)x_{2j-1}(t) - 0.2(j+1)x_{2j-1}(t)x_{2j}(t),\\ \frac{dx_{2j}(t)}{dt} &= 0.1(2j-1)x_{2j-1}(t)x_{2j}(t) - 0.2(j+1)x_{2j}(t),\end{aligned} \tag{25}$$

for $j = 1, \ldots, 5$. Here $dx_{2j-1}(t)/dt$ 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. The initial values are set as $x_{2j-1}(0) = x_{2j}(0)$. The measurement errors are independent Normal$(0, \sigma_j^2)$, where the noise level $\sigma_j \in \{1, 2, \ldots, 10\}$. The time points are evenly distributed in $[0, 100]$, with $n = 200$. In this

Noise level	Method	Nonzero functional $F_{23}(x_3(t))$ Coverage probability	Nonzero functional $F_{23}(x_3(t))$ Confidence band area	Zero functional $F_{12}(x_2(t))$ Coverage probability	Zero functional $F_{12}(x_2(t))$ Confidence band area
$\sigma_j = 0.1$	Linear ODE	0.212	1.550	0.274	2.145
		$(0.205, 0.219)$	$(1.514, 1.586)$	$(0.266, 0.282)$	$(2.113, 2.177)$
	Additive ODE	0.224	1.332	0.292	2.008
		$(0.216, 0.232)$	$(1.290, 1.374)$	$(0.285, 0.299)$	$(1.964, 2.052)$
	Kernel ODE	0.939	1.270	0.928	1.848
		$(0.926, 0.952)$	$(1.250, 1.290)$	$(0.916, 0.940)$	$(1.824, 1.872)$
	Localized kernel ODE	**0.974**	**0.102**	**0.957**	**0.217**
		$(\mathbf{0.968}, \mathbf{0.980})$	$(\mathbf{0.092}, \mathbf{0.112})$	$(\mathbf{0.952}, \mathbf{0.962})$	$(\mathbf{0.210}, \mathbf{0.224})$
$\sigma_j = 0.3$	Linear ODE	0.178	1.827	0.224	2.441
		$(0.167, 0.189)$	$(1.776, 1.878)$	$(0.212, 0.236)$	$(2.382, 2.500)$
	Additive ODE	0.194	1.663	0.262	2.158
		$(0.184, 0.204)$	$(1.596, 1.730)$	$(0.252, 0.272)$	$(2.089, 2.227)$
	Kernel ODE	0.914	1.381	0.911	1.856
		$(0.891, 0.937)$	$(1.350, 1.412)$	$(0.891, 0.931)$	$(1.821, 1.891)$
	Localized kernel ODE	**0.962**	**0.163**	**0.957**	**0.290**
		$(\mathbf{0.952}, \mathbf{0.972})$	$(\mathbf{0.150}, \mathbf{0.176})$	$(\mathbf{0.949}, \mathbf{0.965})$	$(\mathbf{0.281}, \mathbf{0.299})$
$\sigma_j = 0.5$	Linear ODE	0.123	3.541	0.191	2.538
		$(0.103, 0.143)$	$(3.456, 3.626)$	$(0.172, 0.210)$	$(2.447, 2.629)$
	Additive ODE	0.141	2.679	0.225	2.350
		$(0.122, 0.160)$	$(2.586, 2.772)$	$(0.205, 0.245)$	$(2.251, 2.449)$
	Kernel ODE	0.876	1.637	0.902	2.031
		$(0.843, 0.909)$	$(1.595, 1.679)$	$(0.873, 0.931)$	$(1.980, 2.082)$
	Localized kernel ODE	**0.956**	**0.231**	**0.948**	**0.401**
		$(\mathbf{0.943}, \mathbf{0.969})$	$(\mathbf{0.216}, \mathbf{0.246})$	$(\mathbf{0.937}, \mathbf{0.959})$	$(\mathbf{0.389}, \mathbf{0.413})$

Table 1: The NFBLB example: the empirical coverage probability and area of the confidence band, and their 95% confidence intervals, for the varying noise level σ_j. The results are based on 500 data replications.

example, $p = 10$, and there are $p^2 = 100$ functions in (25) to estimate for each ODE of x_{2j-1} and x_{2j}, $j = 1, \ldots, 5$, and in total there are 1000 unknown functions.

In this example, we focus on the performance of recovery of the entire regulatory system through the proposed confidence band coupled with the Benjamini–Hochberg (BH) procedure for multiple testing correction (Benjamini and Hochberg, 1995). Since the ODE equations in (25) only involve the linear and interaction terms, we use the first-order Matérn kernel for the step in (8), and use the linear kernel in (10). As such, the linear and kernel ODEs yield the same estimates. We continue to use the quadratic density for the local weight function $R_h(t)$. We control the FDR at the level of 20%.

We consider three evaluation criteria, the false discovery proportion, the empirical power, and the trajectory prediction accuracy. The false discovery proportion is defined as the proportion of falsely selected edges in the system out of the total number of edges, and the empirical power is defined as the proportion of selected true edges in the system. We also evaluate the prediction accuracy of the entire regulatory effect $\widehat{F}_j(\widehat{x}(t))$ as given in (12), by the squared root of the sum of predictive mean squared errors for $F_j(x_j(t)), j = 1, \ldots, 10$, at

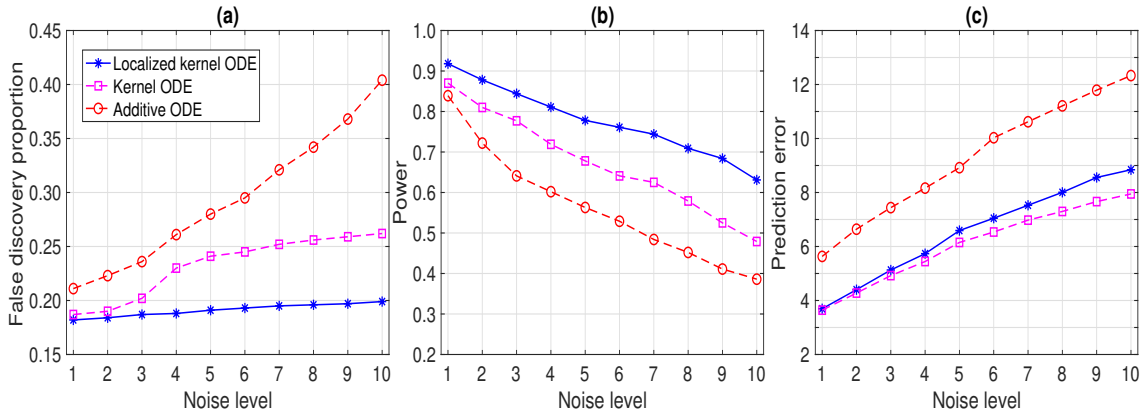

Figure 1: The Lotka-Volterra example: the empirical FDR, power, and trajectory prediction error for the varying noise level σ_j. The results are averaged over 500 data replications.

the unseen "future" time point $t \in [100, 200]$, i.e., $\left\{\sum_{j=1}^{10} \int_{100}^{200} [\widehat{F}_j(\widehat{x}_j(t)) - F_j(x_j(t))]^2 dt\right\}^{1/2}$, where the integral is evaluated at 10000 evenly distributed time points in $[100, 200]$.

Figure 1 reports the results averaged over 500 data replications. We see that our method successfully controls the FDR under the nominal level across all noise levels, whereas the additive and kernel ODEs both suffer some inflations, especially when the noise level is high. Meanwhile, our method achieves the best empirical power. In addition, we see that the prediction error of our localized kernel ODE estimator is slightly worse than that of kernel ODE, which agrees with Theorem 1. Nevertheless, this does not affect the inference performance of our method.

7. Data Applications

In this section, we illustrate our method with two data applications, a gene regulatory network analysis given time-course gene expression data, and a brain effective connectivity analysis given electrocorticography (ECoG) data.

7.1 Gene Regulatory Network

Gene regulation plays a central role in biological activities such as cell growth, development, and response to environmental stimulus (Peng et al., 2009; González et al., 2013). Thanks to the advancement of high-throughput DNA microarray technologies, it becomes feasible to measure the dynamic features of gene expression profiles on a genome scale. Such time-course gene expression data allow investigators to study gene regulatory networks, and ODE modeling is frequently employed for such a purpose (Lu et al., 2011). The data we analyze is the in silico benchmark gene expression data generated by GeneNetWeaver (GNW) using dynamical models of gene regulations and nonlinear ODEs (Schaffter et al., 2011). GNW extracts two regulatory networks of *E.coli*, *E.coli1*, *E.coli2*, and three regulatory networks of yeast, *yeast1*, *yeast2*, *yeast3*, each of which has two values of dimension, $p = 10$ nodes

		$p = 10$				$p = 100$			
		Localized kernel ODE	Kernel ODE	Additive ODE	Linear ODE	Localized kernel ODE	Kernel ODE	Additive ODE	Linear ODE
E.coli1	FDR	**0.182**	0.212	0.241	0.206	**0.191**	0.194	0.231	0.232
		(0.178, 0.186)	(0.207, 0.217)	(0.235, 0.247)	(0.199, 0.213)	(0.187, 0.195)	(0.191, 0.197)	(0.226, 0.236)	(0.229, 0.235)
	Power	**0.793**	0.587	0.547	0.467	**0.823**	0.611	0.512	0.481
		(0.789, 0.797)	(0.582, 0.592)	(0.541, 0.553)	(0.460, 0.474)	(0.818, 0.828)	(0.608, 0.614)	(0.507, 0.517)	(0.478, 0.484)
E.coli2	FDR	**0.193**	0.214	0.332	0.362	**0.186**	0.201	0.312	0.278
		(0.190, 0.196)	(0.210, 0.218)	(0.325, 0.339)	(0.355, 0.369)	(0.181, 0.191)	(0.197, 0.205)	(0.305, 0.319)	(0.278,)
	Power	**0.736**	0.666	0.639	0.569	**0.787**	0.684	0.649	0.575
		(0.733, 0.739)	(0.662, 0.670)	(0.632, 0.646)	(0.562, 0.576)	(0.782, 0.792)	(0.680, 0.688)	(0.642, 0.656)	(0.569, 0.681)
Yeast1	FDR	**0.181**	0.203	0.214	0.336	0.195	**0.193**	0.224	0.288
		(0.177, 0.185)	(0.199, 0.207)	(0.209, 0.219)	(0.330, 0.342)	(0.192, 0.198)	(0.190, 0.196)	(0.216, 0.232)	(0.281, 0.295)
	Power	**0.887**	0.607	0.546	0.442	**0.917**	0.642	0.522	0.498
		(0.883, 0.891)	(0.603, 0.611)	(0.541, 0.551)	(0.436, 0.448)	(0.914, 0.920)	(0.639, 0.645)	(0.514, 0.530)	(0.49810.505)
Yeast2	FDR	**0.174**	0.199	0.262	0.236	**0.189**	0.211	0.241	0.241
		(0.169, 0.179)	(0.195, 0.203)	(0.254, 0.270)	(0.230, 0.242)	(0.185, 0.193)	(0.208, 0.214)	(0.235, 0.247)	(0.236, 0.246)
	Power	**0.744**	0.603	0.570	0.542	**0.729**	0.582	0.535	0.611
		(0.739, 0.749)	(0.599, 0.607)	(0.562, 0.578)	(0.536, 0.548)	(0.725, 0.733)	(0.579, 0.585)	(0.529, 0.541)	(0.606, 0.616)
Yeast3	FDR	0.184	**0.181**	0.189	0.248	**0.190**	0.208	0.196	0.223
		(0.180, 0.188)	(0.177, 0.185)	(0.184, 0.194)	(0.242, 0.254)	(0.185, 0.195)	(0.204, 0.212)	(0.191, 0.201)	(0.219, 0.227)
	Power	**0.845**	0.616	0.573	0.493	**0.812**	0.577	0.539	0.472
		(0.841, 0.849)	(0.612, 0.620)	(0.568, 0.578)	(0.487, 0.499)	(0.807, 0.817)	(0.573, 0.581)	(0.534, 0.544)	(0.468, 0.476)

Table 2: The gene regulatory network example: the empirical FDR and power, and their 95% confidence intervals, for 10 combinations of network structures from GNW. The results are based on 500 data replications.

and $p = 100$ nodes. The system of ODEs for each extracted network is based on a thermodynamic approach, and the resulting ODE system is non-additive and nonlinear (Marbach et al., 2010). For the 10-node network, GNW provides $S = 4$ perturbation experiments, and for the 100-node network, GNW provides $S = 46$ experiments. In each perturbation 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). All the trajectories are measured at $n = 21$ evenly spaced time points in $[0, 1]$. We add independent measurement errors from Normal$(0, 0.025^2)$, which is the same as the DREAM3 competition and the data analysis in Henderson and Michailidis (2014).

We apply the proposed confidence band approach, coupled with the BH procedure at the 20% FDR level, to this data. Table 2 reports the empirical FDR and power for the recovered regulatory network for all ten combinations of network structures, and the results are based on 500 data replications. We compare with the alternative methods of linear, additive, and kernel ODEs, similarly as in Section 6.1. We see clearly that the proposed method performs competitively in all cases. This example 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.

7.2 Brain Effective Connectivity Analysis

Brain effective connectivity refers explicitly to the directional influence that one neural system exerts over another (Friston, 2011), and is of central interest in neuroscience research. Effective connectivity analysis uncovers such directional influences among different brain regions through imaging techniques such as electrocorticographic (ECoG), and modeling techniques such as ODE (Zhang et al., 2015a). The data we analyze is an ECoG study of

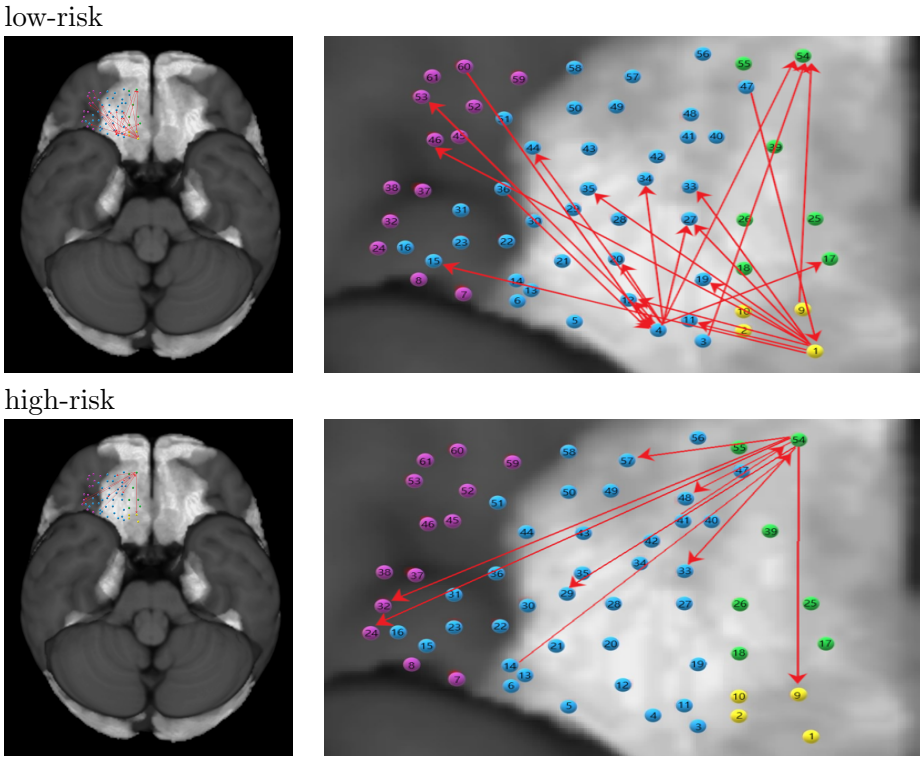

Figure 2: The brain effective connectivity example: the connectivity patterns during the low-risk and high-risk games. The colored nodes correspond to different cytoarchitectural regions of orbitofrontal cortex. Green: Fo1; yellow: Fo2; blue: Fo3; purple: other regions. The left panels are for the entire brain, and the right panels for the enlarged areas.

the brain during decision making (Saez et al., 2018). It consists of the ECoG recordings of $p = 61$ electrodes placed in the orbitofrontal cortex (OFC) region of an epilepsy patient when performing gambling tasks with different levels of winning risk. The patient performed 72 rounds of gambling games in total, half of which are low-risk games, and half are high-risk games. We analyze the low-risk and high-risk games separately, with $S = 36$. The length of the ECoG signals for each round of game is $n = 3001$. See Saez et al. (2018) for more details about the data collection and processing.

We again apply the proposed confidence band approach, coupled with the BH procedure at the 20% FDR level, to this data. To identify the nodes that show different connectivity patterns under two risk groups, we focus on those nodes whose total number of inward and outward edges is no more than 2 in one risk type, but no fewer than 9 in the other risk type. This results in node 1 that is located in the cytoarchitectural region of OFC called Fo2, node 4 located in Fo3, and node 54 located in Fo1 (Henssen et al., 2016). Figure 2 plots the estimated connectivity patterns of these three nodes, first on the entire brain, then in an amplified area. We see that, for nodes 1 and 4, there are many more outward edges during the low-risk games than the high-risk games, whereas for node 54, there are many more outward edges during the high-risk games than the low-risk games. Note that both Fo2 and

Fo3 belong to the posterior OFC, which is more involved in simple reward type decision making, whereas Fo1 belongs to the anterior OFC, which is involved in abstract reward (Kringelbach and Rolls, 2004). Our results suggest that the posterior OFC is more active during the low-risk games, in which the reward is relatively simple and clear. Meanwhile, the anterior OFC tends to more actively influence other nodes during the high-risk games, which involve more calculations and harder decisions. We briefly comment that, the arrow direction in the plot indicates if the estimated effect is from node $x_k(t)$ on $x_j(t)$ or from $x_j(t)$ on $x_k(t)$, $j, k = 1, \ldots, p$. In our analysis, we primarily focus on the total numbers of inward and outward edges.

8. Discussion

In this article, we aim at a central question in ODE modeling; that is, to infer the significance of individual regulatory effect of one signal variable on another. This question is challenging for ODE with unknown regulatory relations and noisy data observations, and remains largely untapped in the literature. We propose a new post-regularization confidence band method, which provides both an uncertainty quantification for the individual regulatory relation, and also a sparse recovery of the entire regulatory system when coupled with a proper FDR control. Our proposal involves two key ingredients: a new localized kernel learning approach that combines reproducing kernel learning with local polynomial learning, and a new debiasing method that tackles infinite-dimensional functionals and additional measurement errors. We establish the theoretical guarantees, and demonstrate the efficacy of the proposed method through numerical analyses.

An interesting extension of the proposed method is to tackle the scenario of multiple experiments or subjects. In this article, we have primarily focused on the scenario of a single experiment or a single subject, and we propose to sum together the objective functions for multiple experiments in Section 4.2. However, in numerous applications, e.g., neuroscience, there may be considerable experiment-to-experiment, or subject-to-subject variability. How to effectively account for such variability is important, and warrants future research.

Acknowledgments

The authors thank two anonymous reviewers and the Action Editor for their invaluable feedback. Xiaowu Dai acknowledges support of the California Center for Population Research as a part of the Eunice Kennedy Shriver National Institute of Child Health and Human Development (NICHD) population research infrastructure grant P2C-HD041022. Lexin Li acknowledges support of NSF grant CIF-2102227, and NIH grants R01AG061303 and R01AG062542.

Appendices

Appendix A. Parameters

We present a list of main parameters in Table S1, along with their meanings and dimensions.

parameter	definition	dimension
$x(t) = (x_1(t), \ldots, x_p(t))^\top$	p variables of interest	$\mathbb{R}^p$
$F = \{F_1, \ldots, F_p\}$	p functionals of regulatory relations	$\mathbb{R}^p$
θ_{j0}	global intercept	$\mathbb{R}^1$
α_{jk,t_0}	regulatory effect of $x_k(t)$ on $x_j(t)$ at $t = t_0$	$\mathbb{R}^1$
H_{jk}	nuisance terms of evaluating effect of $x_k(t)$ on $x_j(t)$	$\mathbb{R}^1$
$c_j = (c_{1j}, \ldots, c_{nj})^\top$	parameters in the optimization (17)	$\mathbb{R}^n$
$\theta_j = \{\theta_{jk'}\}_{k' \neq k} \cup \{\theta_{jk'l}\}_{k' \neq k; l \neq k', k}$	parameters in the optimization (18)	$\mathbb{R}^{(p-1)^2}$

Table S1: List of parameters of the proposed ODE modeling.

Appendix B. Proofs

B.1 Proof of Theorem 1

We divide the proof of this theorem into two parts. We first present the main proof in Section B.1.1, then give two auxiliary lemmas useful for the proof of this theorem in Section B.1.2.

B.1.1 Main proof

Proof For $j = 1, \ldots, p$, write $\widehat{F}_j(\widehat{x}(t_0)) = \widehat{\theta}_{j0} + \widehat{\alpha}_{jk,t_0} + \widehat{H}_{jk}(\widehat{x}(t_0))$ for any $t_0 \in \mathcal{T}$. Write $F_j(x(t_0)) = \theta_{j0} + \alpha_{jk,t_0} + H_{jk}(x(t_0))$. Considering $\widehat{\theta}_{j0}$ that is given by (9), where $\widehat{\theta}_{j0} = \bar{y}_j - \int_{\mathcal{T}} \bar{T}(t)\tilde{F}_j(\widehat{x}(t))dt$, its convergence rate is the same as that of $\widehat{\alpha}_{jk,t_0} + \widehat{H}_{jk}(\widehat{x}(t_0))$. Therefore, we focus our attention on $\widehat{\alpha}_{jk,t_0}$ and $\widehat{H}_{jk}(\widehat{x}(t_0))$ in the subsequent proof.

Recall that $\widehat{\alpha}_{jk,t_0}$ and $\widehat{H}_{jk}$ are obtained from

$$\min_{\alpha_{jk,t_0}, H_{jk}} \left\{ \frac{1}{n} \sum_{i=1}^{n} R_h(t_i - t_0) \left[y_{ij} - \alpha_{jk,t_0} \bar{t}_i - \int_0^{t_i} H_{jk}(\widehat{x}(t))dt \right]^2 + \tau_{nj} J(H_{jk}) \right\},$$

where $J(H_{jk}) \equiv \sum_{k'=1, k' \neq k}^{p} \|\mathcal{P}^l H_{jk}\|_{\mathcal{H}} + \sum_{k'=1, k' \neq k, l}^{p} \sum_{l=1, l \neq k}^{p} \|\mathcal{P}^{k'l} H_{jk}\|_{\mathcal{H}}$, and $\bar{t}_i = t_i - n^{-1} \sum_{i=1}^{n} t_i$. Then we have that,

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^{n} R_h(t_i - t_0) \left\{ \alpha_{jk,t_0} \bar{t}_i + \int_0^{t_i} H_{jk}(x(t))dt + \epsilon_{ij} - \widehat{\alpha}_{jk,t_0} \bar{t}_i - \int_0^{t_i} \widehat{H}_{jk}(\widehat{x}(t))dt \right\}^2 \\ & \qquad + \tau_{nj} J(\widehat{H}_{jk}) \\ \leq \; & \frac{1}{n} \sum_{i=1}^{n} R_h(t_i - t_0) \left\{ \alpha_{jk,t_0} \bar{t}_i + \int_0^{t_i} H_{jk}(x(t))dt + \epsilon_{ij} - \alpha_{jk,t_0} \bar{t}_i - \int_0^{t_i} H_{jk}(\widehat{x}(t))dt \right\}^2 \\ & \qquad + \tau_{nj} J(H_{jk}). \end{aligned}$$

With the rearrangement of the terms, we have that,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} R_h\left(t_i - t_0\right)\left[(\alpha_{jk,t_0} - \widehat{\alpha}_{jk,t_0})\bar{t}_i + \int_0^{t_i}\left\{H_{jk}(x(t)) - \widehat{H}_{jk}(\widehat{x}(t))\right\}dt\right]^2 \\
&\qquad\qquad + \tau_{nj} J(\widehat{H}_{jk}) \\
\leq\ &\frac{2}{n}\sum_{i=1}^{n} R_h\left(t_i - t_0\right)\epsilon_{ij}\left[(\widehat{\alpha}_{jk,t_0} - \alpha_{jk,t_0})\bar{t}_i + \int_0^{t_i}\left\{\widehat{H}_{jk}(\widehat{x}(t)) - H_{jk}(\widehat{x}(t))\right\}dt\right] \\
&+ \frac{1}{n}\sum_{i=1}^{n} R_h\left(t_i - t_0\right)\left[\int_0^{t_i}\left\{H_{jk}(x(t)) - H_{jk}(\widehat{x}(t))\right\}dt\right]^2 + \tau_{nj} J(H_{jk}).
\end{aligned} \tag{S1}$$

By Assumption 2 and the Taylor expansion,

$$(\widehat{H}_{jk} - H_{jk})(\widehat{x}) = (\widehat{H}_{jk} - H_{jk})(x) + \frac{\partial}{\partial t}(\widehat{H}_{jk} - H_{jk})(x)(\widehat{x} - x) + o_p\left(\max_{l=1,\ldots,p}\|\widehat{x}_l - x_l\|_{L_2}\right),$$

where the Fréchet derivative of any $H_{jk}(\cdot) \in \mathcal{H}$ is defined as,

$$\frac{\partial}{\partial t} H_{jk}(x)(\widehat{x} - x) = \sum_{k=1}^{p} \frac{\partial H_{jk}(x)}{\partial x_k}(\widehat{x}_k - x_k).$$

Then the first term on the right-hand-side of (S1) can be written as,

$$\begin{aligned}
&\frac{2}{n}\sum_{i=1}^{n} R_h\left(t_i - t_0\right)\epsilon_{ij}\left[(\widehat{\alpha}_{jk,t_0} - \alpha_{jk,t_0})\bar{t}_i + \int_0^{t_i}\left\{\widehat{H}_{jk}(\widehat{x}(t)) - H_{jk}(\widehat{x}(t))\right\}dt\right] \\
&= \frac{2}{n}\sum_{i=1}^{n} R_h\left(t_i - t_0\right)\epsilon_{ij}\left[(\widehat{\alpha}_{jk,t_0} - \alpha_{jk,t_0})\bar{t}_i + \int_0^{t_i}(\widehat{H}_{jk} - H_{jk})(x(t))dt\right] \\
&+ \frac{2}{n}\sum_{i=1}^{n} R_h\left(t_i - t_0\right)\epsilon_{ij}\left[\int_0^{t_i}\frac{\partial}{\partial t}(\widehat{H}_{jk} - H_{jk})(x(t))\{\widehat{x}(t) - x(t)\}dt + o_p\left(\max_{l=1,\ldots,p}\|\widehat{x}_l - x_l\|_{L_2}\right)\right] \\
&\equiv \Delta_1 + \Delta_2.
\end{aligned}$$

Meanwhile, by the Taylor expansion, the first term on the left-hand-side of (S1) can be written as,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} R_h\left(t_i - t_0\right)\left[(\alpha_{jk,t_0} - \widehat{\alpha}_{jk,t_0})\bar{t}_i + \int_0^{t_i}\left\{H_{jk}(x(t)) - \widehat{H}_{jk}(\widehat{x}(t))\right\}dt\right]^2 \\
=\ &\frac{1}{n}\sum_{i=1}^{n} R_h\left(t_i - t_0\right)\left[(\alpha_{jk,t_0} - \widehat{\alpha}_{jk,t_0})\bar{t}_i + \int_0^{t_i}\left\{H_{jk}(x(t)) - \widehat{H}_{jk}(x(t))\right\}dt\right. \\
&\qquad\left. + \int_0^{t_1}\frac{\partial}{\partial t}\widehat{H}_{jk}(x(t))\{x(t) - \widehat{x}(t)\}dt + o_p\left(\max_{l=1,\ldots,p}\|\widehat{x}_l - x_l\|_{L_2}\right)\right]^2.
\end{aligned}$$

The right-hand-side of the above equation can be rewritten as

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} R_h\left(t_i - t_0\right)\left[(\alpha_{jk,t_0} - \widehat{\alpha}_{jk,t_0})\bar{t}_i + \int_0^{t_i}\left\{H_{jk}(x(t)) - \widehat{H}_{jk}(x(t))\right\}dt\right]^2 \\
&+ \frac{1}{n}\sum_{i=1}^{n} R_h\left(t_i - t_0\right)\left[\int_0^{t_i}\frac{\partial}{\partial t}\widehat{H}_{jk}(x(t))\{x(t) - \widehat{x}(t)\}dt\right]^2 \\
&+ \frac{2}{n}\sum_{i=1}^{n} R_h\left(t_i - t_0\right)\left[(\alpha_{jk,t_0} - \widehat{\alpha}_{jk,t_0})\bar{t}_i + \int_0^{t_i}\left\{H_{jk}(x(t)) - \widehat{H}_{jk}(x(t))\right\}dt\right] \\
&\qquad \times \int_0^{t_i}\frac{\partial}{\partial t}\widehat{H}_{jk}(x(t))\{\widehat{x}(t) - x(t)\}dt + \mathcal{R}_1 \\
&\equiv \widetilde{\Delta}_1 + \widetilde{\Delta}_2 + \widetilde{\Delta}_3 + \mathcal{R}_1,
\end{aligned}
$$

where the remainder term $\mathcal{R}_1$ is of the form,

$$
\begin{aligned}
\mathcal{R}_1 = \frac{1}{n}\sum_{i=1}^{n} R_h\left(t_i - t_0\right)\Bigg(&o_p\left(\max_{l=1,\ldots,p}\|\widehat{x}_l - x_l\|_{L_2}^2\right) - o_p\left(\max_{l=1,\ldots,p}\|\widehat{x}_l - x_l\|_{L_2}\right) \\
&\times\left[(\alpha_{jk,t_0} - \widehat{\alpha}_{jk,t_0})\bar{t}_i + \int_0^{t_i}\left\{H_{jk}(x(t)) - \widehat{H}_{jk}(x(t)) - \frac{\partial}{\partial t}\widehat{H}_{jk}(x(t))\{\widehat{x}(t) - x(t)\}\right\}dt\right]\Bigg).
\end{aligned}
$$

Denote $\Delta_3 \equiv n^{-1}\sum_{i=1}^{n} R_h\left(t_i - t_0\right)\left[\int_0^{t_i}\left\{H_{jk}(x(t)) - H_{jk}(\widehat{x}(t))\right\}dt\right]^2$. Then (S1) becomes,

$$
\widetilde{\Delta}_1 + \widetilde{\Delta}_2 + \widetilde{\Delta}_3 + \mathcal{R}_1 + \tau_{nj}J(\widehat{H}_{jk}) \le \Delta_1 + \Delta_2 + \Delta_3 + \tau_{nj}J(H_{jk}) \tag{S2}
$$

Our proof strategy is to derive the upper and lower bounds for the left and right-hand sides of (S2), respectively, then put them together. We first study the convergence rate of the estimator of the nuisance parameter H_{jk}. We then apply a similar proof procedure to obtain the rate of the estimator of the individual effect α_{jk,t_0}. Together, we obtain the rate for the estimator of the entire regulatory effect $F_j(x(t))$.

Step 1: Bounding the right-hand-side of (S2). We first bound the three terms $\Delta_1, \Delta_2, \Delta_3$ on the right-hand-side of (S2).

For Δ_1, by Lemma 4 and the Minkowski inequality, we have that, as $h = o(1)$,

$$
\begin{aligned}
\Delta_1 \le O_p\Bigg\{&\left\|\widehat{H}_{jk} - H_{jk}\right\|_{L_2}^2 \log^{-2}\left\|\widehat{H}_{jk} - H_{jk}\right\|_{L_2} + h^{2\beta_2} \\
&+ \left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}} + \frac{\log p}{n} + \sqrt{\frac{\log p}{n}}\left\|\widehat{H}_{jk} - H_{jk}\right\|_{L_2}\Bigg\}.
\end{aligned}
$$

For Δ_2, since $\beta_2 > 1$, $\partial K(x,\cdot)/\partial x_k \in \mathcal{H}$, and by the reproducing property, we have,

$$
\frac{\partial(\widehat{H}_{jk} - H_{jk})(x)}{\partial x_k} = \left\langle \widehat{H}_{jk} - H_{jk}, \frac{\partial K(x,\cdot)}{\partial x_k}\right\rangle_{\mathcal{H}} \le \|\widehat{H}_{jk} - H_{jk}\|_{\mathcal{H}}^{1/2}\left\|\frac{\partial K(x,\cdot)}{\partial x_k}\right\|_{\mathcal{H}}^{1/2} < \infty.
$$

Henceforth, $\partial(\widehat{H}_{jk} - H_{jk})(x)/\partial x_k \in \mathcal{H}$ for $k = 1, \ldots, p$. By Assumption 3, we have,

$$
\max_{k=1,\ldots,p}\left\{|\partial(\widehat{H}_{jk} - H_{jk})(x)/\partial x_k|\right\} \le B\|\widehat{H}_{jk} - H_{jk}\|_{L_2}, \quad \text{almost surely.}
$$

By the Cauchy-Schwarz inequality, we have,

$$\begin{aligned}\Delta_2 \leq\ & \frac{2c}{n}\sum_{i=1}^{n}|\epsilon_{ij}|R_h\left(t_i-t_0\right)\int_0^{t_i} B\|\widehat{H}_{jk}(x(t))-H_{jk}(x(t))\|_{L_2}\max_{k=1,\ldots,p}|\widehat{x}_k(t)-x_k(t)|dt \\ & + o_p\left(n^{-1/2}\max_{k=1,\ldots,p}\|x_k-\widehat{x}_k\|_{L_2}^2\right) \\ \leq\ & 2c\max_{k=1,\ldots,p}\|\widehat{x}_k-x_k\|_{L_2}\left\|\widehat{H}_{jk}(x(t))-H_{jk}(x(t))\right\|_{L_2}\frac{1}{n}\sum_{i=1}^{n}|\epsilon_{ij}B|R_h\left(t_i-t_0\right) \\ & + o_p\left(n^{-1/2}\max_{k=1,\ldots,p}\|x_k-\widehat{x}_k\|_{L_2}^2\right) \\ =\ & O_p\left(n^{\frac{-\beta_1}{2\beta_1+1}}\|\widehat{H}_{jk}(x(t))-H_{jk}(x(t))\|_{L_2}\right),\end{aligned}$$

for some constant c, where the last step is due to the strong law of large numbers and Theorem 3 of Dai and Li (2022).

For Δ_3, by the Taylor expansion and Assumption 2, we have,

$$\begin{aligned}\Delta_3 &\leq \frac{c}{n}\sum_{i=1}^{n}R_h\left(t_i-t_0\right)\left[\int_0^{t_i}\frac{\partial}{\partial t}H_{jk}(x(t))\{x(t)-\widehat{x}(t)\}+o_p\left(\max_{k=1,\ldots,p}\|x_k-\widehat{x}_k\|_{L_2}^2\right)dt\right]^2 \\ &\leq c'\|H_{jk}\|_{\mathcal{H}}^2\max_{k=1,\ldots,p}\|x_k-\widehat{x}_k\|_{L_2}^2+c'o_p\left(\max_{k=1,\ldots,p}\|x_k-\widehat{x}_k\|_{L_2}^2\right)=O_p\left(n^{\frac{-2\beta_1}{2\beta_1+1}}\right).\end{aligned} \tag{S3}$$

for some constant c, c', where the second step is by the Jensen's inequality.

Step 2: Bounding the left-hand-side of (S2). We next bound the terms $\widetilde{\Delta}_1, \widetilde{\Delta}_2, \widetilde{\Delta}_3$ and $\mathcal{R}_1$ on the left-hand-side of (S2).

For $\widetilde{\Delta}_1$, by Lemma 5, with probability at least $1-2p^{-c_1}$, for some constant $C>0$,

$$\begin{aligned}\widetilde{\Delta}_1 \geq\ & \left\|H_{jk}-\widehat{H}_{jk}\right\|_{L_2}^2 - C\Bigg\{\left\|H_{jk}-\widehat{H}_{jk}\right\|_{L_2}^2\log^{-2}\left\|H_{jk}-\widehat{H}_{jk}\right\|_{L_2}+h^{2\beta_2} \\ & +\left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}}+\frac{\log p}{n}+n^{-1/2}e^{-p}\Bigg\}.\end{aligned} \tag{S4}$$

For $\widetilde{\Delta}_2$, we can drop this term, because $\widetilde{\Delta}_2 \geq 0$.

For $\widetilde{\Delta}_3$, by the Cauchy-Schwarz inequality,

$$\begin{aligned}\widetilde{\Delta}_3 \geq\ & -2\left(\frac{1}{n}\sum_{i=1}^{n}\left[\int_0^{t_i}\left\{R_h(t_i-t_0)\left[\widehat{H}_{jk}(x(t))-H_{jk}(x(t))\right]\right\}dt\right]^2\right)^{1/2} \\ & \times\left(\frac{1}{n}\sum_{i=1}^{n}\left[\int_0^{t_i}\frac{\partial}{\partial t}\widehat{H}_{jk}(x(t))\{\widehat{x}(t)-x(t)\}dt\right]^2\right)^{1/2}.\end{aligned}$$

Henceforth,

$$\begin{aligned}\widetilde{\Delta}_3 \geq\ & -2\left\|\widehat{F}_j(x(t))-F_j(x(t))\right\|_{L_2}\|F_j\|_{\mathcal{H}}\max_{k=1,\ldots,p}\|x_k-\widehat{x}_k\|_{L_2} \\ =\ & O_p\left(n^{\frac{-\beta_1}{2\beta_1+1}}\left\|\widehat{F}_j(x(t))-F_j(x(t))\right\|_{L_2}\right),\end{aligned}$$

where the second step is due to the Minkowski inequality.

For the remainder term $\mathcal{R}_1$ on the left-hand-side of (S2), by Assumption 2 and the Cauchy-Schwarz inequality, we have,

$$\begin{aligned}\mathcal{R}_1 &= o_p\left(\max_{k=1,\ldots,p}\|x_k-\widehat{x}_k\|_{L_2}\left\|\widehat{H}_{jk}-H_{jk}\right\|_{L_2}+\max_{k=1,\ldots,p}\|x_k-\widehat{x}_k\|_{L_2}^2\|H_{jk}\|_{\mathcal{H}}\right)\\ &= o_p\left(n^{\frac{-\beta_1}{2\beta_1+1}}\left\|\widehat{H}_{jk}-H_{jk}\right\|_{L_2}\right)+o_p\left(n^{\frac{-2\beta_1}{2\beta_1+1}}\right),\end{aligned}$$

where the second step is again due to the Minkowski inequality.

Step 3: Putting the two bounds together. Combining the bounds for each term in (S2), we obtain that, for any $c_1>0$ and $c_2>1$, with probability at least $1-4p^{-c_1}$, there exists a constant $C>0$, such that

$$\begin{aligned}&\left\|H_{jk}-\widehat{H}_{jk}\right\|_{L_2}^2\log^{-2}\left\|H_{jk}-\widehat{H}_{jk}\right\|_{L_2}\\ \leq C\Bigg[&c_2^{-\frac{4\beta_2}{2\beta_2-1}}\left\|H_{jk}-\widehat{H}_{jk}\right\|_{L_2}^2\log^{-2}\left\|H_{jk}-\widehat{H}_{jk}\right\|_{L_2}+c_2^{\frac{4\beta_2}{4\beta_2+1}}\left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta 2+1}}\\ &+(c_1+1)\frac{\log p}{n}+\sqrt{(c_1+1)\frac{\log p}{n}}\left\|H_{jk}-\widehat{H}_{jk}\right\|_{L_2}+h^{2\beta_2}\\ &+n^{-1/2}e^{-p}+n^{\frac{-\beta_1}{2\beta_1+1}}\left\|\widehat{H}_{jk}-H_{jk}\right\|_{L_2}+n^{\frac{-2\beta_1}{2\beta_1+1}}+\tau_{nj}\left\{J(H_{jk})-J(\widehat{H}_{jk})\right\}\Bigg].\end{aligned}$$

Taking c_2 large enough such that $Cc_2^{-4\beta_2/(2\beta_2-1)}\leq 1/2$, then we obtain that,

$$\begin{aligned}\left\|H_{jk}-\widehat{H}_{jk}\right\|_{L_2}^2\log^{-2}\left\|H_{jk}-\widehat{H}_{jk}\right\|_{L_2}\leq 2C\Bigg[&c_2^{\frac{4\beta_2}{4\beta_2+1}}\left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta 2+1}}\\ &+(c_1+1)\frac{\log p}{n}+\sqrt{(c_1+1)\frac{\log p}{n}}\left\|H_{jk}-\widehat{H}_{jk}\right\|_{L_2}+h^{2\beta_2}\\ &+n^{-1/2}e^{-p}+n^{\frac{-\beta_1}{2\beta_1+1}}\left\|\widehat{H}_{jk}-H_{jk}\right\|_{L_2}+n^{\frac{-2\beta_1}{2\beta_1+1}}+\tau_{nj}\left\{J(H_{jk})-J(\widehat{H}_{jk})\right\}\Bigg].\end{aligned}$$

Therefore,

$$\left\|H_{jk}(x(t))-\widehat{H}_{jk}(x(t))\right\|_{L_2}^2=O_p\left\{\left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta 2+1}}+h^{2\beta_2}+\frac{\log p}{n}+n^{-\frac{2\beta_1}{2\beta_1+1}}\right\}. \tag{S5}$$

Step 4: Estimation of α_{jk,t_0}. We now apply a similar proof procedure in the above Steps 1-3 to estimate α_{jk,t_0}. First, by Lemma 4 and the Minkowski inequality, as $h=o(1)$, the right-hand-side of (S2) is bounded by

$$\begin{aligned}&\Delta_1+\Delta_2+\Delta_3+\tau_{nj}J(H_{jk})\\ \leq\ & O_p\Bigg\{\|\widehat{\alpha}_{jk,t}-F_{jk}(x(t))\|_{L_2}^2+h^{2\beta_2}+\left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}}+\frac{\log p}{n}\Bigg\}.\end{aligned}$$

Second, by Lemma 5, with probability at least $1-2p^{-c_1}$, the left-hand-side of (S2) is bounded by, for some constant $C>0$,

$$\begin{aligned}
&\widetilde{\Delta}_1+\widetilde{\Delta}_2+\widetilde{\Delta}_3+\mathcal{R}_1+\tau_{nj}J(\widehat{H}_{jk})\\
&\geq \|\widehat{\alpha}_{jk,t}-F_{jk}(x(t))\|_{L_2}^2-\left\{h^{2\beta_2}+\left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}}+\frac{\log p}{n}+n^{-1/2}e^{-p}\right\}\\
&\quad+o_p\left(n^{\frac{-\beta_1}{2\beta_1+1}}\|\widehat{\alpha}_{jk,t}-F_{jk}(x(t))\|_{L_2}\right)+o_p\left(n^{\frac{-2\beta_1}{2\beta_1+1}}\right).
\end{aligned}$$

Then, putting the above two bounds together, we have that

$$\|\widehat{\alpha}_{jk,t}-F_{jk}(x(t))\|_{L_2}^2=O_p\left\{\left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta 2+1}}+h^{2\beta_2}+\frac{\log p}{n}+n^{-\frac{2\beta_1}{2\beta_1+1}}\right\}. \tag{S6}$$

Step 5: Estimation of $F_j(x(t))$. Combining the bounds (S5) and (S6), we obtain that

$$\left\|F_j(x(t))-\widehat{F}_j(x(t))\right\|_{L_2}^2=O_p\left\{\left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta 2+1}}+h^{2\beta_2}+\frac{\log p}{n}+n^{-\frac{2\beta_1}{2\beta_1+1}}\right\}.$$

This leads to the desired upper bound. Letting $h=O\left((n/\log n)^{-1/(2\beta_2+2)}\right)$, then the localized kernel ODE estimator in (12) satisfies that,

$$\left\|F_j(x(t))-\widehat{F}_j(x(t))\right\|_{L_2}^2=O_p\left\{\left(\frac{n}{\log n}\right)^{-\frac{\beta_2}{\beta_2+1}}+\frac{\log p}{n}+n^{-\frac{2\beta_1}{2\beta_1+1}}\right\}.$$

This completes the proof of Theorem 1. ■

B.1.2 Auxiliary lemmas for Theorem 1

For any $g\in\mathcal{H}$, define the norm, $\|g(x(t))\|_n=\sqrt{(1/n)\sum_{i=1}^n g^2(x(t_i))}$.

Lemma 4 *Suppose that $H_{jk}\in\mathcal{H}$, and the errors $\{\epsilon_{ij}\}_{i=1}^n$ are i.i.d. Gaussian. Then there exists some constant $C>0$, such that, for any $c_1>0$ and $c_2>1$, with probability at least $1-2p^{-c_1}$,*

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n R_h(t_i-t_0)\,\epsilon_{ij}H_{jk}(x(t_i))\\
&\leq Ch^{-1}\left\{\|H_{jk}\|_{L_2}^2\log^{-2}\|H_{jk}\|_{L_2}+h^{2\beta_2}+\left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}}\right.\\
&\qquad\left.+\frac{\log p}{n}+\sqrt{\frac{\log p}{n}}\|H_{jk}\|_{L_2}+n^{-1/2}e^{-p}\right\}.
\end{aligned}$$

Proof Recall the RKHS $\mathcal{H}$ as defined in (4). For notational simplicity, we denote $F_{jl} \equiv F_{jll}$, for $l = 1, \ldots, p$. Recall that $F_{jlr} \in \mathcal{H}_l \otimes \mathcal{H}_r$, $l, r \neq k$ and the corresponding reproducing kernel is K_{lr}. Let $\lambda_\nu(K)$ denote the νth largest eigenvalue of a positive definite operator K. Note that $\lambda_\nu(K_{lr}) = O\left\{(\nu \log^{-1} \nu)^{-2\beta_2}\right\}$, for $\nu \geq 1$ (Bach, 2017). By the Sobolev's embedding theorem, $(K_{lr}R_h)(L_2)$ can be embedded to a Sobolev space $\mathcal{W}$ that corresponds to a reproducing kernel K^* (Cucker and Smale, 2002). Moreover, $\lambda_\nu(K^*) = O\{\lambda_\nu(K_{lr})\}$. Let $\{e_\nu : \nu \geq 1\}$ denote the eigenfunctions of K^*; that is, $K^* e_\nu = \lambda_\nu(K^*) e_\nu$ for $\nu \geq 1$. Denote by $\mathcal{E}_\nu$ and $\mathcal{E}_\nu^\perp$ the linear space spanned by $\{e_s : 1 \leq s \leq \nu\}$ and $\{e_s : s \geq \nu + 1\}$, respectively. By the Courant-Fischer-Weyl min-max principle,

$$\lambda_\nu(K_{lr}R_h) \geq \min_{e \in \mathcal{E}_\nu} \|K_{lr}^{1/2} R_h^{1/2} e\|_{L_2}^2 / \|e\|_{L_2}^2 \geq C \min_{e \in \mathcal{E}_\nu} \|(K^*)^{1/2} e\|_{L_2}^2 / \|e\|_{L_2}^2 \geq C \lambda_\nu(K^*)$$

for some constant $C > 0$. On the other hand,

$$\lambda_\nu(K_{lr}R_h) \leq \min_{e \in \mathcal{E}_{\nu-1}^\perp} \|K_{lr}^{1/2} R_h^{1/2} e\|_{L_2}^2 / \|e\|_{L_2}^2 \leq C' \min_{e \in \mathcal{E}_{\nu-1}^\perp} \|(K^*)^{1/2} e\|_{L_2}^2 / \|e\|_{L_2}^2 \leq C' \lambda_\nu(K^*)$$

for some constant $C' > 0$. Henceforth, together with Assumption 1, we have,

$$\lambda_\nu(K_{lr}R_h) = O\{\lambda_\nu(K^*)\} = O\left\{(\nu \log^{-1} \nu)^{-2\beta_2}\right\}. \tag{S7}$$

Since $\{\epsilon_{ij}\}_{i=1}^n$ are i.i.d. Gaussian, by Lemma 2.2 of Yuan and Zhou (2016) and Corollary 8.3 of van de Geer (2000), we have that, for any $c_1 > 0$, with probability at least $1 - p^{-c_1}$,

$$\begin{aligned}
&\frac{1}{n} \sum_{i=1}^n R_h(t_i - t_0) \epsilon_{ij} H_{jk}(x(t_i)) \\
&\leq 2C_1 n^{-1/2} \sum_{l,r=1; l,r \neq k}^p \left(\|F_{jlr}R_h\|_n \log^{-1} \|F_{jlr}R_h\|_n\right)^{1 - \frac{1}{2\beta_2}} \left(\|F_{jlr}R_h\|_{\mathcal{H}} \log^{-1} \|F_{jlr}R_h\|_{\mathcal{H}}\right)^{\frac{1}{2\beta_2}} \\
&\quad + 2C_1 n^{-1/2} \sqrt{(c_1+1)\log p} \sum_{l,r=1; l,r \neq k}^p \|F_{jlr}R_h\|_n + 2C_1 n^{-1/2} e^{-p} \sum_{l,r=1; l,r \neq k}^p \|F_{jlr}R_h\|_{\mathcal{H}} \\
&\equiv 2C_1(\Delta_4 + \Delta_5 + \Delta_6),
\end{aligned} \tag{S8}$$

for some constant C_1. Next, we bound the three terms $\Delta_4, \Delta_5, \Delta_6$ on the right-hand-side of (S8), respectively.

For Δ_4, by the Young's inequality, there exists $c_2 > 1$ such that,

$$\begin{aligned}
\Delta_4 \leq\ & c_2^{-\frac{4\beta_2}{2\beta_2 - 1}} \sum_{l,r=1; l,r \neq k}^p \left(\|F_{jlr}R_h\|_n \log^{-1} \|F_{jlr}R_h\|_n\right)^2 \\
& + c_2^{\frac{4\beta_2}{2\beta_2+1}} n^{-\frac{2\beta_2}{2\beta_2+1}} h^{-\frac{2}{2\beta_2+1}} \sum_{l,r=1; l,r \neq k}^p \left(\|F_{jlr}\|_{\mathcal{H}} \log^{-1} \|F_{jlr}\|_{\mathcal{H}}\right)^{\frac{2}{2\beta_2+1}}.
\end{aligned}$$

Note that

$$\sum_{l,r=1; l,r \neq k}^p \left(\|F_{jlr}\|_{\mathcal{H}} \log^{-1} \|F_{jlr}\|_{\mathcal{H}}\right)^{\frac{2}{2\beta_2+1}} \leq C_2' \sum_{l,r=1; l,r \neq k}^p \left(\|F_{jlr}\|_{\mathcal{H}} \log^{-1} \|F_{jlr}\|_{\mathcal{H}}\right)^0 \leq C_2,$$

for some constants C_2', C_2, where the last step is due to Assumption 2 that the number of nonzero functional components of H_{jk} is bounded. Henceforth,

$$\Delta_4 \leq c_2^{-\frac{4\beta_2}{2\beta_2-1}} \sum_{l,r=1;l,r\neq k}^{p} \left(\|F_{jlr}R_h\|_n \log^{-1}\|F_{jlr}R_h\|_n\right)^2 + c_2^{\frac{4\beta_2}{4\beta_2+1}} n^{-\frac{2\beta_2}{2\beta 2+1}} h^{-\frac{2}{2\beta_2+1}} C_2. \quad \text{(S9)}$$

By (S7), Theorem 4 of Koltchinskii and Yuan (2010) and Theorem 3 of Fan (1993), there exists some constant $C_3 > 0$, such that, with probability at least $1 - p^{-c_1}$,

$$\sum_{l,r=1;l,r\neq k}^{p} \left(\|F_{jlr}R_h\|_n \log^{-1}\|F_{jlr}R_h\|_n\right)^2 \leq 2C_3^2 \sum_{l,r=1;l,r\neq k}^{p} \left(\|F_{jlr}\|_{L_2} \log^{-1}\|F_{jlr}\|_{L_2}\right)^2$$
$$+2C_3^2 h^{2\beta_2} + 2C_3^2 \left\{ \left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}} + \frac{(c_1+1)\log p}{n} \right\} \sum_{l,r=1;l,r\neq k}^{p} \left(\|F_{jlr}\|_{\mathcal{H}} \log^{-1}\|F_{jlr}\|_{\mathcal{H}}\right)^2.$$

Note that there exists some constant $c_3 > 1$, such that

$$\sum_{l,r=1;l,r\neq k}^{p} \left(\|F_{jlr}\|_{L_2} \log^{-1}\|F_{jlr}\|_{L_2}\right)^2 \leq c_3 \left(\|H_{jk}\|_{L_2} \log^{-1}\|H_{jk}\|_{L_2}\right)^2,$$

where we recall that $H_{jk}(x(t)) = \sum_{l,r=1;l,r\neq k}^{p} F_{jlr}(x_l(t), x_r(t))$. Moreover,

$$\sum_{l,r=1;l,r\neq k}^{p} \left(\|F_{jkl}\|_{\mathcal{H}} \log^{-1}\|F_{jkl}\|_{\mathcal{H}}\right)^2 \leq \sum_{l,r=1;l,r\neq k}^{p} \left(\|F_{jlr}\|_{\mathcal{H}} \log^{-1}\|F_{jlr}\|_{\mathcal{H}}\right)^0 \leq C_2.$$

Then, we have

$$\sum_{l,r=1;l,r\neq k}^{p} \left(\|F_{jlr}R_h\|_n \log^{-1}\|F_{jlr}R_h\|_n\right)^2 \leq 2C_3^2 c_3 \left(\|H_{jk}\|_{L_2} \log^{-1}\|H_{jk}\|_{L_2}\right)^2$$
$$+2C_3^2 h^{2\beta_2} + 2C_2 C_3^2 \left\{ \left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}} + \frac{(c_1+1)\log p}{n} \right\}.$$

Inserting into (S9) yields that

$$\Delta_4 \leq 2C_3^2 c_3 c_2^{-\frac{4\beta_2}{2\beta_2-1}} \left(\|H_{jk}\|_{L_2} \log^{-1}\|H_{jk}\|_{L_2}\right)^2 + 2C_3^2 c_2^{-\frac{4\beta_2}{2\beta_2-1}} h^{2\beta_2}$$
$$+ 2C_2 C_3^2 c_2^{-\frac{4\beta_2}{2\beta_2-1}} \left\{ \left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}} + \frac{(c_1+1)\log p}{n} \right\} + C_2 c_2^{\frac{4\beta_2}{4\beta_2+1}} n^{-\frac{2\beta_2}{2\beta 2+1}} h^{-\frac{2}{2\beta_2+1}}.$$

Since $\beta_2 > 1$ and $h = o(1)$, we have $n^{-\frac{2\beta_2}{2\beta 2+1}} h^{-\frac{2}{2\beta_2+1}} < \left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}}$.

For Δ_5, by Theorem 4 of Koltchinskii and Yuan (2010) again, there exists a constant $C_4 > 0$, such that

$$\begin{aligned}
&\sum_{l,r=1;l,r\neq k}^{p} \|F_{jlr}R_h\|_n \\
&\leq C_4 \sum_{l,r=1;l,r\neq k}^{p} \|F_{jlr}\|_{L_2} + C_4 h^{\beta_2} + C_4 \left\{ \left(\frac{nh}{\log n}\right)^{-\frac{\beta_2}{2\beta_2+1}} + \sqrt{\frac{(c_1+1)\log p}{n}} \right\} \sum_{l,r=1;l,r\neq k}^{p} \|F_{jlr}\|_{\mathcal{H}} \\
&\leq C_4 \sum_{l,r=1;l,r\neq k}^{p} \|F_{jlr}\|_{L_2} + C_4 h^{\beta_2} + C_2 C_4 \left\{ \left(\frac{nh}{\log n}\right)^{-\frac{\beta_2}{2\beta_2+1}} + \sqrt{\frac{(c_1+1)\log p}{n}} \right\}.
\end{aligned}$$

Define the set $\mathcal{Q}_1 \equiv \left\{ l, r = 1, \ldots, p; l, r \neq k : \|F_{jlr}\|_{L_2} > \sqrt{n^{-1}\log p} \right\}$. By the Cauchy-Schwartz inequality, we have,

$$\begin{aligned}
\sum_{l,r\in\mathcal{Q}_1} \|F_{jlr}\|_{L_2} &\leq \mathrm{card}^{1/2}(\mathcal{Q}_1) \cdot \left(\sum_{l,r\in\mathcal{Q}_1} \|F_{jlr}\|_{L_2}^2 \right)^{1/2} \\
&\leq \sum_{l,r=1;l,r\neq k}^{p} \|F_{jlr}\|_{\mathcal{H}}^0 \cdot \left(\sum_{l,r=1;l,r\neq k}^{p} \|F_{jlr}\|_{L_2}^2 \right)^{1/2} \leq C_2 c_4 \|H_{jk}\|_{L_2},
\end{aligned}$$

The constant $c_4 > 1$ satisfies that $\sum_{l,r=1;l,r\neq k}^{p} \|F_{jlr}\|_{L_2}^2 \leq c_4^2 \|H_{jk}\|_{L_2}^2$, where we recall that $H_{jk}(x(t)) = \sum_{l,r=1;l,r\neq k}^{p} F_{jlr}(x_l(t), x_r(t))$. Next, define the set $\mathcal{Q}_2 \equiv \{l, r = 1, \ldots, p; l, r \neq k : \|F_{jkl}\|_{L_2} \leq \sqrt{n^{-1}\log p}\}$. By definition,

$$\sum_{l,r\in\mathcal{Q}_2} \|F_{jlr}\|_{L_2} \leq \sum_{l,r\in\mathcal{Q}_2} \|F_{jlr}\|_{L_2}^0 \sqrt{\frac{\log p}{n}} \leq \sqrt{\frac{\log p}{n}} \sum_{l,r=1;l,r\neq k}^{p} \|F_{jlr}R_h\|_{L_2}^0 \leq C_2 \sqrt{\frac{\log p}{n}}.$$

Combining $\mathcal{Q}_1$ and $\mathcal{Q}_2$ gives that,

$$\sum_{l,r=1;l,r\neq k}^{p} \|F_{jlr}\|_{L_2} \leq \sum_{l,r\in\mathcal{Q}_1} \|F_{jlr}\|_{L_2} + \sum_{l,r\in\mathcal{Q}_2} \|F_{jlr}\|_{L_2} \leq C_2 c_4 \|H_{jk}\|_{L_2} + C_2 \sqrt{\frac{\log p}{n}}.$$

Henceforth, we can bound Δ_5 as,

$$\Delta_5 \leq \sqrt{(c_1+1)} C_4 \left[C_2 c_4 \sqrt{\frac{\log p}{n}} \|H_{jk}\|_{L_2} + C_2 h^{\beta_2} + C_2 \left(\frac{nh}{\log n}\right)^{-\frac{\beta_2}{2\beta_2+1}} \sqrt{\frac{\log p}{n}} + C_2 \frac{\log p}{n} \right].$$

For Δ_6, it can be bounded as,

$$\Delta_6 \leq n^{-1/2} e^{-p} \sum_{l,r=1;l,r\neq k}^{p} \|F_{jlr}\|_{\mathcal{H}}^0 \leq C_2 n^{-1/2} e^{-p}.$$

Combining the bounds for $\Delta_4, \Delta_5, \Delta_6$, and applying the Cauchy-Schwarz inequality completes the proof of Lemma 4. ■

Lemma 5 *Suppose that $H_{jk} \in \mathcal{H}$. Then there exists some constant $C > 0$, such that, for any $c_1 > 0$ and $c_2 > 1$, with probability at least $1 - 2p^{-c_1}$,*

$$\|H_{jk}\|_{L_2}^2 \leq \|H_{jk}R_h^{1/2}\|_n^2 + C\Big\{c_2^{-\frac{4\beta_2}{2\beta_2-1}}\|H_{jk}\|_{L_2}^2 \log^{-2}\|H_{jk}\|_{L_2} + h^{2\beta_2} + c_2^{\frac{4\beta_2}{2\beta_2-1}}\left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta 2+1}} + \sqrt{c_1+1}\frac{\log p}{n} + \sqrt{(c_1+1)\frac{\log p}{n}}(\|H_{jk}\|_{L_2} + h^{\beta_2}) + n^{-1/2}e^{-p}\Big\}.$$

Proof Note that

$$\|H_{jk}\|_{L_2}^2 - \|H_{jk}R_h^{1/2}\|_n^2 \leq \sup_{\substack{g\in\mathcal{H},\|g\|_{\mathcal{H}}^0\leq\|H_{jk}\|_{\mathcal{H}}^0 \\ \|g\|_{L_2}\leq\|H_{jk}\|_{L_2}}} \left(\|g\|_{L_2}^2 - \|gR_h^{1/2}\|_n^2\right). \tag{S10}$$

By the Talagrand's concentration inequality (Talagrand, 1996), with probability at least $1 - e^{-c_1}$, we have that,

$$\begin{aligned} &\sup_{\substack{g\in\mathcal{H},\|g\|_{\mathcal{H}}^0\leq\|H_{jk}\|_{\mathcal{H}}^0 \\ \|g\|_{L_2}\leq\|H_{jk}\|_{L_2}}} \left(\|g\|_{L_2}^2 - \|gR_h^{1/2}\|_n^2\right) \\ &\leq 2\mathbb{E}\sup_{\substack{g\in\mathcal{H},\|g\|_{\mathcal{H}}^0\leq\|H_{jk}\|_{\mathcal{H}}^0 \\ \|g\|_{L_2}\leq\|H_{jk}\|_{L_2}}} \left(\|g\|_{L_2}^2 - \|gR_h^{1/2}\|_n^2\right) + 4\|H_{jk}\|_{L_2}\sqrt{\frac{c_1}{n}} + \frac{16c_1}{n}. \end{aligned} \tag{S11}$$

By the symmetrization inequality for the Rademacher process (van der Vaart and Wellner, 1996) and Theorem 3 of Fan (1993), there exists a constant $C_1 > 0$, such that,

$$\begin{aligned} &\mathbb{E}\sup_{\substack{g\in\mathcal{H},\|g\|_{\mathcal{H}}^0\leq\|H_{jk}\|_{\mathcal{H}}^0 \\ \|g\|_{L_2}\leq\|H_{jk}\|_{L_2}}} \left(\|g\|_{L_2}^2 - \|gR_h^{1/2}\|_n^2\right) \\ &\leq \mathbb{E}\sup_{\substack{g\in\mathcal{H},\|g\|_{\mathcal{H}}^0\leq\|H_{jk}\|_{\mathcal{H}}^0 \\ \|g\|_{L_2}\leq\|H_{jk}\|_{L_2}}} \left\{\frac{1}{n}\sum_{i=1}^n \omega_i R_h(t_i)g^2(x(t_i))\right\} + C_1h^{2\beta_2} \\ &\leq C_1\mathbb{E}\sup_{\substack{g\in\mathcal{H},\|g\|_{\mathcal{H}}^0\leq\|H_{jk}\|_{\mathcal{H}}^0 \\ \|g\|_{L_2}\leq\|H_{jk}\|_{L_2}}} \left\{\frac{1}{n}\sum_{i=1}^n \omega_i R_h(t_i)g(x(t_i))\right\} + Ch^{2\beta_2}, \end{aligned} \tag{S12}$$

where $\omega, \ldots, \omega_n$ are independent random variables drawn from the Rademacher distribution; i.e., $\mathbb{P}(\omega_i = 1) = \mathbb{P}(\omega_i = -1) = 1/2$, for $i = 1, \ldots, n$. The last inequality in (S12) is due to the contraction inequality, and the fact that g^2 is a Lipschitz function. Henceforth,

with Talagrand's concentration inequality, there exists a constant $C_2 > 0$, such that, with probability at least $1 - e^{-c_1}$,

$$\begin{aligned}
&\mathbb{E} \sup_{\substack{g\in\mathcal{H},\|g\|_{\mathcal{H}}^0\le\|H_{jk}\|_{\mathcal{H}}^0 \\ \|g\|_{L_2}\le\|H_{jk}\|_{L_2}}} \left\{\frac{1}{n}\sum_{i=1}^n \omega_i R_h(t_i) g(x(t_i))\right\} \\
&\le C_2 \left[\sup_{\substack{g\in\mathcal{H},\|g\|_{\mathcal{H}}^0\le\|H_{jk}\|_{\mathcal{H}}^0 \\ \|g\|_{L_2}\le\|H_{jk}\|_{L_2}}} \sum_{l,r=1}^p \left\{\frac{1}{n}\sum_{i=1}^n \omega_i R_h(t_i) g_{lr}(x(t_i))\right\} + c_1 h^{\beta_2} + \|H_{jk}\|_{L_2}\sqrt{\frac{c_1}{n}} + \frac{c_1}{n}\right].
\end{aligned} \tag{S13}$$

By Lemma 2.2 of Yuan and Zhou (2016), and the result that the νth eigenvalue of the RKHS $\mathcal{H}$ is of order $(\nu \log^{-1}\nu)^{-2\beta_2}$, for $\nu \ge 1$ (Bach, 2017), there exists a constant $C_3 > 0$, such that, with probability at least $1 - d^{-c_1}$,

$$\begin{aligned}
&\sum_{l,r=1}^p \left\{\frac{1}{n}\sum_{i=1}^n \omega_i R_h(t_i) g_{lr}(x(t_i))\right\} \\
&\le C_3 n^{-1/2} \sum_{l,r=1}^p \Big\{ \left(\|g_{lr}R_h\|_{\mathcal{H}} \log^{-1}\|g_{lr}R_h\|_{\mathcal{H}}\right)^{\frac{1}{2\beta_2}} \left(\|g_{lr}R_h\|_{L_2}\log^{-1}\|g_{lr}R_h\|_{L_2}\right)^{1-\frac{1}{2\beta_2}} \\
&\quad + \|g_{lr}R_h\|_{L_2}\sqrt{(c_1+1)\log p} + e^{-p}\|g_{lr}R_h\|_{\mathcal{H}}\Big\}.
\end{aligned}$$

Following the arguments for bounding Δ_4 in (S8), there exists a constant $C_4 > 0$ and for any $c_2 > 1$, such that,

$$\begin{aligned}
&n^{-1/2} \sup_{\substack{g\in\mathcal{H},\|g\|_{\mathcal{H}}^0\le\|H_{jk}\|_{\mathcal{H}}^0 \\ \|g\|_{L_2}\le\|H_{jk}\|_{L_2}}} \sum_{l,r=1}^p \left(\|g_{lr}R_h\|_{\mathcal{H}}\log^{-1}\|g_{lr}R_h\|_{\mathcal{H}}\right)^{\frac{1}{2\beta_2}} \left(\|g_{lr}R_h\|_{L_2}\log^{-1}\|g_{lr}R_h\|_{L_2}\right)^{1-\frac{1}{2\beta_2}} \\
&\le C_4 \sup_{\substack{g\in\mathcal{H},\|g\|_{\mathcal{H}}^0\le\|H_{jk}\|_{\mathcal{H}}^0 \\ \|g\|_{L_2}\le\|F_j\|_{L_2}}} \sum_{l,r=1}^p \left(\|g_{lr}\|_{L_2}\log^{-1}\|g_{lr}\|_{L_2}\right)^2 + C_4 h^{2m} + C_4\left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta 2+1}} + C_4\frac{\log p}{n} \\
&\le C_4 C_5 \|H_{jk}\|_{L_2}^2 \log^{-2}\|H_{jk}\|_{L_2} + C_4 h^{2m} + C_4\left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta 2+1}} + C_4\frac{\log p}{n}.
\end{aligned}$$

Here the last step is due to $\sum_{l,r=1}^p \left(\|g_{lr}\|_{L_2}\log^{-1}\|g_{lr}\|_{L_2}\right)^2 \le C_5\left(\|H_{jk}\|_{L_2}\log^{-1}\|H_{jk}\|_{L_2}\right)^2$ for some constant $C_5 > 1$. Following the arguments for bounding Δ_5 in (S8), and by Theorem 3 of Fan (1993), there exists a constant $C_6 > 0$, such that,

$$\sum_{l,r=1}^p \|g_{lr}R_h\|_{L_2} \le C_6\left\{\sqrt{\frac{\log p}{n}} + \|H_{jk}\|_{L_2} + h^{\beta_2}\right\}.$$

Henceforth, for some constant $C_7 > 0$,

$$\begin{aligned}
&\sup_{\substack{g\in\mathcal{H},\|g\|_{\mathcal{H}}^0\le\|H_{jk}\|_{\mathcal{H}}^0\\ \|g\|_{L_2}\le\|H_{jk}\|_{L_2}}} \sum_{l,r=1}^{p}\left\{\frac{1}{n}\sum_{i=1}^{n}\omega_i R_h(t_i)g_{lr}(x(t_i))\right\}\\
&\le C_7\left\{c_2^{-\frac{4\beta_2}{2\beta_2-1}}\|H_{jk}\|_{L_2}^2\log^{-2}\|H_{jk}\|_{L_2}+h^{2\beta_2}+c_2^{\frac{4\beta_2}{4\beta_2+1}}\left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta 2+1}}\right\}\\
&\quad+C_7\sqrt{\frac{(c_1+1)\log p}{n}}\left\{\sqrt{\frac{\log p}{n}}+\|H_{jk}\|_{L_2}+h^{\beta_2}\right\}+C_7 n^{-1/2}e^{-p}.
\end{aligned}$$

Together with (S11), (S12), and (S13), we have, with probability at least $1-2e^{-c_1}$,

$$\begin{aligned}
&\sup_{\substack{g\in\mathcal{H},\|g\|_{\mathcal{H}}^0\le\|H_{jk}\|_{\mathcal{H}}^0\\ \|g\|_{L_2}\le\|H_{jk}\|_{L_2}}} \left(\|g\|_{L_2}^2-\|gR_h^{1/2}\|_n^2\right)\\
&\le C_8\left\{c_2^{-\frac{4\beta_2}{2\beta_2-1}}\|H_{jk}\|_{L_2}^2\log^{-2}\|H_{jk}\|_{L_2}+h^{2\beta_2}+c_2^{\frac{4\beta_2}{4\beta_2+1}}\left(\frac{nh}{\log n}\right)^{-\frac{2\beta_2}{2\beta 2+1}}\right\} \qquad \text{(S14)}\\
&\quad+C_8\sqrt{\frac{(c_1+1)\log p}{n}}\left\{\sqrt{\frac{\log p}{n}}+\|H_{jk}\|_{L_2}+h^{\beta_2}\right\}+C_8 n^{-1/2}e^{-p},
\end{aligned}$$

for some constant $C_8 > 0$. Combining (S10) and (S14) completes the proof of Lemma 5. ■

B.2 Proof of Theorem 2

Proof Define the empirical process,

$$\widetilde{Z}_n(t_0) \equiv \sqrt{nh}\cdot\widehat{\sigma}_n^{-1}(t_0)\left[\widehat{F}_{jk}(x_k(t_0))-F_{jk}(x_k(t_0))\right],\quad \forall t_0\in\mathcal{T}.$$

Define $\widetilde{V}^Z \equiv \sup_{t_0\in\mathcal{T}}\widetilde{Z}_n(t_0)$. We divide the proof of this theorem into four steps.

Step 1. We aim to prove the following statement: There exists a Gaussian process $\widetilde{\mathbb{H}}_n(t_0)$, such that $\mathbb{E}[\sup_{t_0\in\mathcal{T}}\widetilde{\mathbb{H}}_n(t_0)]\le C\sqrt{\log n}$, for some constant $C>0$, and a sequence of random variables W_n^0, such that $W_n^0=\sup_{t_0\in\mathcal{T}}\widetilde{\mathbb{H}}_n(t_0)$ and $\mathbb{P}\left(|W_n^0-\widetilde{V}^Z|>\epsilon_{1n}\right)<\delta_{1n}$, for some $(\epsilon_{1n},\delta_{1n})\to 0$, as $n\to\infty$.

We construct the Gaussian process $\widetilde{\mathbb{H}}_n(t_0)$ as

$$\widetilde{\mathbb{H}}_n(t_0)=\frac{1}{\sqrt{nh^{-1}}}\sum_{i=1}^{n}\epsilon_i\frac{R_h(t_i-t_0)R_{t_0,i\cdot}^{\top}\Sigma_{k\cdot}}{\widehat{\sigma}_n(t_0)},$$

where ϵ_i is the error term in (2). Then $\widetilde{\mathbb{H}}_n(t_0)$ is a Gaussian variable conditional on $\{t_1,\ldots,t_n\}$. By the Jensen's inequality, there exists some constant $C>0$, such that,

$$\exp\left[s\mathbb{E}(W_n^0)\right]\le\mathbb{E}\exp\left(sW_n^0\right)=\mathbb{E}\left\{\sup_{t_0\in\mathcal{T}}\exp\left[s\widetilde{\mathbb{H}}_n(t_0)\right]\right\}\le n\exp\left(Cs^2\right),$$

for $s > 0$, where the last inequality follows from the definition of the Gaussian moment generating function. Rewriting this inequality, we have $\mathbb{E}(W_n^0) \leq \log n/s + Cs$. Setting $s = \sqrt{\log n/C}$, we obtain that,

$$\mathbb{E}\left[\sup_{t_0 \in \mathcal{T}} \widetilde{\mathbb{H}}_n(t_0)\right] \leq \sqrt{C \log n}.$$

By the Cauchy-Schwarz inequality, there exists a constant $c > 0$, such that,

$$\widetilde{Z}_n(t_0) - \widetilde{\mathbb{H}}_n(t_0) \leq O\left(\sqrt{N}\{\mathbb{E}[\delta_0^2(X)]\}^{1/2}\right) + O_p(N^{-c}). \tag{S15}$$

Define $V_n^0 \equiv \sup_{t_0 \in \mathcal{T}} \widetilde{\mathbb{H}}_n(t_0)$. Recall that $\widetilde{V}^Z = \sup_{t_0 \in \mathcal{T}} \widetilde{Z}_n(t_0)$. Then by (S15), there exists some constant $C > 0$, such that,

$$\mathbb{P}\left(\left|V_N^0 - \widetilde{V}^Z\right| > CN^{-c}\right) \leq \mathbb{P}\left(\sup_{x \in \mathcal{X}^p} \left|\widetilde{\mathbb{H}}_N(x) - \widetilde{Z}_N(x)\right| > CN^{-c}\right) \leq N^{-1}. \tag{S16}$$

Setting $\epsilon_{1N} = CN^{-c}$, $\delta_{1N} = N^{-1}$, and $W_N^0 \stackrel{d}{=} V_N^0$ completes the proof of Step 1.

Step 2. We aim to prove the following anti-concentration inequality for any $\epsilon > 0$,

$$\sup_{s \in \mathbb{R}} \mathbb{P}\left[\left|\sup_{t_0 \in \mathcal{T}} \left|\widetilde{\mathbb{H}}_n(t_0)\right| - s\right| \leq \bar{\epsilon}\right] \leq C\bar{\epsilon}\sqrt{\log N}.$$

This is true due to the result of Step 1 and Corollary 2.1 of Chernozhukov et al. (2014).

Step 3.We aim to prove the following statement: Letting $c_n(\alpha)$ and $\widehat{c}_n(\alpha)$ be the $(1-\alpha)$-quantiles of $\widetilde{V}^Z$ and V_n^0, respectively, there exist $\tau_n, \epsilon_n, \delta_n > 0$, such that,

$$\mathbb{P}\big[\widehat{c}_n(\alpha) < c_n(\alpha + \tau_n) - \epsilon_n\big] \leq \delta_n, \quad \mathbb{P}\big[\widehat{c}_n(\alpha) > c_n(\alpha - \tau_n) + \epsilon_n\big] \leq \delta_n,$$

and $(\tau_n, \epsilon_n, \delta_n) \to 0$ as $n \to \infty$.

Recall the Gaussian multiplier process $\widehat{\mathbb{H}}_n(t_0)$ in Section 4.1, which is defined as,

$$\widehat{\mathbb{H}}_n(t_0) \equiv \frac{1}{\sqrt{nh^{-1}}} \sum_{i=1}^n \xi_i \cdot \frac{\widehat{\sigma}_j R_h(t_i - t_0) R_{t_0,i\cdot}^\top \Sigma_{k\cdot}}{\widehat{\sigma}_n(t_0)}, \quad \text{for any } t_0 \in \mathcal{T},$$

where $\xi_1, \ldots, \xi_n$ are independent standard normal variables. We consider the process,

$$\widehat{\mathbb{H}}_n^{(1)}(t_0) = \frac{1}{\sqrt{nh^{-1}}} \sum_{i=1}^n \xi_i \cdot \frac{\sigma_j R_h(t_i - t_0) R_{t_0,i\cdot}^\top \Sigma_{k\cdot}}{\widehat{\sigma}_n(t_0)}.$$

Let $\widehat{V}_n = \sup_{t_0 \in \mathcal{T}} \widehat{\mathbb{H}}_n(t_0)$, and $\widehat{V}_n^{(1)} = \sup_{t_0 \in \mathcal{T}} \widehat{\mathbb{H}}_n^{(1)}(t_0)$. Denote $\Delta\mathbb{H}^{(1)}(t_0) = \widehat{\mathbb{H}}_n^{(1)}(t_0) - \widehat{\mathbb{H}}_n(t_0)$. By the triangle inequality,

$$\sup_{t_0 \in \mathcal{T}} \left|\Delta\mathbb{H}^{(1)}(t_0)\right| \leq \left|\sigma_j - \widehat{\sigma}_j\right| \sup_{t_0 \in \mathcal{T}} \sqrt{h} \cdot \widehat{\sigma}_n^{-1}(t_0) \left[\sup_{t_0 \in \mathcal{T}} \mathcal{I}_1^{\mathbb{H}}(t_0) + \sup_{t_0 \in \mathcal{T}} \mathcal{I}_2^{\mathbb{H}}(t_0)\right], \tag{S17}$$

where

$$\mathcal{I}_1^{\mathbb{H}}(t_0) = \frac{1}{n}\sum_{i=1}^{n} R_h(t_i - t_0)\left|R_{t_0,i\cdot}^{\top}\Sigma_{k\cdot} - F_j(x(t_0))\right|$$

$$\mathcal{I}_2^{\mathbb{H}}(t_0) = \frac{1}{n}\sum_{i=1}^{n} R_h(t_i - t_0)F_j(x(t_0)).$$

For $\mathcal{I}_1^{\mathbb{H}}(t_0)$, by the Cauchy-Schwarz inequality, we have that,

$$\begin{aligned}\sup_{t_0\in\mathcal{T}}|\mathcal{I}_1^{\mathbb{H}}(t_0)| &\le \sup_{t_0\in\mathcal{T}}\left(\frac{1}{n}\sum_{i=1}^{n} R_h(t_i - t_0)(\Psi_{i\cdot}^{\top}(\widehat{\theta}_z - \theta_z))^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n} K_h(X_{i1} - z)\right)^{1/2}\\ &\le C\inf_{t_0\in\mathcal{T}}\left(\frac{\log(D_n/p_1(z))}{p_1^2(z)}\right)^{1/2}\cdot\sqrt{m}\left(\sqrt{\frac{m\log(dm)}{nh}} + \frac{m}{nh} + \sqrt{\frac{\log(1/h)}{nh}}\right)^{1/2}\\ &= O(r_{nj}^{1/2}).\end{aligned}$$

For $\mathcal{I}_2^{\mathbb{H}}(t_0)$, we have that,

$$\begin{aligned}\sup_{t_0\in\mathcal{T}}|\mathcal{I}_2^{\mathbb{H}}(t_0)| &\le \sup_{t_0\in\mathcal{T}}\frac{1}{n}\|\Psi^{\top}W_z\mathbf{1}\|_{2,\infty}\|\theta_z\|_1\\ &\le \sup_{t_0\in\mathcal{T}}\frac{1}{n}\|W_z^{1/2}\Psi_{\cdot j}\Psi_{\cdot j}^{\top}W_z^{1/2}\|_2\|W_z^{1/2}\mathbf{1}\|_2\|\theta_z\|_1\\ &\le \frac{C}{m}\cdot\sup_{t_0\in\mathcal{T}}\sqrt{p(z)}\cdot\sqrt{m} = O(1/\sqrt{m}).\end{aligned}$$

Therefore, we have that,

$$\mathbb{P}\left(\left|\widehat{V}_n - \widehat{V}_n^{(1)}\right| > C\sqrt{r_n}m^{1/4}\right) \le n^{-1}.$$

We next bound $|\widehat{\sigma}_j - \sigma_j|$, for $j = 1, \ldots, p$. We start with an upper bound on $|\widehat{\sigma}_j^2 - \sigma_j^2|$. Let $\widehat{\epsilon}_{ij} = y_{ij} - \int_0^{t_i}\widehat{F}_j(\widehat{x}(t))dt$, $i = 1, \ldots, n$. Using the triangle inequality, we have that,

$$\left|\widehat{\sigma}_j^2 - \sigma_j^2\right| \le \left|\frac{1}{n}\sum_{i=1}^{n}\widehat{\epsilon}_{ij}^2 - \frac{1}{n}\sum_{i=1}^{n}\epsilon_{ij}^2\right| + \left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_{ij}^2 - \sigma_j^2\right| \equiv I_1 + I_2. \tag{S18}$$

To bound I_1, we have that,

$$\begin{aligned}I_1 &\le \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{\epsilon}_{ij}^2 - \epsilon_{ij}^2\right| \le \frac{1}{n}\sum_{i=1}^{n}|\widehat{\epsilon}_{ij} - \epsilon_{ij}|\left(|\widehat{\epsilon}_{ij} - \epsilon_{ij}| + 2|\epsilon_{ij}|\right)\\ &\le \left(\frac{1}{n}\sum_{i=1}^{n}[\widehat{\epsilon}_{ij} - \epsilon_{ij}]^2\right)^{1/2}\times\left[2\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_{ij}^2\right)^{1/2} + \left\{\frac{1}{n}\sum_{i=1}^{n}[\widehat{\epsilon}_{ij} - \epsilon_{ij}]^2\right\}^{1/2}\right].\end{aligned}$$

Since $\mathbb{E}(\epsilon_{ij}^2) = \sigma_j^2 < \infty$, we have that,

$$\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_{ij}^2\right)^{1/2} = O_p(1).$$

Let $r_{nj} \equiv \tau_{nj}^{-\frac{1}{2\beta_2}} \frac{\log n}{nh} + \tau_{nj} + h^{2\beta_2} + \frac{\log p}{n} + n^{-\frac{2\beta_1}{2\beta_1+1}}$. By Theorem 1, we have $n^{-1}\sum_{i=1}^{n}[\widehat{\epsilon}_{ij} - \epsilon_{ij}]^2 = O_p(r_{nj})$. Therefore,

$$I_1 \leq O_p(r_{nj}^{1/2}). \tag{S19}$$

To bound I_2, we have that,

$$\mathbb{E}(I_2^2) \leq n^{-1}\mathbb{E}(\epsilon_{ij}^4) = O(n^{-1}),$$

where the last step is due to that ϵ_{ij} is a normal random variable and hence ϵ_{ij} has a bounded fourth moment. Therefore,

$$I_2 \leq O_p(n^{-1/2}). \tag{S20}$$

Combining (S18), (S19), and (S20), we have that,

$$\widehat{\sigma}_j - \sigma_j = O\left(|\widehat{\sigma}_j^2 - \sigma_j^2|\right) \leq O_p(r_{nj}^{1/2}).$$

By (S17), we have that,

$$\sup_{t_0 \in \mathcal{T}} |\Delta\mathbb{H}^{(1)}(t_0)| \leq O_p(r_{nj}^{1/2})$$

Then there exists a constant $C > 0$, such that,

$$\mathbb{P}\left(\left|\widehat{V}_n - \widehat{V}_n^{(1)}\right| > Cr_{nj}^{1/2}\right) \leq \mathbb{P}\left(\sup_{t_0 \in \mathcal{T}} |\Delta\mathbb{H}^{(1)}(t_0)| > Cr_{nj}^{1/2}\right) \leq n^{-1}. \tag{S21}$$

Since $\sigma\xi \stackrel{d}{=} (\epsilon_{1j}, \ldots, \epsilon_{nj})^\top$, we have $\sup_{t_0 \in \mathcal{T}} \widehat{\mathbb{H}}_n^{(1)}(t_0) \stackrel{d}{=} \sup_{t_0 \in \mathcal{T}} \widetilde{\mathbb{H}}_n(t_0)$. That is, $\widehat{V}_n^{(1)} \stackrel{d}{=} V_n^0$. Combining (S16) with (S21), we have that,

$$\mathbb{P}\left(\left|\widehat{V}_n - \widetilde{V}_n^Z\right| > Cn^{-c}\right) \leq n^{-1}.$$

Therefore, by the definition of $\widehat{c}_N(\alpha)$,

$$\mathbb{P}\left(\widetilde{V}_n^Z \leq \widehat{c}_n(\alpha) + Cn^{-c}\right) \geq \mathbb{P}\left(\widehat{V}_n \leq \widehat{c}_n(\alpha)\right) - \mathbb{P}\left(\left|\widehat{V}_n - \widetilde{V}_n^Z\right| > Cn^{-c}\right) \geq 1 - \alpha - n^{-1},$$

which implies that the estimated quantile is lower bounded as,

$$\widehat{c}_n(\alpha) \geq c_n(\alpha + n^{-1}) - Cn^{-c}, \quad \text{for some } c \in (0, c_{\min}].$$

Similarly, we also have $\widehat{c}_n(\alpha) \leq c_n(\alpha - n^{-1}) + Cn^{-c}$. Setting $\tau_n = n^{-1}$, $\epsilon_n = Cn^{-c}$, and $\delta_n = n^{-1}$ completes the proof of Step 3.

Step 4. By verifying the statements in Steps 1 to 3, we now apply Corollary 3.1 of Chernozhukov et al. (2014) and obtain that,

$$\mathbb{P}\left(F_{jk}(x_k(t_0)) \in \mathcal{C}_{n,\alpha},\ \forall t_0 \in \mathcal{T}\right) \geq 1 - \alpha - Cn^{-c}.$$

Therefore, the confidence interval $\mathrm{CI}(f_0(x))$ is asymptotic honest. This completes the proof of Theorem 2. ■

B.3 Proof of Theorem 3

We divide the proof of this theorem into two parts. We first present the main proof in Section B.3.1, then give an auxiliary lemma useful for the proof of this theorem in Section B.3.2.

B.3.1 Main proof

Proof We use the primal-dual witness method to prove that the localized kernel ODE approach selects all significant variables, and includes no insignificant ones. Recall that, by the representer theorem (Wahba, 1990), the selection problem becomes (18); i.e.,

$$\min_{\theta_j}\left\{\frac{1}{n}(z_j - G\theta_j)^\top R_{t_0}(z_j - G\theta_j) + \kappa_{nj}\left(\sum_{k'=1,k'\neq k}^{p}\theta_{jk'} + \sum_{k'=1,k'\neq k}^{p}\sum_{l=1,l\neq k',k}^{p}\theta_{jk'l}\right)\right\}, \tag{S22}$$

subject to $\theta_{jk'} \geq 0, \theta_{jk'l} \geq 0$, where the "response" is $z_j = (y_j - \bar{y}_j) - \widehat{\alpha}_{jk,t_0}\bar{t} - (1/2)n\eta_{nj}c_j$, and the "predictor" is $G \in \mathbb{R}^{n\times(p-1)^2}$. The vector θ_j solves (S22), if it satisfies the Karush-Kuhn-Tucker (KKT) condition,

$$\frac{2}{n}G^\top R_{t_0}(G\theta_j - z_j) + \kappa_{nj}g_j = 0, \quad j = 1,\ldots,p, \tag{S23}$$

where G contains errors in the variables due to the estimated $\widehat{x}(t)$, and

$$g_j = \text{sign}(\theta_j), \text{ if } \theta_j \neq 0, \quad \text{and} \quad |g_j| \leq 1, \text{ otherwise.} \tag{S24}$$

To apply the primal-dual witness method, we next construct an oracle primal-dual pair $(\widehat{\theta}_j, \widehat{g}_j)$ satisfying the KKT conditions (S23) and (S24). Specifically,

(a) We set $\widehat{\theta}_{jk'l} = 0$ for $(k',l) \notin S_j^*$, where S_j^* is as defined in Section 5.2, and $s_j = \text{card}(S_j^*)$.

(b) Let $\widehat{\theta}_{S_j^*}$ be the minimizer of the partial penalized likelihood,

$$(z_j - G_{S_j^*}\theta_{S_j^*})^\top R_{t_0}(z_j - G_{S_j^*}\theta_{S_j^*}) + n\kappa_{nj}\left(\sum_{k'=1,k'\neq k}^{p}\theta_{jk'} + \sum_{k'=1,k'\neq k,l}^{p}\sum_{l=1,l\neq k}^{p}\theta_{jk'l}\right). \tag{S25}$$

(c) Let S_j^c be the complement of S_j^* in $\{(k',l) : k',l = 1,\ldots,k-1,k+1,\ldots,p\}$. We obtain $\widehat{g}_{S_j^c}$ from (S23) by substituting in the values of $\widehat{\theta}_j$ and $\widehat{g}_{S_j^*}$.

Next, we verify the support recovery consistency; i.e.,

$$\max_{(k,l)\in S_j^*}\|\widehat{\theta}_{jkl} - \theta_{jkl}\|_{\ell_2} \leq \frac{2}{3}\theta_{\min},$$

which in turn implies that the oracle estimator $\widehat{\theta}_j$ recovers the support of θ_j exactly.

Note that the subgradient condition for the partial penalized likelihood (S25) is

$$2G_{S_j^*}^\top R_{t_0}(G_{S_j^*}\widehat{\theta}_{S_j^*} - z_j) + n\kappa_{nj}\widehat{g}_{S_j^*} = 0,$$

which implies that

$$2G_{S_j^*}^\top R_{t_0}(G_{S_j^*}\widehat{\theta}_{S_j^*} - G_{S_j^*}\theta_{S_j^*}) + 2G_{S_j^*}^\top R_{t_0}(G_{S_j^*}\theta_{S_j^*} - z_j) + n\kappa_{nj}\widehat{g}_{S_j^*} = 0.$$

Define $\mathcal{R}_{S_j^*} \equiv 2G_{S_j^*}^\top R_{t_0} G_{S_j^*}\theta_{S_j^*} - 2G_{S_j^*}^\top R_{t_0} z_j$. Then,

$$\widehat{\theta}_{S_j^*} - \theta_{S_j^*} = -\left(2G_{S_j^*}^\top R_{t_0} G_{S_j^*}\right)^{-1}(\mathcal{R}_{S_j^*} + n\kappa_{nj}\widehat{g}_{S_j^*}). \tag{S26}$$

For each (k,l), denote the corresponding column of G by G_{kl}. Then for $(k,l) \in S_j^*$,

$$\mathcal{R}_{kl} = 2G_{kl}^\top R_{t_0} G_{S_j^*}\theta_{S_j^*} - 2G_{kl}^\top R_{t_0} z_j. \tag{S27}$$

By Lemma 6, we have $\|\mathcal{R}_{kl}\|_{\ell_2} \le \eta_{\mathcal{R}}$ for any $(k,l) \in S_j^*$. Then,

$$\|\mathcal{R}_{S_j^*}\|_{\ell_2} \le \eta_{\mathcal{R}}\sqrt{s_j}. \tag{S28}$$

By Assumption 5, we have that $\Lambda_{\min}\left(G_{S_j^*}^\top R_{t_0} G_{S_j^*}\right) \ge C_{\min}/2$, for some constant $C_{\min} > 0$. Henceforth,

$$\Lambda_{\max}\left\{\left(2G_{S_j^*}^\top R_{t_0} G_{S_j^*}\right)^{-1}\right\} \le \frac{1}{C_{\min}}.$$

Note that for any $(k,l) \in S_j^*$, $\|\widehat{g}_{jkl}\|_{\ell_2} \le 1$, which implies that,

$$\|\widehat{g}_{S_j^*}\|_{\ell_2} \le \sqrt{s_j}. \tag{S29}$$

Therefore, we have that,

$$\max_{(k,l)\in S_j^*} \|\widehat{\theta}_{jkl} - \theta_{jkl}\|_{\ell_2} \le \|\widehat{\theta}_{S_j^*} - \theta_{S_j^*}\|_{\ell_2} \le \frac{\eta_{\mathcal{R}}\sqrt{s_j}}{C_{\min}} + n\kappa_{nj}\frac{\sqrt{s_j}}{C_{\min}} \le \frac{2}{3}\theta_{\min}.$$

where the last inequality is due to Assumption 6.

Next, we verify the strict dual feasibility; i.e.,

$$\max_{(k,l)\notin S_j^*} |\widehat{g}_{jkl}| < 1,$$

which in turn implies that the oracle estimator $\widehat{\theta}_j$ satisfies the KKT condition of the localized kernel ODE optimization problem.

For any $(k,l) \notin S_j^*$, by (S23), we have,

$$2G_{kl}^\top R_{t_0}(G_{S_j^*}\widehat{\theta}_{S_j^*} - z_j) + n\kappa_{nj}\widehat{g}_{jkl} = 0,$$

which implies that

$$2G_{kl}^\top R_{t_0}(G_{S_j^*}\widehat{\theta}_{S_j^*} - G_{S_j^*}\theta_{S_j^*}) + 2G_{kl}^\top R_{t_0}(G_{S_j^*}\theta_{S_j^*} - z_j) + n\kappa_{nj}\widehat{g}_{jkl} = 0.$$

By (S26) and (S27), we have,

$$n\kappa_{nj}\widehat{g}_{jkl} = G_{kl}^\top R_{t_0} G_{S_j^*}(G_{S_j}^\top G_{S_j^*})^{-1}(\mathcal{R}_{S_j^*} + n\kappa_{nj}\widehat{g}_{S_j^*}) - \mathcal{R}_{kl}.$$

By Assumption 5 again, and by (S28) and (S29), we have that,

$$|\widehat{g}_{jkl}| \le \frac{(\xi_G+1)\sqrt{s_j}}{n\kappa_{nj}}\eta_{\mathcal{R}} + \xi_G\sqrt{s_j}, \quad (k,l) \notin S_j^*.$$

By Assumption 6, we obtain that, $|\widehat{g}_{jkl}| < 1$, for any $(k,l) \notin S_j^*$.

Finally, the selection consistency for S_j^* implies the selection consistency for $\widehat{S}_j$. This completes the proof of Theorem 3. ■

B.3.2 Auxiliary lemma for Theorem 3

The next lemma gives a bound similar to the deviation condition in Loh and Wainwright (2012) and Dai and Li (2022). The difference is that, the noise in the variable $\widehat{x}(t)$ in our setting involves a nonlinear transformation through the kernel $K(\widehat{x}(t), \widehat{x}(s))$. Besides, we have adopted the localized learning.

Lemma 6 *For $j = 1, \ldots, p$, we have,*

$$\|G_{kl}^\top R_{t_0} G_{S_j^*}\theta_{S_j^*} - G_{kl}^\top R_{t_0} z_j\|_{\ell_2} \le \eta_{\mathcal{R}},$$

where

$$\eta_{\mathcal{R}} = O_p\left(\left(\frac{n}{\log n}\right)^{-\frac{\beta_2}{2(\beta_2+1)}} + \left(\frac{\log p}{n}\right)^{1/2} + n^{-\frac{\beta_1}{2\beta_1+1}}\right).$$

Proof Similar to the "predictor" G defined in (18) in Section 3.2, we first construct a noiseless version of the "predictor", $\widetilde{G} \in \mathbb{R}^{n\times p^2}$, whose first p columns are $\widetilde{\Sigma}^{k'}c_j$, the last $p(p-1)$ columns are $\widetilde{\Sigma}^{k'l}c_j$, and $\widetilde{\Sigma}^{k'} = (\widetilde{\Sigma}_{ii'}^{k'}), \widetilde{\Sigma}^{k'l} = (\widetilde{\Sigma}_{ii'}^{k'l})$ are both $n\times n$ matrices whose (i,i')th entries are,

$$\widetilde{\Sigma}_{ii'}^{k'} = \int_{\mathcal{T}}\int_{\mathcal{T}}\{T_i(s) - \bar{T}(s)\}K_{k'}(x(t),x(s))\{T_{i'}(t) - \bar{T}(t)\}dsdt, \;\; 1 \le k' \le p, 1 \le i, i' \le n,$$
$$\widetilde{\Sigma}_{ii'}^{k'l} = \int_{\mathcal{T}}\int_{\mathcal{T}}\{T_i(s) - \bar{T}(s)\}K_{k'l}(x(t),x(s))\{T_{i'}(t) - \bar{T}(t)\}dsdt, \;\; 1 \le k' < l \le p, 1 \le i, i' \le n.$$

Next, we consider the term $\left\|G_{kl}^\top R_{t_0} z_j - G_{kl}^\top R_{t_0} G_{S_j^*}\theta_{S_j^*}\right\|_{\ell_2}$, which can be bounded as,

$$\begin{aligned}\left\|G_{kl}^\top R_{t_0}\left(z_j - G_{S_j^*}\theta_{S_j^*}\right)\right\|_{\ell_2} \le\; & \left\|G_{kl}^\top R_{t_0}\left(\mathbb{E}[z_j] - \widetilde{G}_{S_j^*}\theta_{S_j^*}\right)\right\|_{\ell_2} + \left\|G_{kl}^\top R_{t_0}\left(\widetilde{G}_{S_j^*} - G_{S_j^*}\right)\theta_{S_j^*}\right\|_{\ell_2} \\ & + \left\|G_{kl}^\top R_{t_0}\left(z_j - \mathbb{E}[z_j]\right)\right\|_{\ell_2} \\ \equiv\; & \Delta_7 + \Delta_8 + \Delta_9.\end{aligned} \tag{S30}$$

We next bound the three terms $\Delta_7, \Delta_8, \Delta_9$ on the right-hand-side of (S30), respectively.

For Δ_7, by the Cauchy-Schwarz inequality and Theorem 1, we have,

$$\begin{aligned}\Delta_7^2 &\leq \left\|G_{kl}^\top\right\|_{\ell_2}^2 \left\|R_{t_0}\left(\mathbb{E}[z_j] - \widetilde{G}_{S_j^*}\theta_{S_j^*}\right)\right\|_{\ell_2}^2 \\ &\leq C_1 \left\|R_{t_0}\left(\mathbb{E}[z_j] - \widetilde{G}_{S_j^*}\theta_{S_j^*}\right)\right\|_{\ell_2}^2 = O_p\left(\left(\frac{n}{\log n}\right)^{-\frac{2\beta_2}{2(\beta_2+1)}} + \frac{\log p}{n}\right),\end{aligned}$$

for some constant $C_1 > 0$, where the last step is by (S4).

For Δ_8, again by the Cauchy-Schwarz inequality, we have,

$$\Delta_8^2 \leq \left\|G_{kl}^\top\right\|_{\ell_2}^2 \left\|R_{t_0}\left(\widetilde{G}_{S_j^*} - G_{S_j^*}\right)\theta_{S_j^*}\right\|_{\ell_2}^2 \leq C_2 \left\|R_{t_0}\left(\widetilde{G}_{S_j^*} - G_{S_j^*}\right)\right\|_\infty^2 \left\|\theta_{S_j^*}\right\|_{\ell_1}^2 = O_p\left(n^{-\frac{2\beta_1}{2\beta_1+1}}\right),$$

for some constants $C_2 > 0$, where the last step is by (S3), and the fact that $\|\theta_{S_j^*}\|_{\ell_1}$ is bounded.

For Δ_9, by Lemma 5, we have,

$$\Delta_9^2 = O_p\left(\left(\frac{n}{\log n}\right)^{-\frac{2\beta_2}{2(\beta_2+1)}} + \frac{\log p}{n}\right).$$

Combining the above three bounds, we obtain that,

$$\left\|G_{kl}^\top R_{t_0}(z_j - G_{S_j^*}\theta_{S_j^*})\right\|_{\ell_2} = O_p\left(\left(\frac{n}{\log n}\right)^{-\frac{\beta_2}{2(\beta_2+1)}} + \left(\frac{\log p}{n}\right)^{1/2} + n^{-\frac{\beta_1}{2\beta_1+1}}\right),$$

which completes the proof of Lemma 6. ■

Appendix C. Additional Theoretical Discussions

In this section, we discuss more on the averaging operator introduced in Section 2.1. We also analyze Assumption 5 in Section 5.2 in more detail.

C.1 Averaging Operator

Consider a standard one-way ANOVA model, $Y_{ti} = f_t + \epsilon_{ti}$, where f_t denotes the treatment mean at treatment level $t = 1, \ldots, T$, $i = 1, \ldots, n$ indexes the sample observations, and ϵ_{ti} is an independent normal error. The ANOVA decomposition is written as,

$$f_t = \theta_0 + \alpha_t,$$

where θ_0 is the global mean, and α_t is the treatment effect at level t. The parameters θ_0 and α_t are identifiable through a side condition, where a common choice is that $\sum_{t=1}^T \alpha_t = 0$.

Similarly, the one-way ANOVA model on a continuous domain $\mathcal{T}$ can be cast as $Y_i = f(t_i) + \epsilon_i$, where $t \in \mathcal{T}$. The ANOVA decomposition is written as,

$$f(t) = \mathcal{A}f + (I - \mathcal{A})f,$$

where $\mathcal{A}$ is an averaging operator that "averages out" the covariate t to return a constant function, and I is the identity operator. For instance, with $\mathcal{A}f = \int_{\mathcal{T}} f(t)dt$, one has $f(t) = \int_{\mathcal{T}} f(t)dt + \{f(x) - \int_{\mathcal{T}} f(t)dt\}$, corresponding to $\sum_{t=1}^{T} \alpha_t = 0$ in the standard one-way ANOVA model. Note that applying $\mathcal{A}$ to a constant function returns that constant, hence the name "averaging." It follows that $\mathcal{A}(\mathcal{A}f) = \mathcal{A}f$, or simply, $\mathcal{A}^2 = \mathcal{A}$. Write the constant term $\theta_0 = \mathcal{A}f$, which denotes the global mean. Write the term $\tilde{f} = (I - \mathcal{A})f$, which denotes the treatment effect that satisfies the side condition $\mathcal{A}\tilde{f} = \int_{\mathcal{T}} \tilde{f}(t)dt = 0$. Following this reasoning, we define the averaging operator for our model (3) as

$$\mathcal{A}F_j(x(t)) = \int_{\mathcal{T}} F_j(x(t))dt.$$

Then in the construction of the RKHS $\mathcal{H}$ in model (4), a sufficient side condition is the zero marginal integral for each $k = 1, \ldots, p$, such that

$$\int_{\mathcal{T}} F_{jk}(x_k(t))dt = 0, \quad \text{for any} \;\; k = 1, \ldots, p.$$

Such an averaging operator has been commonly used in the RKHS literature; see, e.g., Wahba et al. (1995); Gu (2013); Lin and Zhang (2006); Dai and Li (2022).

C.2 Discussion on Assumption 5

We further study Assumption 5, and show that the bandwidth h in the localization and R_{t_0} does not affect the validity of this assumption, as long as $h \to 0$ when $n \to \infty$.

We start with the first part of Assumption 5. Note that,

$$\begin{aligned}\Lambda_{\min}\left(G_{S_j^*}^\top R_{t_0} G_{S_j^*}\right) &= \left[\sigma_{\min}\left(R_{t_0}^{1/2} G_{S_j^*}\right)\right]^2 \\ &\geq \left[\sigma_{\min}\left(R_{t_0}^{1/2}\right)\sigma_{\min}(G_{S_j^*})\right]^2 = \Lambda_{\min}(R_{t_0})\Lambda_{\min}(G_{S_j^*}^\top G_{S_j^*}).\end{aligned}$$

Here $\sigma_{\min}$ denotes the minimum singular value and $\Lambda_{\min}$ denotes the minimum eigenvalue. Since the bandwidth $h \to 0$ as $n \to \infty$, we have that $R_h(t_i - t_0) \to \frac{1}{t_i - t_0} c_R$, where $c_R = \lim_{h \to 0} h^{-1} R(h^{-1}) > 0$ is a constant. Therefore, as $n \to \infty$, $\Lambda_{\min}(R_{t_0}) = \min_i \frac{1}{t_i - t_0} c_R > 0$, which does not depend on the bandwidth h. On the other hand, Assumption 3 in Dai and Li (2022) states that $\Lambda_{\min}(G_{S_j^*}^\top G_{S_j^*})$ is lower bounded by a constant. Together, it implies that the bandwidth h does not affect the validity of the first part of this assumption.

Similarly, for the second part of Assumption 5, we have that,

$$\begin{aligned}\left\|G_{kl}^\top R_{t_0} G_{S_j^*} (G_{S_j^*}^\top R_{t_0} G_{S_j^*})^{-1}\right\|_{\ell_2} &\leq \left[\Lambda_{\min}\left(G_{S_j^*}^\top R_{t_0} G_{S_j^*}\right)\right]^{-1} \left\|G_{kl}^\top R_{t_0} G_{S_j^*}\right\|_{\ell_2} \\ &\leq \Lambda_{\max}(R_{t_0}) \left[\Lambda_{\min}(R_{t_0})\right]^{-1} \left[\Lambda_{\min}(G_{S_j^*}^\top G_{S_j^*})\right]^{-1} \left\|G_{kl}^\top G_{S_j^*}\right\|_{\ell_2}.\end{aligned}$$

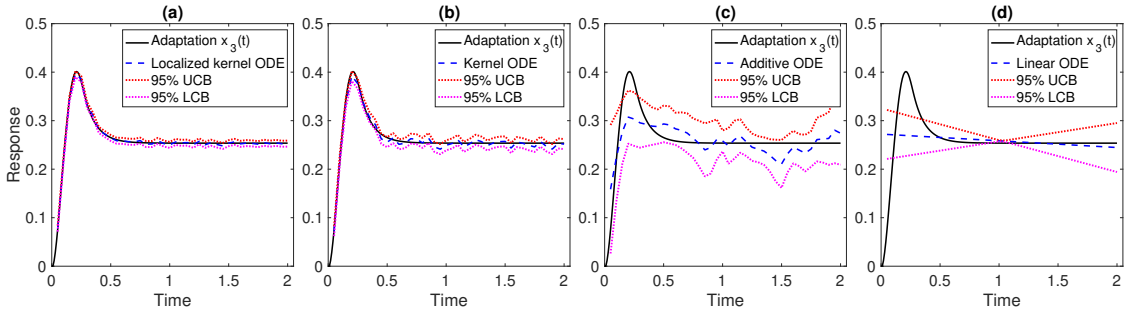

Figure S1: The true (black solid line) and the estimated (blue dashed line) trajectory of $x_3(t)$, with the 95% upper and lower confidence bounds (red dotted lines). The results are averaged over 500 data replications. (a) Localized kernel ODE; (b) Kernel ODE; (c) Additive ODE; (d) Linear ODE.

Since $R_h(t_i - t_0) \to \frac{1}{t_i - t_0} c_R$, $\Lambda_{\max}(R_{t_0}) \left[\Lambda_{\min}(R_{t_0})\right]^{-1} \leq \max_{i,j} \frac{t_i - t_0}{t_j - t_0}$ as $n \to \infty$, which does not depend on the bandwidth h. Again, Assumption 4 in Dai and Li (2022) states that $\max_{(k,l) \notin S_j^*} \left\| G_{kl}^\top G_{S_j^*} (G_{S_j^*}^\top G_{S_j^*})^{-1} \right\|_{\ell_2} \leq \left[\Lambda_{\min}(G_{S_j^*}^\top G_{S_j^*})\right]^{-1} \left\| G_{kl}^\top G_{S_j^*} \right\|_{\ell_2}$ is upper bounded by ξ_G. Together, it implies that the bandwidth h does not affect the validity of the second part of this assumption 5.

Appendix D. Additional Numerical Analyses

In this section, we report some additional numerical results for the simulation study in Section 6. We also carry out a sensitivity analysis to investigate the effect of the choice of the local weight function and bandwidth.

D.1 Additional Results about Enzymatic Regulation Equations

Figure S1 reports the true and estimated trajectory of $x_3(t)$, with 95% upper and lower confidence bounds, of the four ODE methods. 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 localized kernel ODE estimate has a smaller variance than its counterparts, including the kernel ODE method. Additionally, the confidence intervals of localized kernel ODE and kernel ODE achieve the desired coverage for the true trajectory. In contrast, the confidence intervals of additive and linear ODE models mostly fail to include the truth.

Figure S2 reports the coverage probability and confidence band area for the varying noise level of different ODE methods, and is a visualization of Table 1 in Section 6.1. We see that the inference method based on the localized kernel ODE clearly outperforms the alternative solutions with a larger coverage probability and a more tight confidence band.

Figure S3 reports the empirical FDR, power, and trajectory estimation error for the varying noise level when we aim to recover the entire regulatory system through the proposed confidence band coupled with the BH procedure for multiple testing correction at the FDR

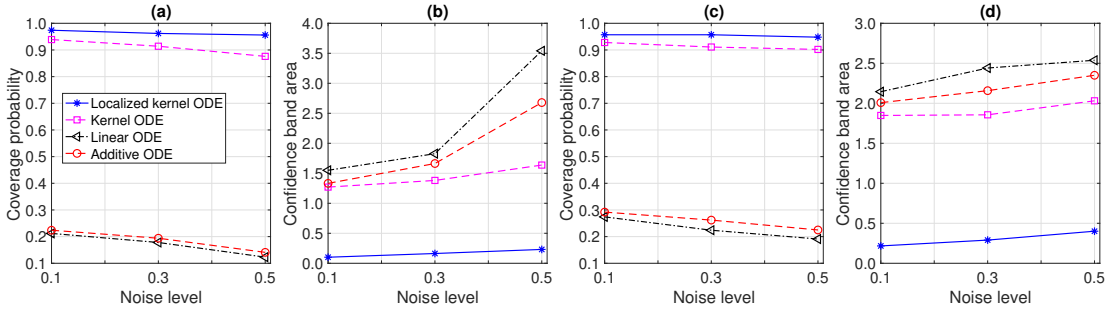

Figure S2: The NFBLB example: the empirical coverage probability and confidence band area for the varying noise level σ_j. The results are averaged over 500 data replications. (a)-(b): Nonzero functional $F_{23}(x_3(t))$; (c)-(d): Zero functional $F_{12}(x_2(t))$.

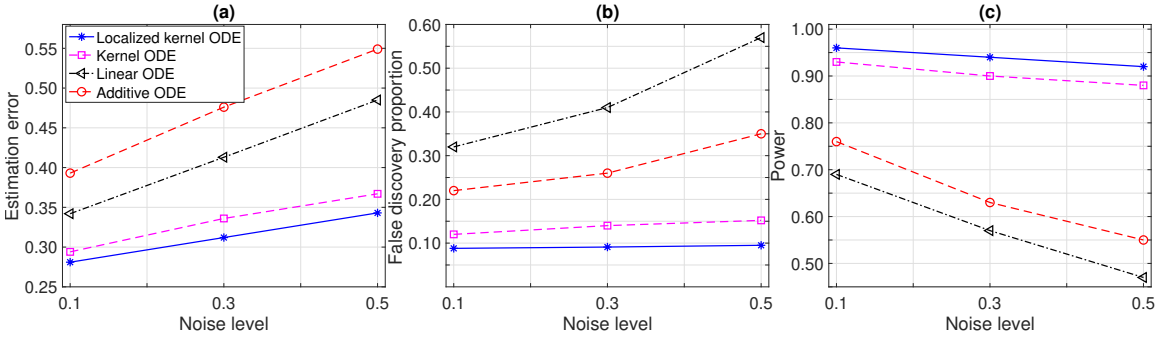

Figure S3: The NFBLB example: the estimation and sparse selection performance for the varying noise level σ_j. The results are averaged over 500 data replications.

level of 10%. Here, the empirical FDR and power are the same as defined in Section 6.2. The estimation accuracy of $\widehat{F}_j(\widehat{x}(t))$ is defined as the squared root of the sum of mean squared errors for $F_j(x_j(t)), j = 1, 2, 3$, at $t \in [0, 2]$, i.e., $\left\{\sum_{j=1}^{3} \int_0^2 [\widehat{F}_j(\widehat{x}_j(t)) - F_j(x_j(t))]^2 dt\right\}^{1/2}$, where the integral is evaluated at 10000 evenly distributed time points in $[0, 2]$. We see that the inference method based on the localized kernel ODE successfully controls the FDR under the nominal level, and outperforms the three alternative solutions in terms of the empirical power and the estimation error.

Figure S4 reports the prediction accuracy of the entire regulatory effect $\widehat{F}_j(\widehat{x}(t))$, as well as the individual regulatory effects $\widehat{F}_{23}(\widehat{x}_3(t))$ and $\widehat{F}_{12}(\widehat{x}_2(t))$. For the entire regulatory effect, the predictor error is computed as the squared root of the sum of predictive mean squared errors for $F_j(x_j(t)), j = 1, 2, 3$, at the unseen "future" time point $t \in [2, 3]$, i.e., $\left\{\sum_{j=1}^{3} \int_2^3 [\widehat{F}_j(\widehat{x}_j(t)) - F_j(x_j(t))]^2 dt\right\}^{1/2}$, where the integral is evaluated at 10000 evenly distributed time points in $[2, 3]$. Similarly, for the individual regula-

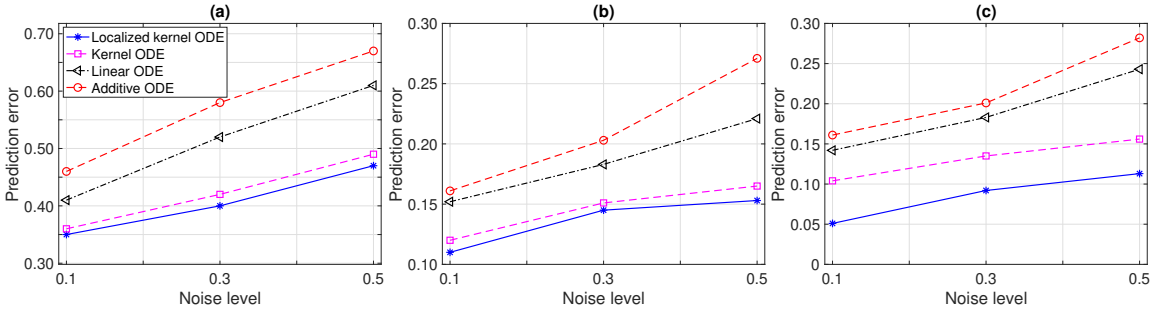

Figure S4: The NFBLB example: the prediction error for the varying noise level σ_j. The results are averaged over 500 data replications. (a) Entire regulatory effect $\widehat{F}_j(x(t)), j = 1, 2, 3$; (b) Individual regulatory effect $\widehat{F}_{23}(\widehat{x}_3(t))$; (c) Individual regulatory effect $\widehat{F}_{12}(\widehat{x}_2(t))$.

tory effects, the predictor error is computed as $\left\{\int_2^3 [\widehat{F}_{23}(\widehat{x}_3(t)) - F_{23}(x_3(t))]^2 dt\right\}^{1/2}$, and $\left\{\int_2^3 [\widehat{F}_{12}(\widehat{x}_2(t)) - F_{12}(x_2(t))]^2 dt\right\}^{1/2}$, respectively. We see that the prediction error of the localized kernel ODE estimator for the *entire* regulatory effect is comparable with that of kernel ODE, which agrees with Theorem 1. Moreover, Figure S4(b) and (c) show that the proposed method outperforms the kernel ODE estimator in predicting the *individual* regulatory effect. This is because the kernel ODE estimator primarily targets the sum of all individual effects, whereas our proposed method directly estimates the individual functional that measures the regulatory effect of one signal variable on another.

D.2 Sensitivity Analysis

We carry out a sensitivity analysis regarding the local weight function $R_h(t)$ and the bandwidth h using the enzymatic regulation equations example in Section 6.1. We show that the inference results are not sensitive to the choice of the weight function or the bandwidth.

First, we consider three different local weight functions: the quadratic weight, the cubic weight, and the Gaussian weight,

$$\begin{aligned} R_h^{(1)}(t) &= (15/16) \cdot (1 - t^2/h^2)^2 \mathbf{1}(|t| < h), \\ R_h^{(2)}(t) &= (1 - t^2/h^2)^3 \mathbf{1}(|t| < h), \\ R_h^{(3)}(t) &= \exp(-t^2/2h^2). \end{aligned}$$

We couple them with the proposed localized kernel ODE method, while we continue to choose the bandwidth h using tenfold cross-validation. We couple the proposed confidence band with the BH procedure at the FDR level of 10%. We consider three evaluation criteria, the false discovery proportion, the empirical power, and the estimation error, the same as those used in Figure S3. Figure S5 reports the performance of the localized kernel ODE method with the three local weight functions, denoted as Local-KODE-1, 2, 3, respectively, plus the kernel ODE, linear ODE, and additive ODE methods. We see that the performances

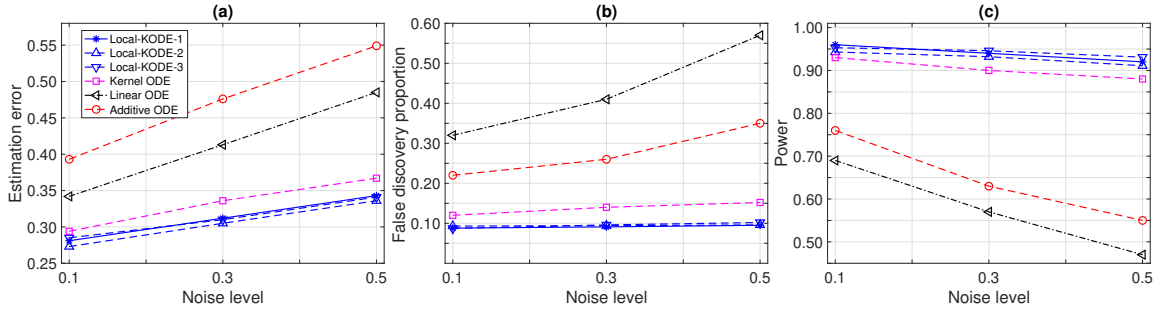

Figure S5: Sensitivity analysis: the estimation and sparse selection performance for localized kernel ODE with three different local weight functions, plus kernel ODE, linear ODE, and additive ODE. The results are averaged over 500 data replications.

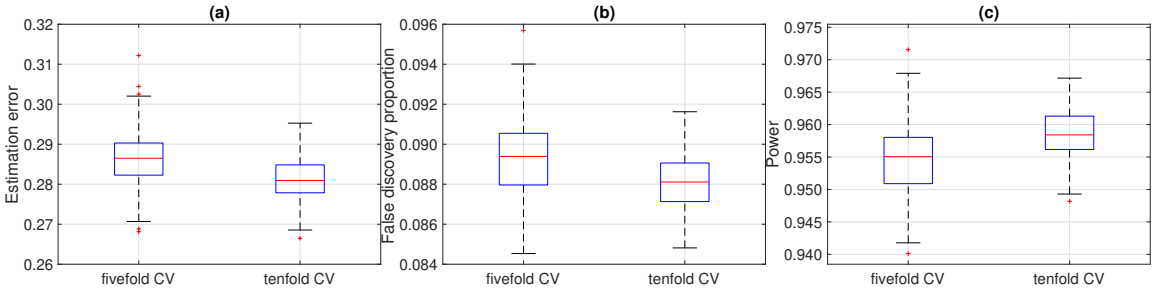

Figure S6: Sensitivity analysis: the estimation and sparse selection performance for localized kernel ODE with two different choices of kernel bandwidth. The boxes range from the lower to the upper quartile, and the whiskers extend to the most extreme data point that is no more than 1.5 times the interquartile range from the box. The results are collected over 500 data replications.

of the three localized kernel ODE methods are fairly close. The false discovery proportions differ at most 0.7%, the empirical powers differ at most 1.7%, and the estimation errors differ at most 4.2%, across different noise levels. Besides, they all outperform the alternative solutions. These results show that the proposed localized kernel ODE method is relatively robust to the choice of the local weight function.

Next, we consider the selection of bandwidth h. We experiment with fivefold cross-validation and tenfold cross-validation of minimizing the residual sums of squares. We use the quadratic local weight function, and fix the noise level of the enzymatic regulation equations example at $\sigma_j = 0.1, j = 1, 2, 3$. Figure S6 reports the estimation and sparse selection performance of the localized kernel ODE method. We see that the performances under these two different choices of the bandwidth h are close. For the medians, the false discovery proportions differ at most 0.15%, the empirical powers differ at most 0.3%, and

the estimation errors differ at most 1.8%. These results show that the proposed localized kernel ODE method is relatively robust to the choice of bandwidth.

References

Tom M. Apostol. *Calculus*. John Wiley & Sons, New York, 1967.

Nachman Aronszajn. Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68(3):337–404, 1950.

Francis Bach. On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18(1):714–751, 2017.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B (Methodological)*, 57(1):289–300, 1995.

Jiguo Cao and Hongyu Zhao. Estimating dynamic models for gene regulation networks. *Bioinformatics*, 24:1619–1624, 2008.

Xuefei Cao, Björn Sandstede, and Xi Luo. A functional data method for causal dynamic network modeling of task-related fMRI. *Frontiers in Neuroscience*, 13:127, 2019.

Shizhe Chen, Ali Shojaie, and Daniela M. Witten. Network reconstruction from high-dimensional ordinary differential equations. *Journal of the American Statistical Association*, 112:1697–1707, 2017.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Anti-concentration and honest, adaptive confidence bands. *Annals of Statistics*, 42(5):1787–1818, 2014.

Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *American Mathematical Society Bulletin*, 39:1–49, 2002.

Xiaowu Dai and Hengzhi He. Continuous time analysis of dynamic matching in heterogeneous networks. *arXiv preprint arXiv:2302.09757*, 2023.

Xiaowu Dai and Lexin Li. Kernel ordinary differential equations. *Journal of the American Statistical Association*, 117(540):1711–1725, 2022.

Itai Dattner and Chris A. J. Klaassen. Optimal rate of direct estimators in systems of ordinary differential equations linear in functions of the parameters. *Electronic Journal of Statistics*, 9:1939–1973, 2015.

Jianqing Fan. Local linear regression smoothers and their minimax efficiencies. *Annals of Statistics*, 21(1):196 – 216, 1993.

Jianqing Fan and Irene Gijbels. *Local Polynomial Modelling and Its Applications*. Monographs on Statistics and Applied Probability (Vol. 66), London: Chapman & Hall, 1996.

Karl J. Friston. Functional and effective connectivity: A review. *Brain Connectivity*, 1(1): 13–36, 2011.

Evarist Giné and Joel Zinn. Bootstrapping general empirical measures. *Annals of Probability*, 18(2):851–869, 1990.

Javier González, Ivan Vujačić, and Ernst Wit. Inferring latent gene regulatory network kinetics. *Statistical Applications in Genetics and Molecular Biology*, 12(1):109–127, 2013.

Javier González, Ivan Vujačić, and Ernst Wit. Reproducing kernel hilbert space based estimation of systems of ordinary differential equations. *Pattern Recognition Letters*, 45: 26–32, 2014.

Chong Gu. *Smoothing Spline ANOVA Models*. Springer Science & Business Media, 2013.

James Henderson and George Michailidis. Network reconstruction using nonparametric additive ode models. *PLOS One*, 9:1–15, 2014.

Anton Henssen, Karl Zilles, Nicola Palomero-Gallagher, Axel Schleicher, Hartmut Mohlberg, Fatma Gerboga, Simon B. Eickhoff, Sebastian Bludau, and Katrin Amunts. Cytoarchitecture and probability maps of the human medial orbitofrontal cortex. *Cortex*, 75:87–112, 2016.

Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

Jianhua Z. Huang. Projection estimation in multiple regression with application to functional anova models. *Annals of Statistics*, 26(1):242–272, 1998.

Eugene M. Izhikevich. *Dynamical Systems in Neuroscience*. MIT Press, 2007.

Adel Javanmard and Andrea Montanari. Confidence intervals and hypothesis testing for high-dimensional regression. *Journal of Machine Learning Research*, 15:2869–2909, 2014.

Vladimir Koltchinskii and Ming Yuan. Sparsity in multiple kernel learning. *Annals of Statistics*, 38(6):3660–3695, 2010.

Morten L. Kringelbach and Edmund T. Rolls. The functional neuroanatomy of the human orbitofrontal cortex: evidence from neuroimaging and neuropsychology. *Progress in Neurobiology*, 72(5):341–372, 2004.

Yun Li, Ji Zhu, and Naisyin Wang. Regularized semiparametric estimation for ordinary differential equations. *Technometrics*, 57(3):341–350, 2015.

Hua Liang and Hulin Wu. Parameter estimation for differential equation models using a framework of measurement error in regression models. *Journal of the American Statistical Association*, 103(484):1570–1583, 2008.

Yi Lin and Hao H. Zhang. Component selection and smoothing in multivariate nonparametric regression. *Annals of Statistics*, 34:2272–2297, 2006.

Po-Ling Loh and Martin J. Wainwright. High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Annals of Statistics*, 40:1637–1664, 2012.

Junwei Lu, Mladen Kolar, and Han Liu. Kernel meets sieve: Post-regularization confidence bands for sparse additive model. *Journal of the American Statistical Association*, 115 (532):2084–2099, 2020.

Tao Lu, Hua Liang, Hongzhe Li, and Hulin Wu. High-dimensional ODEs coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. *Journal of the American Statistical Association*, 106:1242–1258, 2011.

Wenzhe Ma, Ala Trusina, Hana El-Samad, Wendell A. Lim, and Chao Tang. Defining network topologies that can achieve biochemical adaptation. *Cell*, 138:760–773, 2009.

Daniel Marbach, Thomas Schaffter, Claudio Mattiussi, and Dario Floreano. Generating realistic in silico gene networks for performance assessment of reverse engineering methods. *Journal of Computational Biology*, 16:229–239, 2009.

Daniel Marbach, Robert J. Prill, Thomas Schaffter, Claudio Mattiussi, Dario Floreano, and Gustavo Stolovitzky. Revealing strengths and weaknesses of methods for gene network inference. *Proceedings of the national academy of sciences*, 107(14):6286–6291, 2010.

Nicolai Meinshausen and Peter Bühlmann. High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462, 2006.

Hongyu Miao, Hulin Wu, and Hongqi Xue. Generalized ordinary differential equation models. *Journal of the American Statistical Association*, 109(508):1672–1682, 2014.

Frederik Vissing Mikkelsen and Niels Richard Hansen. Learning large scale ordinary differential equation systems. *arXiv preprint arXiv:1710.09308*, 2017.

Yang Ning and Han Liu. A general theory of hypothesis tests and confidence regions for sparse high dimensional models. *Annals of Statistics*, 45(1):158 – 195, 2017.

Jean D. Opsomer and David Ruppert. Fitting a bivariate additive model by local polynomial regression. *Annals of Statistics*, 25(1):186–211, 1997.

Jie Peng, Pei Wang, Nengfeng Zhou, and Ji Zhu. Partial correlation estimation by joint sparse regression models. *Journal of the American Statistical Association*, 104(486):735–746, 2009.

Xin Qi and Hongyu Zhao. Asymptotic efficiency and finite-sample properties of the generalized profiling estimation of parameters in ordinary differential equations. *Annals of Statistics*, 38(1):435–481, 2010.

Jim O. Ramsay, Giles Hooker, David Campbell, and Jiguo Cao. Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 69(5):741–796, 2007.

Garvesh Raskutti, Martin J. Wainwright, and Bin Yu. Minimax rates of estimation for high-dimensional linear regression over ℓ_q-balls. *IEEE Transactions on Information Theory*, 57(10):6976–6994, 2011.

Pradeep Ravikumar, Martin J. Wainwright, and John D. Lafferty. High-dimensional Ising model selection using l_1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

Ignacio Saez, Jack Lin, Arjen Stolk, Edward Chang, Josef Parvizi, Gerwin Schalk, Robert T. Knight, and Ming Hsu. Encoding of multiple reward-related computations in transient and sustained high-frequency activity in human OFC. *Current Biology*, 28(18):2889–2899, 2018.

Thomas Schaffter, Daniel Marbach, and Dario Floreano. Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, 27:2263–2270, 2011.

Bernhard Scholkopf and Alexander J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, 2018.

Xiaotong Shen and Wing Hung Wong. Convergence rate of sieve estimates. *Annals of Statistics*, pages 580–615, 1994.

Michel Talagrand. New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126(3):505–563, 1996.

Sara van de Geer. *Empirical Processes in M-Estimation*. Cambridge University Press, 2000.

Sara van de Geer, Peter Bühlmann, Ya'acov Ritov, and Ruben Dezeure. On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3):1166–1202, 2014.

Aad van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer-Verlag, New York, 1996.

James M. Varah. A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific and Statistical Computing*, 3:28–46, 1982.

Vito Volterra. Variations and fluctuations of the number of individuals in animal species living together. *ICES Journal of Marine Science*, 3(1):3–51, 1928.

Grace Wahba. *Spline Models for Observational Data*. SIAM, Philadelphia, 1990.

Grace Wahba, Yuedong Wang, Chong Gu, Ronald Klein, and Barbara Klein. Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Annals of Statistics*, 23:1865–1895, 1995.

Tom Wansbeek and Erik Meijer. Measurement error and latent variables. *A Companion to Theoretical Econometrics. Oxford: Basil Blackwell*, pages 162–179, 2001.

Hulin Wu, Tao Lu, Hongqi Xue, and Hua Liang. Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. *Journal of the American Statistical Association*, 109:700–716, 2014.

Leqin Wu, Xing Qiu, Ya-Xiang Yuan, and Hulin Wu. Parameter estimation and variable selection for big systems of linear ordinary differential equations: A matrix-based approach. *Journal of the American Statistical Association*, 114(526):657–667, 2019.

Hongqi Xue, Hongyu Miao, and Hulin Wu. Sieve estimation of constant and time-varying coefficients in nonlinear ordinary differential equation models by considering both numerical error and measurement error. *Annals of Statistics*, 38(4):2351, 2010.

Ming Yuan and Ding-Xuan Zhou. Minimax optimal rates of estimation in high dimensional additive models. *Annals of Statistics*, 44(6):2564–2593, 2016.

Cun-Hui Zhang and Stephanie S. Zhang. Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society: Series B.*, 76(1):217–242, 2014.

Tingting Zhang, Jingwei Wu, Fan Li, Brian Caffo, and Dana Boatman-Reich. A dynamic directional model for effective brain connectivity using electrocorticographic (ECoG) time series. *Journal of the American Statistical Association*, 110(509):93–106, 2015a.

Tingting Zhang, Qiannan Yin, Brian Caffo, Yinge Sun, and Dana Boatman-Reich. Bayesian inference of high-dimensional, cluster-structured ordinary differential equation models with applications to brain connectivity studies. *Annals of Applied Statistics*, 11:868–897, 2017.

Xianyang Zhang and Guang Cheng. Simultaneous inference for high-dimensional linear models. *Journal of the American Statistical Association*, 112(518):757–768, 2017.

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

Peng Zhao and Bin Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

Hongxiao Zhu, Fang Yao, and Hao H. Zhang. Structured functional additive regression in reproducing kernel Hilbert spaces. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 76:581–603, 2014.