Computational Statistics and Data Analysis 173 (2022) 107483

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

www.elsevier.com/locate/csda

Dynamical modeling for non-Gaussian data with high-dimensional sparse ordinary differential equations

Muye Nanshan[a], Nan Zhang[a,*], Xiaolei Xun[b], Jiguo Cao[c]

[a] *School of Data Science, Fudan University, Shanghai, China*
[b] *Global Statistics and Data Science, BeiGene, China*
[c] *Department of Statistics and Actuarial Science, Simon Fraser University, Burnaby, British Columbia, Canada*

ARTICLE INFO

Article history:
Received 7 September 2021
Received in revised form 23 March 2022
Accepted 26 March 2022
Available online 5 April 2022

Keywords:
Dynamic system
Generalized linear model
Ordinary differential equations
Parameter cascade
Penalized likelihood
Profiled estimation

ABSTRACT

Ordinary differential equations (ODE) have been widely used for modeling dynamical complex systems. For high-dimensional ODE models where the number of differential equations is large, it remains challenging to estimate the ODE parameters and to identify the sparse structure of the ODE models. Most existing methods exploit the least-square based approach and are only applicable to Gaussian observations. However, as discrete data are ubiquitous in applications, it is of practical importance to develop dynamic modeling for non-Gaussian observations. New methods and algorithms are developed for both parameter estimation and sparse structure identification in high-dimensional linear ODE systems. First, the high-dimensional generalized profiling method is proposed as a likelihood-based approach with ODE fidelity and sparsity-inducing regularization, along with efficient computation based on parameter cascading. Second, two versions of the two-step collocation methods are extended to the non-Gaussian set-up by incorporating the iteratively reweighted least squares technique. Simulations show that the profiling procedure has excellent performance in latent process and derivative fitting and ODE parameter estimation, while the two-step collocation approach excels in identifying the sparse structure of the ODE system. The usefulness of the proposed methods is also demonstrated by analyzing three real datasets from Google trends, stock market sectors, and yeast cell cycle studies.

1. Introduction

Ordinary differential equations (ODE) are widely used for complex dynamic system modeling in biology, engineering, econometrics, and other scientific and social applications. For example, massive gene expression profiles are available with the advancement of second-generation sequencing technology. Modeling their dynamics using gene regulatory networks has drawn significant interest from both biomedical and statistical research communities (Stuart et al., 2003; Yuan and Kendziorski, 2006; Hecker et al., 2009; Polynikis et al., 2009; Lu et al., 2011; Wu et al., 2014). In computational sociology, public opinion sensing and trend analysis have emerged from the advent of the big data revolution (Dodds et al., 2011; Sloan and Morgan, 2015). Massive datasets, such as Google searches or Twitter posts, are collected daily or even hourly, which enables social scientists to extract interesting temporal or spatial patterns via dynamic modeling. The main purpose of this

[*] Corresponding author at: 220 Handan Rd., 200433, Shanghai, China.
E-mail address: zhangnan@fudan.edu.cn (N. Zhang).

https://doi.org/10.1016/j.csda.2022.107483
0167-9473/© 2022 Elsevier B.V. All rights reserved.

article is to propose new methods and algorithms to estimate the ODE parameters and to identify the sparse structure for high-dimensional ODE models with non-Gaussian observations.

A general first-order ODE system can be described as

$$\boldsymbol{\theta}'(t) = f(\boldsymbol{\theta}(t), \boldsymbol{\beta}), \tag{1.1}$$

where the vector $\boldsymbol{\theta}(t) = (\theta_1(t), \dots, \theta_p(t))^\top$ collects p processes while $\boldsymbol{\theta}'(t)$ is the first-order derivative of $\boldsymbol{\theta}(t)$, function $f = (f_1, \dots, f_p)$ describes the dependence between processes and their derivatives, $\boldsymbol{\beta}$ is the vector of ODE parameters to be estimated. Typically, the processes are indexed with time t and some initial conditions, for example, $\boldsymbol{\theta}(0) = \boldsymbol{\theta}_0$, are assumed for the ODE system (1.1) as well.

In practice, observations from the dynamic system are measured according to the realizations of latent processes $\boldsymbol{\theta}(t)$ at discrete time points. Estimation of ODE parameters from noisy data remains a challenging problem (Ramsay et al., 2007; Wu et al., 2014; Hall and Ma, 2014; Chen et al., 2017; Wu et al., 2019; Dai and Li, 2021). In general, parameter estimation procedures fall into three categories. The first approach is based on a data fitting process by nonlinear least squares. Given a set of initial ODE parameters, the ODE solutions are approximated by numerical methods, for example, the Runge-Kutta algorithm. Then the ODE parameters are updated with the nonlinear least squares. This approach is computationally intensive and can be potentially inaccurate due to iterative numerical approximations. The second approach is the two-step collocation, where the basis expansions are exploited to approximate the ODE solutions. Varah (1982) proposed to fit the processes via data smoothing methods, followed by a second stage of minimizing a least-square criterion based on the ODE system to estimate the ODE parameters. Because of its computational advantage, two-step collocation gains much popularity in the development of methodology and applications (Liang and Wu, 2008; Lu et al., 2011; Brunel et al., 2014; Wu et al., 2014; Dattner and Klaassen, 2015) and is further improved by iterative principal differential analysis (Ramsay, 1996; Poyton et al., 2006). However, the performance of two-step procedures relies heavily on the smoothing step, while the amount of roughness regularization is hard to control. The third approach is the generalized profiling procedure (Ramsay et al., 2007), which also represents ODE solutions with basis expansion as with two-step collocation methods. The essential difference is the inclusion of an ODE-induced penalty that controls the fidelity of the processes to the ODE system. The basis coefficients and ODE parameters are then estimated simultaneously from a penalized criterion using the parameter cascading algorithm (Cao and Ramsay, 2007). From a theoretical perspective, Qi and Zhao (2010) derived an upper bound on the uniform norm of the difference between the true underlying solutions and their approximations, and proved the consistency and asymptotic normality of the estimation procedure.

More recently, there has been growing interest in high-dimensional ODE systems where the number of processes p is large. For instance, the high-dimensional time-course gene expression data enables biomedical researchers to model the regulatory behaviors via a large-scale directed graphical network model. Such a task is called network recovery. The ODE system (1.1) naturally serves for this purpose by relating the dynamics of each process with all the processes in the system, and a sparse network structure can be further imposed. Lu et al. (2011) considered the high-dimensional linear ODE for dynamic gene regulatory network identification and applied the smoothly clipped absolute deviation (Fan and Li, 2001) approach for variable selection. Wu et al. (2014) further relaxed the linear assumption and investigated a sparse additive ODE model using a two-stage procedure coupled with the adaptive group Lasso technique (Wang and Leng, 2008) to deal with nonlinear effects. Chen et al. (2017) proposed an efficient procedure using the integrated form of the ODE to bypass numerical difficulty in the derivative estimation and adopted the group Lasso (Yuan and Lin, 2006) for variable selection. Wu et al. (2019) recently developed a matrix factorization based approach to ultra-high dimensional linear ODE models for parameter estimation and variable selection. To our best knowledge, existing procedures for high-dimensional ODE models are two-stage approaches.

Besides, most of the existing work assumes that observations of the ODE system are contaminated with Gaussian noises. Therefore, least-squares estimation is conveniently adopted. However, non-Gaussian observations are commonly encountered in real applications, for example, short read count data from RNA sequencing (Nagalakshmi et al., 2008), bisulfite sequencing data for DNA methylation analysis (Cokus et al., 2008), and direction of change in the stock price over time (Huang et al., 2005). The literature on non-Gaussian data analysis with the ODE system is rare. Miao et al. (2014) developed a likelihood-based parameter estimation and inference for generalized ODE models. Its extension to high-dimensional ODE models, however, is still unknown.

Motivated by network recovery tasks for time-course non-Gaussian data, this paper focuses on the parameter estimation and sparse structure identification for high-dimensional linear ODE systems with a likelihood-based approach. To facilitate versatile analysis of non-Gaussian data, we assume the observations follow a distribution from the exponential family, where $\theta_j(t)$ is known as the canonical parameter in the context of generalized linear models (McCullagh and Nelder, 1989; Wood, 2017). Assume that $t \in [0, 1]$ without loss of generality. Given a set of discrete time points $t_1, \dots, t_n$, denote by y_{ij} the measurement according to the jth latent process $\theta_j(t)$ at time $t = t_i$, $j = 1, \dots, p$. Then, the conditional distribution of y_{ij} given $\theta_j(t_i)$ admits a density function as

$$f(y_{ij} \mid \theta_j(t_i)) = \exp\left\{ \frac{y_{ij}\theta_j(t_i) - b(\theta_j(t_i))}{a(\phi)} + c(y_{ij}, \phi) \right\},$$

where $a > 0, b, c$ are known functions, ϕ is either known or considered as a nuisance parameter. Let $(y_{1j}, \dots, y_{nj})^\top$ be the vector of observations from the latent process $\theta_j(t)$, and correspondingly the canonical parameter vector be $(\theta_j(t_1), \dots, \theta_j(t_n))^\top$. Imposing a linear structure on the general model (1.1), we investigate in this work the modeling of the dynamics among latent processes $\{\theta_j(t) : j = 1, \dots, p\}$ with a high-dimensional linear ODE system, that is

$$\theta_j'(t) = \gamma_{j0} + \sum_{k=1}^{p} \gamma_{jk}\theta_k(t), \qquad j = 1, \dots, p. \tag{1.2}$$

In this article, we develop new methods and algorithms for both parameter estimation and sparse structure identification in high-dimensional linear ODE systems. First, we propose the high-dimensional generalized profiling method along with a computationally efficient procedure based on parameter cascading (Ramsay et al., 2007; Cao and Ramsay, 2007). It solves a hierarchical optimization for parameter estimation and variable selection: an outer optimization concerning the ODE parameters under sparsity regularization is performed subject to an inner optimization where latent processes expanded with basis functions are fitted by minimizing a weighted sum of data fitting and ODE fidelity criteria given ODE parameters. In particular, we regularize the structural ODE parameters based on individual differential equation and mitigate the computational burden for parameter estimation in the high-dimensional ODE system. Moreover, there are two tuning parameters involved in our procedure: one controls the balance between data fitting and ODE fidelity in the inner optimization while the other regularizes the sparsity or model complexity in the outer optimization. Their interaction may affect the overall convergence performance of the procedure in a complicated way. Due to the non-convexity nature of our objective function, we carefully design the tuning and stopping rules according to the performance of parameter estimation to help escape local minima (Carey and Ramsay, 2021). The global convergence of the proposed algorithm is analyzed. Next, we extend the two-step collocation methods (Wu et al., 2014; Chen et al., 2017), which are recently proposed for high-dimensional ODE models with Gaussian observations, to the non-Gaussian set-up. Two versions, corresponding to the vanilla collocation (Varah, 1982) and the graph reconstruction via additive differential equations (GRADE) (Chen et al., 2017), are developed under the likelihood-based framework. Efficient computation is feasible by applying the iteratively reweighted least squares technique (Wood, 2017). Finally, we apply the proposed methods to simulated and real data sets. In general, the profiling method is more efficient than two-step collocation methods in estimating the latent processes, their derivatives, and the structural ODE parameters, while one two-step collocation method excels in identifying the sparse structure of the ODE system. To sum up, the proposed methods present a versatile toolbox for parameter estimation and sparse structure identification in high-dimensional linear ODE systems.

The remainder of the article is organized as follows. Our profiled estimation approach is developed in Section 2. Detailed computational procedure and its global convergence are discussed in Section 3. In Section 4, we extend two-step collocation methods to model non-Gaussian observations. Section 5 compares empirical performance of the proposed methods. We analyze three real data examples in Section 6 with dynamical modeling approaches. Section 7 concludes the article and Appendix collects some technical details.

2. High-dimensional generalized profiling

This section introduces the proposed approach for simultaneous parameter estimation and sparse structure identification in a high-dimensional linear ODE model for non-Gaussian data under the penalized likelihood estimation framework.

Denote by $\Gamma = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_p)$ the parameter matrix of the ODE model (1.2), where $\boldsymbol{\gamma}_j = (\gamma_{j0}, \dots, \gamma_{jp})^\top \in \mathbb{R}^{p+1}$, for $j = 1, \dots, p$. These ODE parameters Γ are of primary interest in order to understand the network structure, called structural parameters hereafter. On the other hand, the latent processes θ_j's are treated as nuisance parameters. Denote by y_{ij} and $\theta_j(t_i)$ the observation and the canonical parameter of the jth latent process at time t_i, respectively. Under the profiling scheme (Ramsay et al., 2007), an intermediate fit of latent processes $\widehat{\boldsymbol{\theta}}(t; \Gamma) = (\widehat{\theta}_1(t; \Gamma), \dots, \widehat{\theta}_p(t; \Gamma))$ minimizes the following penalized likelihood criterion,

$$-\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left\{y_{ij}\theta_j(t_i) - b(\theta_j(t_i))\right\} + \lambda_\theta \sum_{j=1}^{p}\int_0^1 \left\{\theta_j'(t) - \gamma_{j0} - \sum_{k=1}^{p}\gamma_{jk}\theta_k(t)\right\}^2 \mathrm{d}t, \tag{2.1}$$

where the likelihood part measures fidelity to data, the ODE fidelity part measures the extent to which latent processes fail to satisfy the ODE system, and the tuning parameter λ_θ controls the amount of regularization. Furthermore, with $\widehat{\boldsymbol{\theta}}(t; \Gamma)$ plugged in, an estimate of the structural parameters can be obtained by minimizing a data fitting criterion with respect to Γ,

$$-\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\{y_{ij}\widehat{\theta}_j(t_i; \Gamma) - b(\widehat{\theta}_j(t_i; \Gamma))\}. \tag{2.2}$$

The generalized profiling procedure proceed iteratively with a non-decreasing sequence of λ_θ under certain rules such that the fitted processes adhere to the ODE. Identifiable issue and asymptotic behavior of the estimation procedure are studied by Ramsay et al. (2007) and Qi and Zhao (2010).

Although the generalized profiling method provides a computationally efficient treatment for the challenging ODE parameter estimation, it can only handle relatively small-scale models (Wu et al., 2019). On the one hand, for a p-dimensional linear ODE system, we have $p^2 + p$ ODE parameters to estimate in (2.2). If we further approximate the latent process $\theta_j(t)$ by basis expansion $\mathbf{c}_j^\top \mathbf{h}_j(t)$, where $\mathbf{h}_j(t)$ is an m_j-dimensional basis vector and $\mathbf{c}_j$ is the coefficient vector, then (2.1) becomes

$$-\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left\{y_{ij}\mathbf{c}_j^\top \mathbf{h}_j(t_i) - b(\mathbf{c}_j^\top \mathbf{h}_j(t_i))\right\} + \lambda_\theta \sum_{j=1}^{p}\int_0^1 \left\{\mathbf{c}_j^\top \mathbf{h}_j'(t) - \gamma_{j0} - \sum_{k=1}^{p}\gamma_{jk}\mathbf{c}_j^\top \mathbf{h}_k(t)\right\}^2 \mathrm{d}t,$$

and the total number of nuisance parameters $\sum_{j=1}^{p} m_j$ can be huge. Therefore, a direct application of the standard generalized profiling procedure to parameter estimation for high-dimensional linear ODE is computationally demanding. On the other hand, the structural parameters obtained from (2.2) indeed infer an interaction network among the latent processes, in the sense that a nonzero γ_{jk} implies that $\theta_k(t)$ has an effect on the change of $\theta_j(t)$. For better interpretation and to avoid potential over-fitting, it is reasonable to introduce some sparsity for the structural parameters. For example, the Lasso and its variants (Tibshirani, 1996; Yuan and Lin, 2006; Zou, 2006), the smoothly clipped absolute deviation (SCAD) (Fan and Li, 2001) and the minimax concave penalty (MCP) (Zhang, 2010) have been extensively studied and used to recover probabilistic graphical structures (Yuan and Lin, 2007; Fan et al., 2009; Voorman et al., 2014).

To address the above computational issues, we first notice that the data fidelity term in the penalized criterion (2.1) can be decomposed into sums of the likelihood for p individual processes. Meanwhile, the penalty term, being a squared L_2 norm of differential equations, does not admit such decomposable property. Therefore, we propose to regularize the estimate of θ_j only by the corresponding jth differential equation. Specifically, when estimating θ_j given other $\{\theta_k : k \neq j\}$ at their most recent updates, we obtain $\widehat{\theta}_j(t; \boldsymbol{\gamma}_j)$ by minimizing

$$G_j(\theta_j; \boldsymbol{\gamma}_j) = -\frac{1}{n}\sum_{i=1}^{n}\{y_{ij}\theta_j(t_i) - b(\theta_j(t_i))\} + \lambda_{\theta,j}\int_0^1 \left\{\theta_j'(t) - \gamma_{j0} - \sum_{k=1}^{p}\gamma_{jk}\theta_k(t)\right\}^2 \mathrm{d}t, \tag{2.3}$$

for $j = 1, \ldots, p$. For simplicity, we use the same tuning parameter for individual sub-problems, that is $\lambda_{\theta,j} = \lambda_\theta$ for $j = 1, \ldots, p$. Optimizing G_j involves only $p + 1$ structural parameters in the vector $\boldsymbol{\gamma}_j$ and hence the computational complexity is greatly reduced. The benefit of using (2.3) is justified from two aspects. First, it is computationally infeasible to estimate a large number of ODE parameters jointly by directly applying the original generalized profiling criterion (2.1) to the high-dimensional ODE system. Our new formulation decouples the dependency of $\widehat{\boldsymbol{\theta}}(t; \Gamma)$ on the matrix Γ into individual dependencies of $\theta_j(t; \boldsymbol{\gamma}_j)$ on the vector $\boldsymbol{\gamma}_j$. Second, from the perspective of penalized estimation, it improves the estimation for the latent process and the ODE structural parameters by employing differential equations to regularize data smoothing.

We remark on the potential risk of employing (2.3) instead of (2.1) when estimating the latent processes. Note that (2.1) aggregates all the differential equations to update the latent processes altogether such that the estimates will follow the ODE system jointly. In contrast, our method uses a single differential equation to regularize the estimation of each latent process. When the tuning parameter λ_θ increases, the parallel updating procedure (2.3) over $j = 1, \ldots, p$, is expected to achieve an approximation in a marginal way to the joint estimation by (2.1). The simulation example introduced in Section S1 of the Supplementary Material shows that the approximation by (2.3) performs reasonably well, although the joint method (2.1) has a more accurate estimate for ODE parameters.

Next, to induce sparsity to the structural parameter matrix, we estimate $\boldsymbol{\gamma}_j$ by minimizing

$$H_j(\boldsymbol{\gamma}_j) = -\frac{1}{n}\sum_{i=1}^{n}\{y_{ij}\widehat{\theta}_j(t_i; \boldsymbol{\gamma}_j) - b(\widehat{\theta}_j(t_i; \boldsymbol{\gamma}_j))\} + \mathrm{PEN}_{\lambda_{\gamma,j}}(\boldsymbol{\gamma}_j), \tag{2.4}$$

where the penalty function $\mathrm{PEN}_{\lambda_{\gamma,j}}(\boldsymbol{\gamma}_j)$ with tuning parameter $\lambda_{\gamma,j} > 0$ induces sparsity for the structural parameter of the jth differential equation. Here we also assume for simplicity that $\lambda_{\gamma,j} = \lambda_\gamma$ for $j = 1, \ldots, p$. If the fitted structural parameter vector $\widehat{\boldsymbol{\gamma}}_j$ is zero, then we say other latent processes have no impact on $\theta_j(t)$. Any zero element in $\widehat{\boldsymbol{\gamma}}_j$ implies that the corresponding process has no influence on $\theta_j(t)$. The amount of sparsity regularization is typically determined by Bayesian information criterion (BIC) type principles, which have been adopted in other ODE parameter estimation approaches (Wu et al., 2014; Chen et al., 2017).

Our new profiling estimation procedure for high-dimensional linear ODE systems consists of two objective functions (2.3) and (2.4), which are referred to as inner and outer criteria, respectively. Such a multi-criterion optimization problem is challenging due to non-convexity and non-differentiability. Specifically, we approximate the latent processes with basis expansion in the inner optimization, and basis coefficients can be solved efficiently with the Newton-Raphson method. However, the dependence of $\widehat{\theta}_j(t; \boldsymbol{\gamma}_j)$ on $\boldsymbol{\gamma}_j$ is complicated and in general non-linear, which leads to the non-convexity of H_j. Moreover, the sparsity-inducing penalty in H_j is non-differentiable at zero, making the Gauss-Newton scheme adopted by Ramsay et al. (2007) invalid under this scenario.

Algorithm 1: High-dimensional linear ODE for non-Gaussian data.

```
Input: Observations $\{y_{ij} : i = 1, \dots, n;\ j = 1, \dots, p\}$, initial ODE parameters $\Gamma^{(0)} = (\boldsymbol{\gamma}_1^{(0)}, \dots, \boldsymbol{\gamma}_p^{(0)})$, and fixed tuning parameters $\lambda_\theta$ and $\lambda_\gamma$.
Output: Estimated ODE parameters $\widehat{\Gamma} = (\widehat{\boldsymbol{\gamma}}_1, \dots, \widehat{\boldsymbol{\gamma}}_p)$.

repeat
    At step $s \geq 1$, the current estimate is $\widehat{\Gamma}^{(s)} = (\widehat{\boldsymbol{\gamma}}_1^{(s)}, \dots, \widehat{\boldsymbol{\gamma}}_p^{(s)})$.
    for $1 \leq j \leq p$ do
        Update $\boldsymbol{\gamma}_j$ via the profiling procedure.
        repeat
            1. Given current $\widetilde{\boldsymbol{\gamma}}_j$, obtain the basis coefficient estimate $\mathbf{c}_j^*(\widetilde{\boldsymbol{\gamma}}_j)$ for the $j$th latent process in the inner optimization.
            2. Apply basis expansion and update $\boldsymbol{\gamma}_j$ via minimizing the penalized reweighted least squares.
        until $\widetilde{\boldsymbol{\gamma}}_j$ converges, and set $\boldsymbol{\gamma}_j^{(s+1)} = \widetilde{\boldsymbol{\gamma}}_j$.
    end
until Estimated ODE parameter $\widehat{\Gamma}$ converges.
```

Recent advances of derivative-free optimization algorithms (Powell, 2006; Zhang et al., 2010) may provide a viable solution. Nevertheless, they are in spirit joint optimization algorithms designed for general purpose and are thus not tailored for our specific problem. In contrast, our profiling procedure enjoys not only estimation efficiency but also algorithmic efficiency due to the use of analytical expressions of derivatives. Computational details are presented in the next section. In brief, after obtaining an estimate $\widehat{\theta}_j(t; \boldsymbol{\gamma}_j)$ given the structural parameters, we linearize the likelihood component in (2.4) and formulate the outer optimization as a parameter estimation problem for a penalized generalized linear model. Therefore, the structural parameters can be readily updated by the iterative reweighted least-squares (IRLS). Through an iterative scheme between inner and outer optimizations, our profiling procedure provides ODE parameter estimates and latent process fits and identifies the sparse structure of the ODE model.

3. Computation

In this section, we provide computational details of our profiling procedure for high-dimensional linear ODE and analyze its global convergence. Minimizing criteria in (2.3) and (2.4) are referred as inner and outer optimizations. The structural parameters $\Gamma = (\boldsymbol{\gamma}_1, \dots, \boldsymbol{\gamma}_p)$ is of our primary interest, while the latent process fits by the inner optimization is regarded as a nuisance parameter. In our profiling scheme, whenever $\boldsymbol{\gamma}_j$ changes by minimizing H_j in the outer, latent process fits are then updated by solving the inner criterion G_j. Details are provided in Algorithm 1. In addition, two tuning parameters are involved in the profiling procedure, and their complex interaction affects the overall algorithmic performance because of the non-convexity of the optimization. In the following, we split the discussion into inner and outer parts. Then we discuss the practical strategy of tuning parameter selection and the global convergence of the proposed algorithm.

3.1. Inner optimization

The inner procedure aims at finding an accurate estimate for latent processes given the structural parameter Γ. Similar to the two-step collocation method (Varah, 1982) and the generalized profiling (Ramsay et al., 2007), we represent latent processes by basis expansion. Suppose $\mathbf{h}_j(t) = (\phi_{j1}(t), \dots, \phi_{jm_j}(t))$ is a set of basis functions for the jth process such that $\theta_j(t) = \mathbf{c}_j^\top \mathbf{h}_j(t)$. Choices of basis functions include polynomials, truncated power functions and splines. In our numerical study, we use B-spline due to its numerical stability and excellent empirical performance. For notation simplicity, we use the same basis $\mathbf{h}(t)$ for all latent processes.

Although critical in optimization, the basis coefficients $\mathbf{c}_j$, $j = 1, \dots, p$, are often not of direct concern and thus considered as nuisance parameters. Observing that $G_j(\theta_j; \boldsymbol{\gamma}_j)$ is convex with respect to the basis coefficients $\mathbf{c}_j$, we can apply the Newton-Raphson scheme directly. When the Hessian of $G_j(\theta_j; \boldsymbol{\gamma}_j)$ is invertible, we can start with an initial guess of $\mathbf{c}_j$ and iteratively obtain

$$\mathbf{c}_j^{(r+1)} = \mathbf{c}_j^{(r)} - \left(\left. \frac{\partial^2 G_j}{\partial \mathbf{c}_j \partial \mathbf{c}_j^\top} \right|_{\mathbf{c}_j^{(r)}} \right)^{-1} \left(\left. \frac{\partial G_j}{\partial \mathbf{c}_j} \right|_{\mathbf{c}_j^{(r)}} \right), \quad r \geq 1.$$

Analytical expressions of the derivatives involved in the above updating rule are given in Appendix A.

3.2. Outer optimization

The outer optimization is designed for updating $\boldsymbol{\gamma}_j$ with a regularized likelihood objective function (2.4). Denote by $\mathbf{c}_j^*(\boldsymbol{\gamma}_j)$ the optimal basis coefficients for $\theta_j(t; \boldsymbol{\gamma}_j)$ obtained from the inner optimization given the current $\boldsymbol{\gamma}_j$. Observing that the dependence of $\mathbf{c}_j^*(\boldsymbol{\gamma}_j)$ on $\boldsymbol{\gamma}_j$ is implicit and possibly complicated, we propose to linearize the likelihood component in

(2.4) and transform the optimization to finding the maximum likelihood estimate of a generalized linear model. The solution can then be readily obtained by the iteratively reweighted least squares (IRLS), see Wood (2017) for more detail.

Let $\widetilde{\boldsymbol{\gamma}}_j$ be the most recent update of $\boldsymbol{\gamma}_j$. First, we linearize the $\mathbf{c}_j^*(\boldsymbol{\gamma}_j)$ at $\widetilde{\boldsymbol{\gamma}}_j$ which,

$$\mathbf{c}_j^*(\boldsymbol{\gamma}_j) \approx \mathbf{c}_j^*(\widetilde{\boldsymbol{\gamma}}_j) + \left.\frac{\partial \mathbf{c}_j^*(\boldsymbol{\gamma}_j)}{\partial \boldsymbol{\gamma}_j^{\top}}\right|_{\widetilde{\boldsymbol{\gamma}}_j} (\boldsymbol{\gamma}_j - \widetilde{\boldsymbol{\gamma}}_j), \tag{3.1}$$

where the derivative $\partial \mathbf{c}_j^*/\partial \boldsymbol{\gamma}_j$ is explicitly derived using the implicit function theorem in Appendix A. Hence, $\widehat{\theta}_j(t;\boldsymbol{\gamma}_j)$ in its basis expansion form can be approximated by a linear function of $\boldsymbol{\gamma}_j$. As a result, the outer objective function (2.4) now becomes a penalized likelihood function of a generalized linear model. Second, we apply the IRLS and update our estimate of $\boldsymbol{\gamma}_j$. Let $\widetilde{\theta}_j(t) = \widehat{\theta}_j(t;\widetilde{\boldsymbol{\gamma}}_j)$ be latent process fit given the structural parameter $\widetilde{\boldsymbol{\gamma}}_j$. Based on the theory of generalized linear models, the observation Y_j according to the latent process $\widetilde{\theta}_j(t)$ admits properties of $\mathrm{E}(Y_j|\widetilde{\theta}_j(t)) = b'(\widetilde{\theta}_j(t)) = \widetilde{\mu}_j(t)$ and $\mathrm{var}(Y_j|\widetilde{\theta}_j(t)) = b''(\widetilde{\theta}_j(t))a(\phi) = \widetilde{v}_j(t)a(\phi)$, where functions a, b and parameter ϕ follow from the exponential family specification. Write $\widetilde{u}_{ij} = -y_{ij} + b'(\widetilde{\theta}_j(t_i)) = -y_{ij} + \widetilde{\mu}_j(t_i)$ and $\widetilde{w}_{ij} = b''(\widetilde{\theta}_j(t_i)) = \widetilde{v}_j(t_i)$. The IRLS algorithm applies a quadratic approximation to the log-likelihood, that is, at $\theta_j = \widetilde{\theta}_j$,

$$-y_{ij}\widehat{\theta}_j(t_i;\boldsymbol{\gamma}_j) + b(\widehat{\theta}_j(t_i;\boldsymbol{\gamma}_j)) \approx \frac{1}{2}\widetilde{w}_{ij}\left\{\widetilde{y}_{ij} - \widehat{\theta}_j(t_i;\boldsymbol{\gamma}_j)\right\}^2 + C_{ij},$$

where $\widetilde{y}_{ij} = \widetilde{\theta}_j(t_i) - \widetilde{u}_{ij}/\widetilde{w}_{ij}$ and C_{ij} is independent of $\widetilde{\theta}_j(t_i)$. In conjunction with the linear approximation of $\theta_j(t_i;\boldsymbol{\gamma}_j)$, it amounts to solving a penalized linear least squares to update the estimate for structural parameter $\boldsymbol{\gamma}_j$. Efficient algorithms are available for different sparsity penalty choices PEN$(\cdot)$.

3.3. Tuning parameter selection

There are two tuning parameters involved in our profiling procedure, which jointly affect the algorithmic performance. On the one hand, λ_θ in the inner optimization controls the amount of regularization regarding the differential equations. We define the aggregated ODE fidelity criterion as

$$\sum_{j=1}^{p}\int_0^1 \left\{\theta_j'(t) - \gamma_{j0} - \sum_{k=1}^{p}\gamma_{jk}\theta_k(t)\right\}^2 \mathrm{d}t. \tag{3.2}$$

Small λ_θ makes optimizing $H_j(\boldsymbol{\gamma}_j)$ with respect to $\boldsymbol{\gamma}_j$ more robust to initial guesses, but yields bad approximations to ODE solutions. Large λ_θ gives rise to a difficult optimization problem where $H_j(\boldsymbol{\gamma}_j)$ is usually not convex and can have many local optima (Ramsay et al., 2007; Qi and Zhao, 2010; Carey and Ramsay, 2021). On the other hand, λ_γ in the outer optimization induces a sparse network structure for latent processes with better interpretation, and existing methods such as information criteria can be adopted for tuning. Based on the above discussion, we propose to fix λ_γ in the outer optimization first, iteratively select a proper λ_θ in the inner optimization, and then determine the best λ_γ via the Bayesian information criterion. In detail, suppose we choose λ_γ from a sequence of candidate values. Then, we initialize λ_θ with a small value and moderately increase it via an iterative scheme. At each iteration, $\widehat{\boldsymbol{\theta}}$ and $\widehat{\Gamma}$ are repeatedly estimated for the current λ_θ, which are then used as initial values in the fitting procedure with the next larger λ_θ. The iterative scheme stops when the estimated ODE parameters converge, and thus λ_θ is decided. The change of the estimated ODE parameters should be small when there is only a moderate increase in λ_θ. Therefore, with a conservatively increasing sequence of λ_θ, every estimated $\widehat{\Gamma}$ is much likely to be a proper initialization for the next iteration. Details of the iterative selecting scheme for λ_θ given a fixed λ_γ are as follows.

(1) Start with a small positive $\lambda_\theta^{(0)}$. Choose $\Delta^{(0)}$ as an initial incremental factor.

(2) At the uth iteration where $u \geq 0$, obtain the fitted latent processes $\widehat{\boldsymbol{\theta}}^{(u)}$ and $\widehat{\Gamma}^{(u)}$ via our profiling procedure, and evaluate the ODE fidelity (3.2) based on the estimates.

   (a) If the absolute percentage of change in the ODE fidelity (3.2) is below a threshold constant, then we update $\lambda_\theta^{(u+1)} = \lambda_\theta^{(u)} \times \Delta^{(u)}$.

   (b) Otherwise, we need to downsize the incremental factor, for example, set $\Delta^{(u)} = \Delta^{(u-1)}/2$, which ensures that the ODE fidelity (3.2) varies little among iterations.

(3) When the successive ODE parameter estimates are closed enough, we stop iteration; otherwise, repeat previous steps.

Our iterative tuning strategy treats λ_θ as a function of λ_γ. Hence, after λ_θ is selected for each fixed λ_γ from a sequence of candidate values, we can evaluate the following BIC and choose the best λ_γ,

$$\mathrm{BIC}(\lambda_\gamma) = -\frac{1}{np}\sum_{i=1}^{n}\sum_{j=1}^{p}\left\{y_{ij}\widehat{\theta}_j(t_i;\lambda_\gamma) - b(\theta_j(t_i;\lambda_\gamma))\right\} + k(\lambda)\log(n),$$

where $\widehat{\theta}_j(t;\lambda_\gamma)$ emphasizes the dependence on λ_γ, and $k(\lambda_\gamma)$ denotes the number of non-zero elements in the resultant ODE parameter $\widehat{\Gamma}(\lambda_\gamma)$.

3.4. Global convergence

Suppose that the estimated latent process $\widehat{\theta}_j(t;\boldsymbol{\gamma}_j)$ from the inner optimization is a smooth function of $\boldsymbol{\gamma}_j$, where $j=1,\ldots,p$. Let $H(\Gamma)=\sum_{j=1}^{p}H_j(\boldsymbol{\gamma}_j)$ be the objective function in the outer optimization for a given tuning parameter λ_γ, where $\Gamma=(\boldsymbol{\gamma}_1,\ldots,\boldsymbol{\gamma}_p)$. Algorithm 1 is essentially a block coordinate descent method because it minimizes $H(\Gamma)$ by iteratively updating $\boldsymbol{\gamma}_j$. Write $H_j(\boldsymbol{\gamma}_j)=\ell_j(\boldsymbol{\gamma}_j)+\mathrm{PEN}_{\lambda_\gamma}(\boldsymbol{\gamma}_j)$, where $\ell_j(\boldsymbol{\gamma}_j)$ is the likelihood term and $\mathrm{PEN}_{\lambda_\gamma}(\boldsymbol{\gamma}_j)$ is assumed to be convex. As described in Section 3.2, the outer optimization is equivalent to updating $\boldsymbol{\gamma}_j$ to $\boldsymbol{\gamma}_j+\mathbf{d}_j(\boldsymbol{\gamma}_j)$, where the descent direction $\mathbf{d}_j(\boldsymbol{\gamma}_j)$ is the solution to

$$\min_{\mathbf{d}} \nabla\ell_j(\boldsymbol{\gamma}_j)^\top\mathbf{d} + \frac{1}{2}\mathbf{d}^\top Q_j(\boldsymbol{\gamma}_j)\mathbf{d} + \mathrm{PEN}_{\lambda_\gamma}(\boldsymbol{\gamma}_j+\mathbf{d}),$$

where $\nabla\ell_j(\boldsymbol{\gamma}_j)$ is the gradient of $\ell_j(\boldsymbol{\gamma}_j)$ and

$$Q_j(\boldsymbol{\gamma}_j) = \frac{1}{n}\sum_{i=1}^{n}\left\{b''(\widehat{\theta}_j(t_i;\boldsymbol{\gamma}_j))\frac{\partial\widehat{\theta}_j(t_i;\boldsymbol{\gamma}_j)}{\partial\boldsymbol{\gamma}_j}\left(\frac{\partial\widehat{\theta}_j(t_i;\boldsymbol{\gamma}_j)}{\partial\boldsymbol{\gamma}_j}\right)^\top\right\}$$

is a positive definite matrix approximating the Hessian $\nabla^2\ell_j(\boldsymbol{\gamma}_j)$.

We follow Tseng and Yun (2009) to establish the global convergence. Because the actual value of the Hessian $\nabla^2\ell_j(\boldsymbol{\gamma}_j)$ is identical to its expected value under canonical links (McCullagh and Nelder, 1989), the IRLS method described in Section 3.2 remains the same when the Hessian is replaced by the expected Hessian. Then it follows from Lemma S1 in the Supplementary Material that

$$H_j(\boldsymbol{\gamma}_j+\mathbf{d}_j(\boldsymbol{\gamma}_j)) - H_j(\boldsymbol{\gamma}_j) \le -\mathbf{d}_j^\top(\boldsymbol{\gamma}_j)\left[Q_j(\boldsymbol{\gamma}_j) - \frac{1}{2}\mathrm{E}\{\nabla^2\ell_j(\boldsymbol{\gamma}_j)\}\right]\mathbf{d}_j(\boldsymbol{\gamma}_j) + o(\|\mathbf{d}_j(\boldsymbol{\gamma}_j)\|^2). \quad (3.3)$$

Some algebra yields that

$$Q_j(\boldsymbol{\gamma}_j) - \frac{1}{2}\mathrm{E}\{\nabla^2\ell_j(\boldsymbol{\gamma}_j)\} = \frac{1}{2}Q_j(\boldsymbol{\gamma}_j) + \frac{1}{2n}\sum_{i=1}^{n}\left\{b'(\theta_j^*(t_i)) - b'(\widehat{\theta}_j(t_i,\boldsymbol{\gamma}_j))\right\}\frac{\partial^2\widehat{\theta}_j(t_i,\boldsymbol{\gamma}_j)}{\partial\boldsymbol{\gamma}_j\partial\boldsymbol{\gamma}_j^\top},$$

where $\theta_j^*(t)$ is the true latent process. The above matrix is positive definite because $b'(\cdot)$ is continuous, provided that $\widehat{\theta}_j(t_i,\boldsymbol{\gamma}_j)$ is sufficiently close to the truth. It follows from (3.3) that $H_j(\boldsymbol{\gamma}_j)$ decreases along the iterations and will eventually converge because it is lower-bounded. Moreover, the sequence of descent directions converges to zero due to (3.3). According to Theorem 1(e) and Lemma 2 of Tseng and Yun (2009), every cluster point of the iterative estimates by Algorithm 1 exhibits exact zero descent direction, which implies it is indeed a stationary point of $H(\Gamma)$.

Finally, we remark that the above analysis cannot be directly applied to a non-convex $\mathrm{PEN}_{\lambda_\gamma}(\cdot)$ such as the SCAD penalty. However, the non-convex penalty can be numerically approximated by local linear or quadratic functions (Fan et al., 2020). We would anticipate a similar convergence result but with more involved technical details, which is not pursued in this paper.

4. Two-step collocation methods for non-Gaussian data

Collocation methods have been exploited for both parameter estimation and network reconstruction for various ODE models. In this section, we extend the popular two-step collocation method for high-dimensional linear ODE with non-Gaussian observations. In the large literature on collocation, Varah (1982); Ramsay et al. (2007); Dattner and Klaassen (2015), and Wu et al. (2019) consider the linear case while recently the nonparametric additive structure is investigated by Henderson and Michailidis (2014); Wu et al. (2014) and Chen et al. (2017). Most existing methods are proposed for Gaussian observations and adopt the least square loss function for estimation. In the following, we present two versions of the two-step collocation method for high-dimensional ODE models with non-Gaussian observations: the vanilla collocation based on Varah (1982) and an extension from graph reconstruction via additive differential equations (GRADE) by Chen et al. (2017).

The vanilla two-step method first fits smoothing estimates $\widehat{\boldsymbol{\theta}}(t)$ to the latent processes with maximum likelihood estimation, and then obtain the structural parameter $\boldsymbol{\gamma}$ with the estimated processes and their derivatives plugged in. The procedure solves the following optimization problems,

$$\widehat{\boldsymbol{\gamma}}_j = \arg\min_{\gamma_{j0},\boldsymbol{\gamma}_j} \int_0^1 \left| \frac{\mathrm{d}\widehat{\theta}_j(t)}{\mathrm{d}t} - \gamma_{j0} - \sum_{k=1}^{p} \gamma_{jk}\widehat{\theta}_k(t) \right|^2 \mathrm{d}t + \mathrm{PEN}_{\lambda_\gamma}(\boldsymbol{\gamma}_j), \tag{4.1}$$

with

$$\widehat{\theta}_j(t) = \arg\min_{\theta\in\mathcal{H}} -\frac{1}{n}\sum_{i=1}^{n}\{y_{ij}\theta(t_i) - b(\theta(t_i))\}, \quad 1 \le j \le p, \tag{4.2}$$

where $\mathcal{H}$ is a proper reproducing kernel Hilbert space, and the exponential family smoothing splines can be adopted (Wahba et al., 1995; Gu, 2013; Ma et al., 2017). The performance of the vanilla two-step collocation method relies on the estimation accuracy of $\widehat{\theta_j}(t)$ and its derivatives. Although statistical convergence has been established, it is in practice hard to tune the smoothing procedure to achieve the optimality (Liang and Wu, 2008; Brunel et al., 2014).

Another extension is based on the GRADE method (Chen et al., 2017). It avoids the derivative estimation issue in the vanilla collocation method, and instead considers the ODE fidelity term in its integral form. Similar to the vanilla two-step method, the GRADE method first obtains the smoothing estimates of latent processes from observations as in (4.2). Using integrated basis functions $\widehat{\Theta}_j(t) = \int_0^t \widehat{\theta}_j(t)\,\mathrm{d}t$, $j = 1, \ldots, p$, one can express

$$\widetilde{\theta}_j(t) = C_{j0} + \gamma_{j0}\, t + \sum_{k=1}^{p} \gamma_{jk}\widehat{\Theta}_k(t),$$

according to the integrated differential equations. Finally, we solve the following optimization problems to obtain

$$\widehat{\boldsymbol{\gamma}}_j = \arg\min_{C_{j0},\gamma_{j0},\boldsymbol{\gamma}_j} \frac{1}{n}\sum_{i=1}^{n}\left\{y_{ij}\,\widetilde{\theta}_j(t_i) - b(\widetilde{\theta}_j(t_i))\right\} + \mathrm{PEN}_{\lambda_\gamma}(\boldsymbol{\gamma}_j). \tag{4.3}$$

The GRADE method is initially developed for nonparametric additive ODE models and naturally adapts to the linear case. The use of an integrated form of ODE facilitates investigating the asymptotic behavior of the estimator and enhancing its robustness to the smoothing effect in the first step (Dattner and Klaassen, 2015; Chen et al., 2017). Both the two-step collocation methods proposed in this section involve maximizing the likelihood function for exponential family distributions, which can be efficiently solved with the iteratively reweighted least squares technique as in Section 3.2.

We compare the two-step collocation methods with the high-dimensional generalized profiling (HDGP) procedure in Section 5. For process and derivative estimation, since HDGP balances both the data and ODE fidelities, it usually results in reasonable fits and more accurate ODE parameter estimates due to the more accurate derivatives. For sparse structure identification, GRADE achieves the best accuracy, which is consistent with the motivation of GRADE for network reconstruction (Chen et al., 2017). In summary, HDGP is a better choice for process fitting and ODE parameter estimation, while GRADE excels in sparse structure identification.

5. Simulation studies

This section compares the empirical performance of three dynamical modeling approaches: the high-dimensional generalized profiling (HDGP) procedure and the two-step collocation methods proposed in Section 4, namely the GRADE and the vanilla two-step method, respectively.

Consider the ODE system studied by Chen et al. (2017) which consists of eight processes in four pairs, for $k = 1, \ldots, 4$,

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

It is clear that the ODE solutions take the form of sine and cosine functions with varying frequencies, whereas no interaction exists across pairs. For the kth pair, the initial state is $\sin(y_k)$ and $\cos(y_k)$, where y_k is sampled from $N(0, 1)$. The latent processes $\boldsymbol{\theta}(t) = (\theta_1(t), \ldots, \theta_8(t))^\top$ described by the above ordinary differential equations are used to generate observations from Gaussian, Poisson and Bernoulli distributions. Denote by $t_1, \ldots, t_n$ time points from $[0, 1]$. For Gaussian distribution, y_{ij} is sampled from $N(\theta_j(t_i), \sigma^2)$ with known variance σ^2, and the sample size n for each process is set to be 100 and 500. For Poisson distribution, we draw 500 and 1000 samples from Poisson$(\lambda_j(t_i))$ where the intensity process $\lambda_j(t) = \exp\{\theta_j(t)\}$. For Bernoulli distribution, 1500 and 2500 samples are generated with probability of success $p_j(t) = \exp\{\theta_j(t)\}/[1 + \exp\{\theta_j(t)\}]$. Sample sizes for Poisson and Bernoulli distributions are larger than Gaussian, as in those cases more observations are generally required to ensure reasonable estimates according to the theory of generalized linear model.

We use the smoothing spline fitting as an initialization for the profiling procedure, which also corresponds to the first stage of two-step collocation methods. The order of B-spline functions in HDGP is set as 6, and the number of knots is half of that of time points. Both HDGP and GRADE require numerical integration to evaluate ODE fidelity and integrated basis representations, respectively. For sparsity penalty choices, we consider the Lasso penalty $\mathrm{PEN}_{\lambda_\gamma}(\boldsymbol{\gamma}_j) = \lambda_\gamma \|\boldsymbol{\gamma}_j\|_1$ and the SCAD penalty $\mathrm{PEN}_{\lambda_\gamma}(\boldsymbol{\gamma}_j) = \sum_{k=1}^{p} p_{\lambda_\gamma}(|\gamma_{jk}|)$, where the function $p_\lambda(\cdot)$ is defined on $[0, \infty)$ as

$$p_\lambda(u) = \begin{cases} \lambda u, & \text{if } 0 \le u \le \lambda \\ -(u^2 - 2a\lambda u + \lambda^2)/(a-1), & \text{if } \lambda < u < a\lambda \\ (a+1)\lambda^2/2 & \text{if } u \ge a\lambda, \end{cases}$$

and a suggested value for a is 3.7 according to Fan and Li (2001). Algorithmic convergence is demonstrated when the difference between successive ODE parameter estimates is small enough. It works well for two-step collocation methods. However, due to the complex interaction between inner and outer optimizations, HDGP may not yield sparse ODE parameter estimates at the declaration of convergence. To address this numerical issue, we manually set ODE parameter estimates below a constant threshold as zero. Based on our empirical studies, a recommended value for the threshold is the root-mean-square of the initial estimate $\widehat{\Gamma}$ multiplied by a factor 0.01.

Simulation results are evaluated using three types of criteria. The first two criteria concern about process and derivative estimates, which are evaluated by the mean squared errors (MSE) of $\boldsymbol{\theta}(t)$ and $\boldsymbol{\theta}'(t)$,

$$\mathrm{MSE}(\widehat{\boldsymbol{\theta}}(t)) = \frac{1}{np}\sum_{j=1}^{p}\sum_{i=1}^{n}\left\{\widehat{\theta}_j(t_i) - \theta_j(t_i)\right\}^2,$$

$$\mathrm{MSE}(\widehat{\boldsymbol{\theta}}'(t)) = \frac{1}{np}\sum_{j=1}^{p}\sum_{i=1}^{n}\left\{\widehat{\theta}_j'(t_i) - \theta_j'(t_i)\right\}^2.$$

Second, we measure how well the structural parameters are estimated by their root-mean-square error (RMSE). Third, true positive rate (TPR) and false positive rate (FPR) are used to quantify how well the sparse structure is identified, where we refer to non-zero structural parameters as positive cases and otherwise as negative cases.

Table 1 displays the averaged evaluations over 50 repeated experiments using the Lasso penalty, while the true positive rates are omitted because they are all equal to one for all three methods. Under each simulation set-up, increasing the number of observations always leads to reduced errors and tighter confidence intervals in terms of the process fit and the parameter estimation. For process and derivative fitting, the smoothing splines method, as the first stage of two-step collocation methods, often produces accurate estimates of the latent process itself, but is less efficient in the derivative fitting. In contrast, the inner optimization of HDGP balances the data and ODE fidelities, resulting in reasonable process fitting and improved derivative fitting. For ODE parameter estimation, HDGP delivers the smallest error due to the more accurate derivatives. Interestingly, GRADE has much worse performance than the other two under this criterion. One partial reason is that GRADE only uses structural parameters in the integrated basis representation (4.3) instead of the explicit form of differential equations. For sparse structure identification, GRADE achieves the best accuracy, as it discovers all non-zero structural parameters with the fewest false positives. It is consistent with the motivation of GRADE for network reconstruction (Chen et al., 2017). In summary, HDGP is a better choice for process fitting and ODE parameter estimation, while GRADE excels in sparse structure identification.

We next investigate the effects of different noise levels and choices of sparse penalty. Under the above Gaussian set-up with 500 observations for each process. The signal-to-noise ratio (SNR) is defined as the ratio between the sample standard deviation of $\{\theta_j(t_i)\}_{i=1}^n$ and the noise standard deviation σ. We set the signal-to-noise ratio as 3, 10, 30, and infinity, where the infinite ratio means that no noise is added. Both Lasso and SCAD penalties are considered. Fig. 1 presents the performance evaluations over 50 repeated experiments. In general, all methods perform better over all criteria when the signal-to-noise ratio increases. The top row of Fig. 1 corresponding to the Lasso penalty provides the consistent result as in Table 1, which indicates that HDGP has a comparable process fit and better derivative estimation, especially when the noise level is low. Moreover, HDGP performs the best for estimating structural parameters, while the vanilla two-step method also provides satisfactory results. In contrast, even when there is no noise, the bias of ODE parameter estimates by GRADE is still large and RMSE is almost constant. For sparse structure identification, GRADE outperforms the other methods under a wide range of noise levels. HDGP and the vanilla two-step method only have high accuracy when the signal level is high. The bottom row of Fig. 1 displays simulation results when the SCAD penalty is used for inducing sparsity for the ODE system. Compared with the results with Lasso, overall performances in process, derivative, and ODE parameter estimations are improved mainly due to the unbiasedness property of SCAD penalty (Fan and Li, 2001). More interestingly, the poor performance of GRADE in ODE parameter estimation is greatly enhanced, and now it delivers comparable estimation results as the other two. Due to the oracle property enjoyed by the SCAD penalty (Fan and Li, 2001), we recommend it for better performance in parameter estimation.

Table 1
Performance of HDGP, GRADE and the vanilla two-step method evaluated based on the process estimates (MSE($\widehat{\boldsymbol{\theta}}(t)$)), derivative estimates (MSE($\widehat{\boldsymbol{\theta}}'(t)$)), non-zero parameter estimation (RMSE), and sparse structure estimates (FPR). The 95% confidence intervals are given in parentheses.

	N	Method	MSE ($\widehat{\boldsymbol{\theta}}(t)$)	MSE ($\widehat{\boldsymbol{\theta}}'(t)$)	RMSE ($\widehat{\Gamma}$)	FPR
Gaussian	100	HDGP	0.011 (0.0097,0.0124)	**3.01** (2.48, 3.53)	**0.58** (0.52,0.66)	0.44 (0.43,0.46)
		GRADE	**0.005** (0.0042,0.0049)	5.23 (4.67, 5.87)	2.97 (2.89,3.05)	**0.00** (-,-)
		vanilla	**0.005** (0.0043,0.0049)	5.23 (4.74, 5.93)	0.62 (0.54,0.72)	0.89 (0.84,0.92)
	500	HDGP	0.002 (0.0017,0.0023)	**0.48** (0.41, 0.57)	**0.28** (0.25,0.30)	0.44 (0.42,0.45)
		GRADE	**0.001** (0.0010,0.0011)	1.82 (1.75, 1.88)	0.84 (0.82,0.85)	**0.01** (0.01,0.02)
		vanilla	**0.001** (0.0010,0.0011)	1.82 (1.75, 1.88)	0.34 (0.32,0.37)	0.67 (0.64,0.71)
Poisson	500	HDGP	0.024 (0.0222,0.0259)	**6.24** (5.60, 6.93)	**1.70** (1.54,1.91)	0.58 (0.56,0.61)
		GRADE	0.024 (0.0232,0.0252)	12.27 (11.66,13.06)	2.03 (1.86,2.19)	**0.37** (0.34,0.41)
		vanilla	0.024 (0.0232,0.0252)	12.27 (11.63,13.00)	1.86 (1.70,2.07)	0.98 (0.97,0.98)
	1000	HDGP	**0.011** (0.0105,0.0121)	**2.70** (2.43, 2.97)	**1.04** (0.94,1.14)	0.57 (0.55,0.59)
		GRADE	0.013 (0.0128,0.0141)	8.20 (7.70, 8.73)	1.41 (1.31,1.53)	**0.32** (0.28,0.36)
		vanilla	0.013 (0.0127,0.0141)	8.20 (7.71, 8.67)	1.18 (1.09,1.29)	0.97 (0.96,0.97)
Bernoulli	1500	HDGP	0.031 (0.0260,0.0357)	**8.18** (6.74, 9.73)	**1.77** (1.48,1.97)	0.54 (0.52,0.58)
		GRADE	0.031 (0.0277,0.0333)	15.97 (14.77,16.99)	3.32 (3.01,3.61)	**0.24** (0.21,0.28)
		vanilla	0.032 (0.0285,0.0376)	22.28 (17.21,32.11)	2.28 (1.70,3.22)	0.94 (0.90,0.96)
	2500	HDGP	**0.019** (0.0169,0.0216)	**5.05** (4.33, 6.03)	**1.55** (1.39,1.74)	0.59 (0.55,0.62)
		GRADE	0.020 (0.0193,0.0210)	12.71 (11.91,13.67)	2.57 (2.34,2.79)	**0.20** (0.16,0.24)
		vanilla	0.020 (0.0193,0.0211)	12.71 (11.88,13.69)	1.67 (1.46,1.87)	0.98 (0.97,0.99)

Table 2
Three categories of keywords selected for the analysis of Google Trends data.

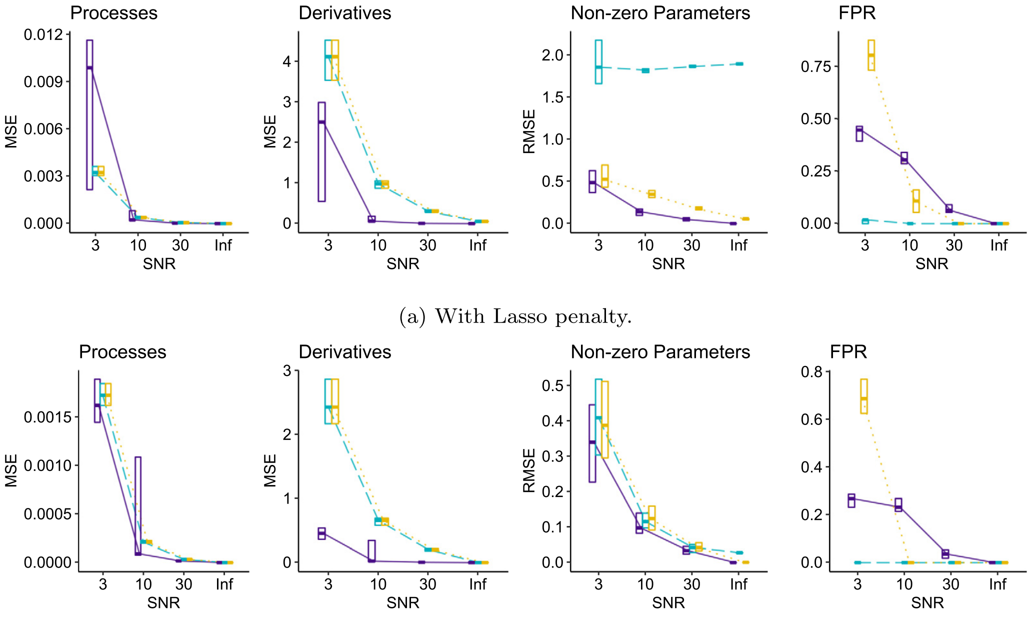
Category	Keyword
COVID-19 related	coronavirus, mask, quarantine, vaccine, WHO (5 words)
Countries or districts	Africa, Antarctica, Arctic, Australia, Brazil, Canada, China, India, Iran, Italy, Japan, Russia, the United States (13 words)
Noise words	cat, cloud, desert, dog, game, sun (6 words)

6. Real data analysis

6.1. Google trends data analysis

Google Trends provides a publicly accessible online portal to analyze the popularity of search queries. In this study, we attempt to apply our method to model the interactions among a number of trending keywords during the recent pandemic of Coronavirus disease 2019 (COVID-19). In Table 2, we list 24 keywords and cluster them into three categories. The first category consists of five keywords about specific terminologies such as mask and quarantine. The second category includes not only the countries with the most confirmed cases as of January 2021, such as the United States, India, and Brazil but also the districts like Antarctica, which is the last continent to report confirmed cases due to the remoteness and sparse population. We also include the last category of noise keywords with no apparent relationship to the pandemic.

The Google Trends data used in our study cover the range from January 20 to September 20 in 2020. The keyword popularity is measured by an integer index calculated by normalizing and rounding the keyword count in an unbiased



(a) With Lasso penalty.

(b) With SCAD penalty.

Fig. 1. Performance of HDGP (purple solid), GRADE (blue dashed), and the vanilla two-step method (yellow dotted) for Gaussian observations at different noise levels. The boxes identify the medians and the quartiles of each criterion for 50 repeated experiments. Top and bottom rows correspond to Lasso and SCAD penalties, respectively. (For interpretation of the colors in the figure(s), the reader is referred to the web version of this article.)

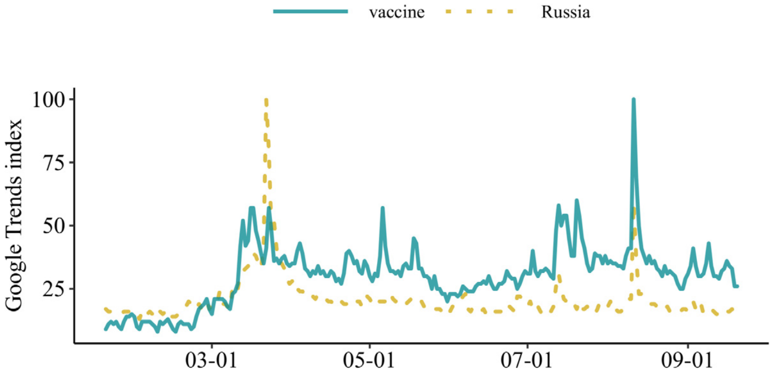
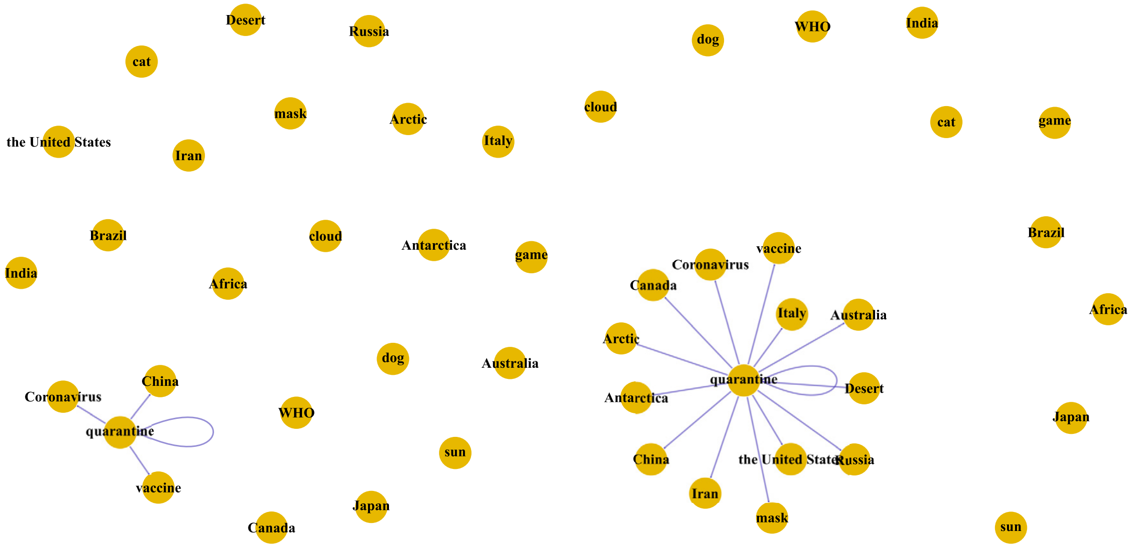


Fig. 2. Daily Google Trends indices of keyword *Russia* and *vaccine* from January to September 2020.

searching requests sample. We observe that the daily trend indices have several sharp peaks, see Fig. 2 for an illustrative example. Direct modeling for the mean trends will result in abrupt high values near the peaks and undersmooth other relatively flat regions. Therefore, it is more appropriate to assume the indices follow Poisson distributions, and we apply the proposed method to model the latent processes of intensity parameters with ODEs.

To better exhibit different stages of the pandemic, we consider three time periods: from January 20 to March 19, from March 20 to June 19, and from June 20 to September 20. For each period, our method is applied to fit the trending processes with a series of sparsity parameter λ_γ's. Figs. 3a and 3b display two networks with different sparsity parameters in the first period (from January to March). Keyword *quarantine* has the highest degree in both networks. During the COVID-19 pandemic, quarantines or self-quarantines are enacted by multiple governmental actors to prevent the rapid spread of the virus. It is of no surprise to become the top-ranked trending keyword. The other three keywords in Fig. 3a are *coronavirus*, *China* and *vaccine*, which stand for the virus's name, the country where the first case was identified, and the immunization method. The top four keywords represent the major trending focus at the early stage of the pandemic. In Fig. 3b, more affected countries such as Australia, Italy, and the United States, are involved when the sparsity parameter is decreased. In contrast, noise keywords are isolated in both networks, indicating no connection to the trending topics. More interestingly, we investigate the evolution of network structure for the trending keywords along the progression of the COVID-19 pandemic. Table 3 lists the top four keywords in three time periods where the keyword with the highest



(a) $\lambda_\gamma = 10^{-0.5}$

(b) $\lambda_\gamma = 10^{-1}$

Fig. 3. Recovered networks of the trending keywords during the first period (from January 20 to March 19) with different values of sparsity parameters.

Table 3
Top four keywords in the recovered networks during three periods. The keyword of the highest degree is in boldface.

Period	Keywords
January 20 – March 19	**quarantine**, China, coronavirus, vaccine
March 20 – June 19	**Italy**, China, Iran, Russia
June 20 – September 20	**coronavirus**, the United States, vaccine, mask

degree is in boldface. From the first period to the second, the keyword *Italy* emerges as the new top word. According to the WHO report, on March 19, Italy overtook China as the country with the most reported deaths, and announced the national lockdown in March. Turning to the third period, *China* and *Italy* drop out of the top list. Both countries had successfully slowed down the domestic infections and reduced daily new cases significantly. As preventive measures including wearing face masks in public are advised and several promising vaccines are being developed, *mask* and *vaccine* are among the top trending keywords.

6.2. Analysis of stock price change directions

In the year 2020, the stock market experienced enormous volatility due to the coronavirus pandemic. Many companies have suffered massive price drops, while others have witnessed substantial increases. We collect the stock price indices for 40 companies during 251 trading days spanning from January 1 to December 30, 2020. Our goal in this study is to characterize the change direction patterns of stock prices, taking into account the dynamic interactions among the stocks. To this end, the original price indices are coded as binary data to denote an increase or decrease. We group the companies into eight categories based on the Global Industry Classification Standard. Details are provided in Table 4.

The high-dimensional ODE system is built up for the latent success probability processes. Our sparsity tuning procedure leads to $\lambda_\gamma = 10^{-2.1}$ and the fitted model achieves an ODE fidelity below 10^{-6}. Fig. 4 displays the fitted probabilities of a daily stock price increase for all categories. We notice some interesting results from the result. First, all categories have the low fitted probabilities around March. It corresponds to the 2020 stock market crash, during which multiple circuit breakers were triggered on fears of the COVID-19 coronavirus. Since the crash, some sectors recovered and re-entered a bull market through December. Online retail companies made huge profits as health concerns changed customers' shopping habits. Information technology companies benefited from the growing demands for information services and electronics devices. For example, the shifts towards remote working had raised the number of Zoom's daily users to an unprecedented one. In contrast, sectors like energy, hotels, and air transportation experienced the most severe hit by the COVID-19 pandemic. Although there were signs of recovery in the fourth quarter, these industries are still under the tremendous impact of the COVID-19 recession.

Table 4
Companies selected in eight categories for stock price data analysis.

Group	Category	Companies
1	Information Technology	Adobe, Apple, Microsoft, Salesforce, Zoom
2	Electric Vehicle	BYD, Kandi, Nio, Tesla, Workhorse
3	Pharmaceutical	AbbVie, Eli lilly, Moderna, Novartis, Pfizer
4	Consumer Services & Retail	Ascena, J. C. Penney, Kohl's, Macy's, Nordstrom
5	Online Retail Shopping	Amazon, Best Buy, Target, Walmart, Wayfair
6	Hotels	Hilton, Marriott, Wyndham, Wynn, Park
7	Air Transportation	Boeing, Airbus, Delta Air Lines, Southwest Airlines, United Airline
8	Energy	Chevron, Conocophillips, Exxon Mobil, Schlumberger, Valero Energy

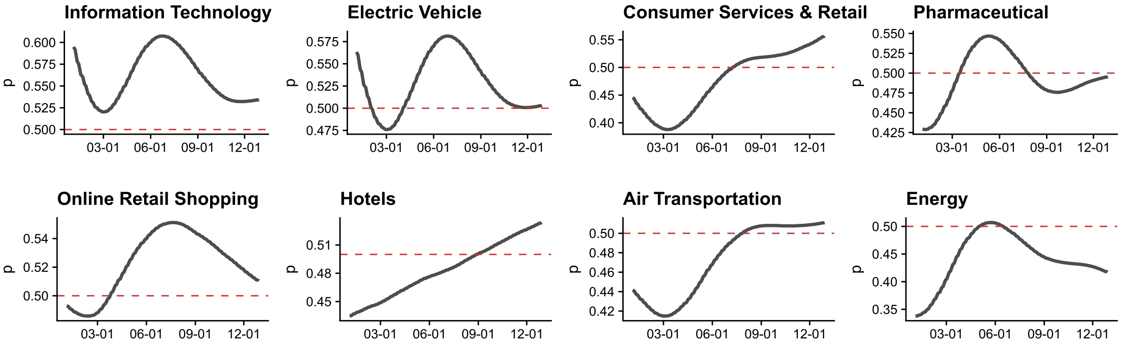


Fig. 4. The fitted probability processes of daily price increase for the eight categories. The red dashed line denotes $p = 0.5$.

6.3. Analysis of yeast cell cycle-regulated genes

The cell cycle is a fundamental biological process consisting of cell growth, duplication of genetic information, distribution of chromosomes, and cell division (Cho et al., 1998). Spellman et al. (1998) analyzed the expression levels of 6,178 yeast genes at 7-minute intervals for 119 minutes. The experiments were carried out in the cell cultures with three independent synchronization methods. A score was calculated for each gene to indicate their similarities to those cell-cycle regulated genes already known. Due to missingness in data, we choose 72 out of 93 genes identified by Spellman et al. (1998) in the *alpha* factor-based synchronized experiment, and model the dynamic relationship between the mean profiles of these 72 genes using an ODE system under Gaussian assumption for gene expression level. The proposed method is applied to identify the sparse structure of the gene regulatory network. The result is shown in Fig. 5, which excludes 12 isolated genes. This suggests that although those genes get involved in the cell cycle, their regulated transcriptions are not absolutely required. Among the 60 genes in Fig. 5, 116 regulations (i.e., directed edges) are discovered. The average number of regulations for each gene is around three, while more than 80% genes have regulations fewer than five. Genes with high network degrees are identified as central hub nodes. For example, CLN3 (node 1) has the largest number of regulations in Fig. 5. According to the Saccharomyces Genome Database (Cherry et al., 1998), the encoded protein CLN3p is known as a cell cycle regulator and promotes the G1/S transition (Nasmyth, 1993). More interestingly, the positive or negative signs of our estimated structural parameters naturally imply the potential promotion or inhibition between genes, respectively. Our result suggests that CHS1 (node 62) promotes the expression of POL30 (node 12), which regulates DNA replication in the G1 phase. Meanwhile, it suppresses the expression of FAR1 (node 30), which is a CDC28p kinase inhibitor functioning in the G2/M transition.

7. Conclusions and discussion

In this article, we have proposed a new profiling procedure for both parameter estimation and sparse structure identification for high-dimensional linear ODE models with non-Gaussian observations. Our method involves a hierarchical optimization scheme: the inner optimization balances the data fitting and ODE fidelity to improve estimation efficiency, while the outer optimization induces a sparse structure for better model interpretation. Besides, we extend two-step collocation methods to the non-Gaussian observation setting and compare them with the proposed profiling procedure via comprehensive studies.

One limitation of our work is that only the linear ODE system is under consideration. We are aware of the recent development of two-step collocation to sparse additive ODE systems (Henderson and Michailidis, 2014; Wu et al., 2014;

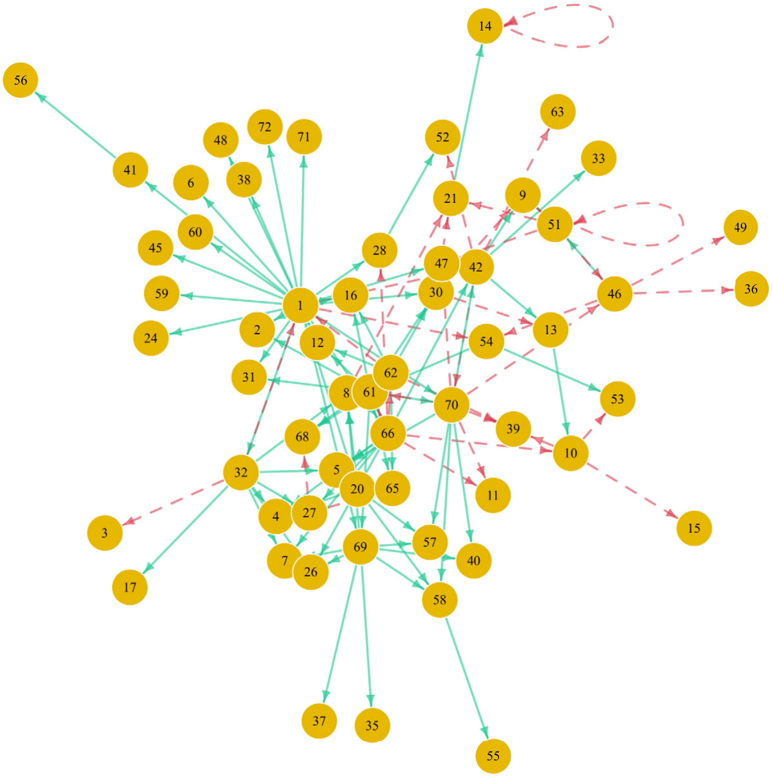


Fig. 5. The recovered network of the yeast cell cycle. Yellow nodes represent genes, and the green-solid or red-dashed edges indicate potential promotion or suppression effects.

Chen et al., 2017) and a more general functional ANOVA extension (Dai and Li, 2021). Although our hierarchical optimization is not restricted to the linear ODE, the extension to nonlinear ODE systems is not straightforward. For instance, a common strategy to handle additive ODE models is to expand the nonlinear components with basis function. However, due to the profiling nature, the range of collocation bases for latent processes needs to be controlled within a compact interval, which may not be easily overcome. Another future research is on the statistical properties such as uniform bound on the approximations to the true solutions, asymptotic normality of the estimators. Despite existing theory established for the standard generalized profiling (Qi and Zhao, 2010), it is still a challenging problem due to high dimensionality, and we leave the systematic study to future work.

Acknowledgements

The authors thank the Editor, the Associate Editor, and the anonymous reviewers for their constructive comments that helped significantly improve this work. Dr. Zhang's research is supported by the National Natural Science Foundation of China under Grants 72101058 and 11690014; and the Science and Technology Commission of Shanghai Municipality under Grant 17JC1420200. Dr. Cao's research is supported by the Discovery grant (RGPIN-2018-06008) from the Natural Sciences and Engineering Research Council of Canada (NSERC).

Appendix A. Derivatives

We provide the analytical expressions of the derivatives used in the computation (Section 3).

Derivatives of G_j in inner optimization

Write $G_j(\theta_j; \boldsymbol{\gamma}_j)$ in the inner optimization

$$G_j = -\frac{1}{n}\sum_{i=1}^{n}\left\{y_{ij}\theta_j(t_i) - b(\theta_j(t_i))\right\} + \lambda_\theta \int_{\mathcal{T}} \left\{\theta_j'(t) - \gamma_{j0} - \sum_{k=1}^{p}\gamma_{jk}\,\theta_k(t)\right\}^2 \mathrm{d}t,$$

then the first derivative is

$$\frac{\partial G_j}{\partial \mathbf{c}_j} = -\frac{1}{n}\sum_{i=1}^{n}\{y_{ij}\mathbf{h}(t_i) - b'(\theta_j(t_i))\mathbf{h}(t_i)\} + 2\lambda_\theta \int_{\mathcal{T}} \left\{\frac{\mathrm{d}\theta_j(t)}{\mathrm{d}t} - \gamma_{j0} - \sum_{k=1}^{p}\gamma_{jk}\,\theta_k(t)\right\}\left\{\frac{\mathrm{d}\mathbf{h}(t)}{\mathrm{d}t} - \gamma_{jj}\,\mathbf{h}(t)\right\}\mathrm{d}t,$$

and the second derivative is

$$\frac{\partial^2 G_j}{\partial \mathbf{c}_j \partial \mathbf{c}_j^\top} = \frac{1}{n}\sum_{i=1}^{n}\{b''(\theta_j(t_i))\mathbf{h}(t_i)\mathbf{h}(t_i)^\top\} + 2\lambda_\theta \int_{\mathcal{T}} \left\{\frac{\mathrm{d}\mathbf{h}(t)}{\mathrm{d}t} - \gamma_{jj}\mathbf{h}(t)\right\}\left\{\frac{\mathrm{d}\mathbf{h}(t)}{\mathrm{d}t} - \gamma_{jj}\mathbf{h}(t)\right\}^\top \mathrm{d}t.$$

For $k=0$,

$$\frac{\partial^2 G_j}{\partial \mathbf{c}_j \partial \gamma_{j0}} = -2\lambda_\theta \int_{\mathcal{T}} \left\{\frac{\mathrm{d}\mathbf{h}(t)}{\mathrm{d}t} - \gamma_{jj}\mathbf{h}(t)\right\}\mathrm{d}t,$$

for $k=1,\ldots,p$ and $k \neq j$,

$$\frac{\partial^2 G_j}{\partial \mathbf{c}_j \partial \gamma_{jk}} = -2\lambda_\theta \int_{\mathcal{T}} \left\{\frac{\mathrm{d}\mathbf{h}(t)}{\mathrm{d}t} - \gamma_{jj}\mathbf{h}(t)\right\}\theta_k(t)\,\mathrm{d}t,$$

for $k=j$,

$$\frac{\partial^2 G_j}{\partial \mathbf{c}_j \partial \gamma_{jj}} = -2\lambda_\theta \int_{\mathcal{T}} \left\{\frac{\mathrm{d}\mathbf{h}(t)}{\mathrm{d}t} - \gamma_{jj}\mathbf{h}(t)\right\}\theta_j(t)\,\mathrm{d}t - 2\lambda_\theta \int_{\mathcal{T}} \left\{\frac{\mathrm{d}\theta_j(t)}{\mathrm{d}t} - \gamma_{j0} - \sum_{k=1}^{p}\gamma_{jk}\theta_k(t)\right\}\mathbf{h}(t)\,\mathrm{d}t.$$

Derivative of $\mathbf{c}_j^*$ in outer optimization

Write $\mathbf{c}_j^*(\boldsymbol{\gamma}_j)$ as $\mathbf{c}_j^*$ for simplicity. Since G_j has zero-gradient at $\mathbf{c}_j^*$, then

$$\left.\frac{\partial G_j}{\partial \mathbf{c}_j}\right|_{\mathbf{c}_j^*} = 0.$$

Taking the derivative with respect to $\boldsymbol{\gamma}_j$ on both sides, we have

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\gamma}_j^\top}\left(\left.\frac{\partial G_j}{\partial \mathbf{c}_j}\right|_{\mathbf{c}_j^*}\right) = \left.\frac{\partial^2 G_j}{\partial \mathbf{c}_j \partial \boldsymbol{\gamma}_j^\top}\right|_{\mathbf{c}_j^*} + \left(\left.\frac{\partial^2 G_j}{\partial \mathbf{c}_j \partial \mathbf{c}_j^\top}\right|_{\mathbf{c}_j^*}\right)\frac{\partial \mathbf{c}_j^*(\boldsymbol{\gamma}_j)}{\partial \boldsymbol{\gamma}_j^\top} = 0.$$

Suppose $\partial^2 G_j/(\partial \mathbf{c}_j \partial \mathbf{c}_j^\top)|_{\mathbf{c}_j^*}$ is non-singular, we have the following expression of the derivative

$$\frac{\partial \mathbf{c}_j^*(\boldsymbol{\gamma}_j)}{\partial \boldsymbol{\gamma}_j^\top} = -\left(\left.\frac{\partial^2 G_j}{\partial \mathbf{c}_j \partial \mathbf{c}_j^\top}\right|_{\mathbf{c}_j^*}\right)^{-1}\left(\left.\frac{\partial^2 G_j}{\partial \mathbf{c}_j \partial \boldsymbol{\gamma}_j^\top}\right|_{\mathbf{c}_j^*}\right).$$

Both matrices on the right-hand side have been explicitly derived, the derivative of $\mathbf{c}_j^*$ follows.

Appendix B. Supplementary material

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.csda.2022.107483.

References

Brunel, N.J., Clairon, Q., d'Alché Buc, F., 2014. Parametric estimation of ordinary differential equations with orthogonality conditions. J. Am. Stat. Assoc. 109, 173–185.

Cao, J., Ramsay, J.O., 2007. Parameter cascades and profiling in functional data analysis. Comput. Stat. 22, 335–351.

Carey, M., Ramsay, J.O., 2021. Fast stable parameter estimation for linear dynamical systems. Comput. Stat. Data Anal. 156, 107124.

Chen, S., Shojaie, A., Witten, D.M., 2017. Network reconstruction from high-dimensional ordinary differential equations. J. Am. Stat. Assoc. 112, 1697–1707.

Cherry, J.M., Adler, C., Ball, C., Chervitz, S.A., Dwight, S.S., Hester, E.T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., 1998. Sgd: Saccharomyces genome database. Nucleic Acids Res. 26, 73–79.

Cho, R.J., Campbell, M.J., Winzeler, E.A., Steinmetz, L., Conway, A., Wodicka, L., Wolfsberg, T.G., Gabrielian, A.E., Landsman, D., Lockhart, D.J., 1998. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol. Cell 2, 65–73.

Cokus, S.J., Feng, S., Zhang, X., Chen, Z., Merriman, B., Haudenschild, C.D., Pradhan, S., Nelson, S.F., Pellegrini, M., Jacobsen, S.E., 2008. Shotgun bisulphite sequencing of the arabidopsis genome reveals DNA methylation patterning. Nature 452, 215–219.

Dai, X., Li, L., 2021. Kernel ordinary differential equations. J. Am. Stat. Assoc.

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

Dodds, P.S., Harris, K.D., Kloumann, I.M., Bliss, C.A., Danforth, C.M., 2011. Temporal patterns of happiness and information in a global social network: hedonometrics and Twitter. PLoS ONE 6, e26752.

Fan, J., Feng, Y., Wu, Y., 2009. Network exploration via the adaptive lasso and scad penalties. Ann. Appl. Stat. 3, 521–541.

Fan, J., Li, R., 2001. Variable selection via nonconcave penalized likelihood and its oracle properties. J. Am. Stat. Assoc. 96, 1348–1360.

Fan, J., Li, R., Zhang, C.H., Zou, H., 2020. Statistical Foundations of Data Science. Chapman and Hall/CRC.

Gu, C., 2013. Smoothing Spline ANOVA Models, vol. 297, 2nd ed. Springer, New York.

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

Hecker, M., Lambeck, S., Toepfer, S., Van Someren, E., Guthke, R., 2009. Gene regulatory network inference: data integration in dynamic models—a review. Biosystems 96, 86–103.

Henderson, J., Michailidis, G., 2014. Network reconstruction using nonparametric additive ODE models. PLoS ONE 9, e94003.

Huang, W., Nakamori, Y., Wang, S.Y., 2005. Forecasting stock market movement direction with support vector machine. Comput. Oper. Res. 32, 2513–2522.

Liang, H., Wu, H., 2008. Parameter estimation for differential equation models using a framework of measurement error in regression models. J. Am. Stat. Assoc. 103, 1570–1583.

Lu, T., Liang, H., Li, H., Wu, H., 2011. High-dimensional odes coupled with mixed-effects modeling techniques for dynamic gene regulatory network identification. J. Am. Stat. Assoc. 106, 1242–1258.

Ma, P., Zhang, N., Huang, J.Z., Zhong, W., 2017. Adaptive basis selection for exponential family smoothing splines with application in joint modeling of multiple sequencing samples. Stat. Sin. 27, 1757–1777.

McCullagh, P., Nelder, J., 1989. Generalized Linear Models, second edition. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Taylor & Francis.

Miao, H., Wu, H., Xue, H., 2014. Generalized ordinary differential equation models. J. Am. Stat. Assoc. 109, 1672–1682.

Nagalakshmi, U., Wang, Z., Waern, K., Shou, C., Raha, D., Gerstein, M., Snyder, M., 2008. The transcriptional landscape of the yeast genome defined by RNA sequencing. Science 320, 1344–1349.

Nasmyth, K., 1993. Control of the yeast cell cycle by the Cdc28 protein kinase. Curr. Opin. Cell Biol. 5, 166–179.

Polynikis, A., Hogan, S., di Bernardo, M., 2009. Comparing different ODE modelling approaches for gene regulatory networks. J. Theor. Biol. 261, 511–530.

Powell, M.J., 2006. The NEWUOA software for unconstrained optimization without derivatives. In: Large-Scale Nonlinear Optimization. Springer, pp. 255–297.

Poyton, A., Varziri, M.S., McAuley, K.B., McLellan, P.J., Ramsay, J.O., 2006. Parameter estimation in continuous-time dynamic models using principal differential analysis. Comput. Chem. Eng. 30, 698–708.

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

Ramsay, J.O., 1996. Principal differential analysis: data reduction by differential operators. J. R. Stat. Soc., Ser. B, Stat. Methodol. 58, 495–508.

Ramsay, J.O., Hooker, G., Campbell, D., Cao, J., 2007. Parameter estimation for differential equations: a generalized smoothing approach. J. R. Stat. Soc., Ser. B, Stat. Methodol. 69, 741–796.

Sloan, L., Morgan, J., 2015. Who tweets with their location? Understanding the relationship between demographic characteristics and the use of geoservices and geotagging on Twitter. PLoS ONE 10, e0142209.

Spellman, P.T., Sherlock, G., Zhang, M.Q., Iyer, V.R., Anders, K., Eisen, M.B., Brown, P.O., Botstein, D., Futcher, B., 1998. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol. Biol. Cell 9, 3273–3297.

Stuart, J.M., Segal, E., Koller, D., Kim, S.K., 2003. A gene-coexpression network for global discovery of conserved genetic modules. Science 302, 249–255.

Tibshirani, R., 1996. Regression shrinkage and selection via the lasso. J. R. Stat. Soc., Ser. B, Stat. Methodol. 58, 267–288.

Tseng, P., Yun, S., 2009. A coordinate gradient descent method for nonsmooth separable minimization. Math. Program. 117, 387–423.

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

Voorman, A., Shojaie, A., Witten, D., 2014. Graph estimation with joint additive models. Biometrika 101, 85–101.

Wahba, G., Wang, Y., Gu, C., Klein, R., Klein, B., 1995. Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy: the 1994 Neyman memorial lecture. Ann. Stat. 23, 1865–1895.

Wang, H., Leng, C., 2008. A note on adaptive group lasso. Comput. Stat. Data Anal. 52, 5277–5286.

Wood, S.N., 2017. Generalized Additive Models: an Introduction with R. CRC Press.

Wu, H., Lu, T., Xue, H., Liang, H., 2014. Sparse additive ordinary differential equations for dynamic gene regulatory network modeling. J. Am. Stat. Assoc. 109, 700–716.

Wu, L., Qiu, X., Yuan, Y.x., Wu, H., 2019. Parameter estimation and variable selection for big systems of linear ordinary differential equations: a matrix-based approach. J. Am. Stat. Assoc. 114, 657–667.

Yuan, M., Kendziorski, C., 2006. Hidden Markov models for microarray time course data in multiple biological conditions. J. Am. Stat. Assoc. 101, 1323–1332.

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

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

Zhang, C.H., 2010. Nearly unbiased variable selection under minimax concave penalty. Ann. Stat. 38, 894–942.

Zhang, H., Conn, A.R., Scheinberg, K., 2010. A derivative-free algorithm for least-squares minimization. SIAM J. Optim. 20, 3555–3576.

Zou, H., 2006. The adaptive lasso and its oracle properties. J. Am. Stat. Assoc. 101, 1418–1429.