Statistica Sinica Preprint No: SS-2016-0113R1	
Title	Modeling subject-specific nonautonomous dynamics
Manuscript ID	SS-2016-0113
URL	http://www.stat.sinica.edu.tw/statistica/
DOI	10.5705/ss.202016.0113
Complete List of Authors	Debashis Paul Siyuan Zhou and Jie Peng
Corresponding Author	Debashis Paul
E-mail	debpaul@ucdavis.edu
Notice: Accepted version subject to English editing.	

Statistica Sinica: Newly accepted Paper
(accepted version subject to English editing)

Modeling subject-specific nonautonomous dynamics

Siyuan Zhou[1], Debashis Paul[2] and Jie Peng[3]

1 Shields Ave., Department of Statistics, University of California, Davis, CA 95616

[1]*email:* zhousiyuan12@gmail.com, [2]*email:* debpaul@ucdavis.edu, [3]*email:* jiepeng@ucdavis.edu

Abstract

We consider modeling non-autonomous dynamical systems for a group of subjects. The proposed model involves a common baseline gradient function and a multiplicative time-dependent subject-specific effect which accounts for phase and amplitude variations in the rate of change across subjects. The baseline gradient function is represented in a spline basis and the subject-specific effect is modeled as a polynomial in time with random coefficients. We establish appropriate identifiability conditions and propose an estimator based on the hierarchical likelihood. We prove consistency and asymptotic normality of the proposed estimator under a regime of moderate-to-dense observations per subject. Simulation studies and an application to the Berkeley Growth Data demonstrate the effectiveness of the proposed methodology.

Keywords : Ordinary differential equation (ODE); gradient function; nonlinear mixed effects models; hierarchical likelihood; Levenberg-Marquardt method; phase variation.

1 Introduction

Continuous time smooth dynamical systems arise in modeling biological, physical, or chemical processes such as, growth of organisms, synthesis of chemicals, disease progression or dynamics of ecological systems. Many of these processes are modeled through systems of ordinary differential equations (ODEs). Most of the existing approaches assume known functional forms of the dynamical systems that are determined by a small number of parameters. However, due to insufficient knowledge, sometimes a more flexible approach to modeling the gradient function of the dynamical system is necessary. Moreover, if observations are on a group of subjects, it may be beneficial to combine information across different subjects. This can be achieved by including subject-specific effects into the model which enables estimation of the population level dynamics, while also estimating the dynamics for each individual. Mixed-effects models for ODEs have been used in pharmacokinetics (Li et al., 2002) and in disease dynamics (Huang et al., 2006; Guedj et al., 2007; Huang and Lu, 2008), where the ODE is assumed to have a known parametric form. Recently, Wang et al. (2014) considered a semiparametric mixed-effects ODE model assuming a parametric ODE and the estimation of is performed by imposing a penalty on the trajectories represented by splines.

In this paper, we model the dynamics for a group of subjects simultaneously by ordinary differential equations, with a common "baseline" dynamics depending on the current state and represented in a spline basis, and time-dependent subject-specific random effects that

capture both amplitude and phase variation. The observed data for the i-th subject is the sample trajectory $X_i(\cdot)$ measured at a set of time points in a finite time interval with measurement errors. In many studies, the rate of change $X_i'(\cdot)$ is assumed to be a function of the state $X_i(\cdot)$ alone, i.e., the dynamics follows an autonomous system. However, many dynamics, especially those arising from biological systems, often display certain phase-variation in addition to amplitude-variation across subjects. This is prominent in the dynamics of human growth where some individuals start puberty earlier while for others the growth rate peaks up at a later age. Since the defining feature of an autonomous system is that the rate of change at any given time is a function of the state at that time, but is not affected by the time itself, thus an autonomous system is inadequate in describing phase variations.

In order to model the phase variation in an interpretable way, we propose a system of nonautonomous ODEs where the gradient function is the product of two parts: a common time-independent fixed effects part (referred to as the *baseline gradient function*), and a time-dependent random effects part, capturing the phase and amplitude variations: $X_i'(t) = e^{Z(t,\boldsymbol{\theta}_i)} g(X_i(t))$, where $\boldsymbol{\theta}_i$ is a random vector representing the unobserved subject-specific effects and $Z(\cdot, \boldsymbol{\theta})$ is a function of time which captures both amplitude and phase variations. We represent the common gradient function $g(\cdot)$ in a spline basis and model $Z(t, \boldsymbol{\theta})$ as a polynomial in t. Decoupling of these two components requires an appropriate identifiability constraint that is discussed in Section 2. Moreover, to avoid singularity in the solution of the ODE, we also assume that g is either strictly positive or negative on its domain.

We propose an estimator based on the framework of hierarchical likelihood (Lee et al., 2006). The model is fitted using the Levenberg-Marquardt nonlinear optimization procedure.

The hierarchical likelihood-based estimation is computationally a much cheaper alternative to the commonly used maximum likelihood procedure for nonlinear mixed effects models (cf. Jiang (2007)) due to nonlinearity in the ODEs and lack of closed form solutions. We adopt an asymptotic framework in which the baseline gradient function g is assumed to be exactly represented by a large but fixed number of basis functions, while for model fitting, we regularize g through adding a roughness penalty to the objective function. Under an appropriate identifiability constraint, we prove consistency of the proposed estimator of g when the measurements become dense within a time interval as the number of subjects n increases. We also prove that, when the number of measurements per subject grows faster than $\sqrt{n}$, the proposed estimator has an asymptotic normal distribution. The latter result can be used to determine confidence sets for the baseline gradient function. Finally, we apply the proposed method to the Berkeley growth data and show that valuable insights about human growth dynamics can be obtained through modeling the growth trajectories at a population level. The proposed method also provides an alternative framework for functional data analysis when such data are characterized by monotone sample trajectories.

Among related works, in Paul et al. (2011) we considered a model with $Z(t, \boldsymbol{\theta}_i) = \boldsymbol{\theta}_i$ to capture subject-specific amplitude variations in autonomous ODEs, even though no theoretical analysis was presented. The current proposal is seen as an extension of that model to nonnautonomous ODEs. Also, in Paul et al. (2016) we considered nonparametric estimation of g based on a single trajectory. Both the methodology and theoretical analysis presented in this paper are substantially different from that in Paul et al. (2016).

Rest of the paper is organized as follows. The model is described in Section 2 and the

model fitting procedure in Section 3. The asymptotic theory is established in Section 4. A simulation study is reported in Section 5 and the application to Berkeley growth data is described in Section 6. Proof details are given in the Appendix. Further details and additional numerical and graphical summaries are available in the Supplementary Material.

2 Model

In this section, we describe the proposed model and then discuss the identifiability constraint. We assume that the true trajectory $X_i(\cdot)$, corresponding to the i-th subject, follows the ODE:

$$X_i'(t) = e^{Z(t,\boldsymbol{\theta}_i)} g(X_i(t)), \quad t \in [0,1], \quad X_i(0) = a_i, \qquad i = 1, \ldots, n. \tag{1}$$

For simplicity of exposition, throughout we treat the initial conditions $a_i := X_i(0)$, $i = 1, \ldots, n$, as known, even though they can also be treated as random effects and estimated in a similar fashion as $\boldsymbol{\theta}_i$'s. We further assume that, the baseline gradient function g in (1) is represented by a finite set of spline functions whose combined support covers the range of the observed trajectories:

$$g(x) = g_{\boldsymbol{\beta}}(x) = \sum_{j=1}^{M} \beta_j \phi_j(x) \tag{2}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_M)^T$ is unknown and $\boldsymbol{\Phi}(\cdot) = (\phi_1(\cdot), \ldots, \phi_M(\cdot))^T$ is a set of spline functions. In this paper, we use a cubic B-spline basis with equally spaced knots and a large fixed M. Larger values of M provides a more accurate approximation to g while leading to higher variability of the estimator. To address this, we regularize g by adding a ridge-type roughness penalty term to the objective function (see (8)) to achieve bias-variance trade-off.

We model the subject-specific effect as a polynomial in t with random coefficients:

$$Z(t, \boldsymbol{\theta_i}) = \theta_{i1} + \theta_{i2} t + \cdots + \theta_{ip} t^{p-1}, \quad i = 1, \cdots, n, \tag{3}$$

with the working assumption that $\boldsymbol{\theta}_i = (\theta_{i1}, \ldots, \theta_{ip})^T \overset{i.i.d.}{\sim} N(\boldsymbol{\mu}_\theta, \Sigma_\theta)$, $\boldsymbol{\mu}_\theta = (\mu_1, \ldots, \mu_p)^T$ and $\Sigma_\theta = \text{diag}(\sigma^2_{\theta_1}, \ldots, \sigma^2_{\theta_p})$ with $\sigma^2_{\theta_k} > 0$ for $k = 1, \ldots, p$. Larger values of p increase model variability and consequently require a finer grid for numerically solving of the ODEs to overcome numerical instability. Instead of monomials in t, an orthogonal polynomial basis may be used to improve computational stability. The key feature of the random effect $Z(t, \boldsymbol{\theta}_i)$ needed for theoretical derivations is that $Z(t, \boldsymbol{\theta}_i)$ is linear in the parameter $\boldsymbol{\theta}_i$.

Finally, the observed data are modeled as:

$$Y_{ij} = X_i(T_{ij}) + \varepsilon_{ij}, \quad j = 1, \ldots, m_i, \quad i = 1, \ldots, n, \tag{4}$$

with the working assumption that ε_{ij}'s are i.i.d. $N(0, \sigma^2_\varepsilon)$ and T_{ij}s are i.i.d. following a distribution with density f_T supported on $[0, 1]$. We also assume the observational errors ε_{ij}'s and the random parameters $\boldsymbol{\theta}_i$'s are independent.

The model specified by equations (1), (2) and (3) is not identifiable without additional constraints. The reason is that the following transformation of the parameters

$$\boldsymbol{\beta} \mapsto e^c \boldsymbol{\beta} \ \text{ and } \ (\theta_{i1}, \ldots, \theta_{ip}) \mapsto (-c + \theta_{i1}, \ldots, \theta_{ip}), \qquad c \in \mathbb{R}, \tag{5}$$

leaves the trajectories determined by the model invariant. This also suggests that a natural way to impose identifiability is to ensure that either the scale of $\boldsymbol{\beta}$ or the mean of θ_{i1} is kept fixed at a given value. However, the constraint $\mu_1 := \mathbb{E}(\theta_{i1}) = 0$ is not effective in ensuring the asymptotic identifiability of the system, which can be seen from the asymptotic analysis

in Section 4. Therefore, we impose identifiability through a constraint on $\boldsymbol{\beta}$, namely

$$\sum_{j=1}^{M} \beta_j = 1. \tag{6}$$

In Section 4.2, we prove that (6), together with some technical conditions on the sampling design, ensures asymptotic identifiability of the parameters.

3 Model Fitting

A common approach for fitting mixed effects models is to integrate out the random effects and then maximize the resulting marginal likelihood with respect to the fixed effects. However, this approach is computationally impractical here as it involves integrating out a random parameter (i.e $\{\boldsymbol{\theta}_i\}_{i=1}^n$) in the solution of a nonlinear ODE which does not have an analytical form and can only be numerically evaluated. Instead, we adopt a hierarchical likelihood (henceforth, H-likelihood) approach (Lee et al., 2006), which is a first order approximation to the marginal likelihood. The H-likelihood approach involves specifying a working model for the distribution of random effects and then maximizing the resulting joint likelihood for the fixed and random effects. This can also be viewed as a penalized maximum likelihood procedure. For dense measurements, the H-likelihood based estimate of the fixed effects closely approximates the MLE or its second order approximation through Laplace's method (Jiang, 2007).

3.1 Penalized loss function

Let $X_i(\cdot;\boldsymbol{\theta}_i,\boldsymbol{\beta})$ denote the solution to (1) with the gradient function $g(\cdot) \equiv g_{\boldsymbol{\beta}}(\cdot)$ specified in (2) and $Z(\cdot,\boldsymbol{\theta}_i)$ specified in (3). Then the negative joint log likelihood of the observed data $\mathbf{Y} := (Y_{ij} : 1 \le i \le m_i; 1 \le i \le n)$ and the random effects $\boldsymbol{\Theta} := (\boldsymbol{\theta}_1, \cdots, \boldsymbol{\theta}_n)$ is given by (up to multiplicative and additive constants):

$$\sum_{i=1}^{n}\sum_{j=1}^{m_i} \ell_{ij}(\boldsymbol{\theta}_i,\boldsymbol{\beta}),$$

$$\text{with,} \qquad \ell_{ij}(\boldsymbol{\theta}_i,\beta) = (Y_{ij} - X_i(t_{ij};\boldsymbol{\theta}_i,\boldsymbol{\beta}))^2 + \sigma_\varepsilon^2(\boldsymbol{\theta}_i - \boldsymbol{\mu}_\theta)^T\Sigma_\theta^{-1}(\boldsymbol{\theta}_i - \boldsymbol{\mu}_\theta)/m_i. \tag{7}$$

The trajectory $X_i(\cdot;\boldsymbol{\theta}_i,\boldsymbol{\beta})$ and its gradients with respect to $\boldsymbol{\theta}_i$ and $\boldsymbol{\beta}$ can be numerically computed using the Runge-Kutta method, as described in Paul et al. (2011).

In order to achieve a bias-variance trade-off and higher computational stability, we use a fixed large M and impose a roughness penalty on g of the form $\lambda_\beta \int_{d_0}^{d_1}(g''(x))^2dx$. Here $[d_0, d_1]$ is the range covered by the trajectories (i.e., the domain of g) and $\lambda_\beta \ge 0$ is a penalty parameter. This formulation is related to penalized spline regression (Ruppert, 2002; Yu and Ruppert, 2002). Under (2), we have

$$\int_{d_0}^{d_1}(g''(x))^2dx = \boldsymbol{\beta}^T\mathbf{H}\boldsymbol{\beta}, \quad \text{with} \quad \mathbf{H} = \int_{d_0}^{d_1}\boldsymbol{\Phi}''(x)(\boldsymbol{\Phi}''(x))^Tdx$$

where $\boldsymbol{\Phi}(x) = (\phi_1(x), \ldots, \phi_M(x))^T$. Then the penalized loss function is:

$$L(\boldsymbol{\theta},\boldsymbol{\beta}) = \sum_{i=1}^{n}\sum_{j=1}^{m_i}\ell_{ij}(\boldsymbol{\theta}_i,\boldsymbol{\beta}) + \lambda_\beta\boldsymbol{\beta}^T\mathbf{H}\boldsymbol{\beta}. \tag{8}$$

The proposed estimator $(\hat{\boldsymbol{\theta}},\hat{\boldsymbol{\beta}})$ is the minimizer of loss function (8):

$$(\widehat{\boldsymbol{\theta}},\widehat{\boldsymbol{\beta}}) := \underset{(\boldsymbol{\theta},\boldsymbol{\beta})}{\operatorname{argmin}}\ L(\boldsymbol{\theta},\boldsymbol{\beta}), \quad \text{subject to} \quad \sum_k \beta_k = 1. \tag{9}$$

Estimator of the gradient function g is then: $\widehat{g}(x) = \sum_{j=1}^{M} \hat{\beta}_j \phi_j(x)$, and estimated trajectories X_i's can be evaluated by solving (1) with $\boldsymbol{\theta}_i$'s replaced by $\widehat{\boldsymbol{\theta}}_i$'s and g replaced by $\widehat{g}$.

3.2 Fitting algorithm

We use the Levenberg-Marquardt algorithm for nonlinear regression (Nocedal and Wright, 2006) to minimize (8). It involves iteratively updating $\boldsymbol{\theta}_i$'s and $\boldsymbol{\beta}$. At each step, we need to evaluate $X_i(\cdot; \boldsymbol{\theta}_i, \boldsymbol{\beta})$ and its partial derivatives with respect to $\boldsymbol{\theta}_i$ and $\boldsymbol{\beta}$. Since these are not available in close forms, a 4^{th} order Runge-Kutta method is used to evaluate these functions on a fine grid. More details are given in a Supplementary Material.

Let $\boldsymbol{\theta}_i^c$'s and $\boldsymbol{\beta}^c$ denote the current estimates of $\boldsymbol{\theta}_i$ and $\boldsymbol{\beta}$, at each updating step of the Levenberg-Marquardt algorithm. Then μ_k is estimated as the mean of the θ_{ik}^c's. The variances σ_k^2s and σ_ε^2 can be viewed as either unknown parameters or as tuning parameters. In the former case, they are estimated as the empirical variances of θ_{ik}^c's and current residuals $\tilde{\varepsilon}_{ij} = Y_{ij} - X_i(t_{ij}; \boldsymbol{\theta}_i^c, \boldsymbol{\beta}^c)$, respectively. In the latter case, these parameters can be selected similarly as the penalty parameter λ_β through cross validation. The penalty parameter λ_β controls the trade-off between fidelity to the data and the complexity of the model. We use an approximate leave-one-subject-out cross-validation score $\widetilde{CV}$ for choosing λ_β, described in Section S.6 of the Supplementary Material.

4 Asymptotic theory

In this section, we present results on the asymptotic behavior of the proposed estimator of the baseline gradient function g under the model specified by (1) – (4). For simplicity of

exposition, let $\phi_1, \ldots, \phi_M$ be the B-spline basis functions of degree ≥ 3, with equally spaced knots, and combined support $[d_0, d_1]$. The asymptotic theory remains valid for any well-conditioned basis with twice continuously differentiable basis functions supported on this interval. In order to avoid singularity in the solution of the ODE, throughout we assume that g is either strictly positive or strictly negative on its domain.

Throughout we assume that the initial conditions are randomly distributed and known. In order to simplify the derivations, we treat $\boldsymbol{\mu}_{\boldsymbol{\theta}} = \mathbb{E}(\boldsymbol{\theta}_i)$ as known and, without loss of generality, equal to zero. If the mean $\boldsymbol{\mu}_{\boldsymbol{\theta}}$ is unknown, it would result in additional terms in the expression for the Fisher information with respect to $\boldsymbol{\beta}$, which can be neglected asymptotically under assumptions **A1**–**A4** and **F1** (see Sections 4.1 and 4.2), as indicated in Section S.3.1 of the Supplementary Material.

We establish consistency and asymptotic normality of the proposed estimator $\widehat{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$ under the identifiability constraint (6). For consistency, we assume that m_i's, the number of measurements per subject, increase to infinity uniformly as the sample size n increases. For asymptotic normality, we further assume that $\min_{1 \leq i \leq n} m_i$ grows faster than $\sqrt{n}$. The asymptotic theory presented here differs from the standard theory of nonlinear mixed effects models (cf. Jiang (2007)) due to the use of H-likelihood estimator rather than marginal likelihood estimator, and the imposition of the identifiability constraint on $\boldsymbol{\beta}$. Proofs of the asymptotic results make use of the profile H-likelihood with respect to $\boldsymbol{\beta}$, where the profiling is done by substituting θ_i's by their local optimizers as a function of $\boldsymbol{\beta}$.

4.1 Assumptions

A0 The true parameter $\boldsymbol{\beta}_0$ satisfies $\sum_{k=1}^{M} \beta_{k,0} = 1$, and $g_{\boldsymbol{\beta}_0}(x) > 0$ for all $x \in (d_0, d_1)$.

A1 The distribution F_θ of $\boldsymbol{\theta}_i$'s and the distribution F_a of a_i's have bounded support.

A2 The measurement times T_{ij} are randomly distributed on $[0, 1]$ with a density f_T that is bounded above and below (away from zero). Also, the noise ε_{ij} are i.i.d. $N(0, \sigma_\varepsilon^2)$.

A3 Let $\underline{m} = \min_{1 \le i \le n} m_i$ and $\overline{m} = \max_{1 \le i \le n} m_i$. Then $\underline{m} \to \infty$ as $n \to \infty$ so that $\overline{m}/\underline{m}$ remains bounded.

Condition **A0** reduces one degree of freedom in the parameter and can always be achieved through a recentering of θ_{i1}. Condition **A1** helps to ensure the boundedness of the trajectories and their derivatives. Moreover, **A0** and a refinement of conditions **A0** and **A1** (condition **F1** in Section A.1) are needed to prove that Fisher information matrix (19) associated with the profile H-likelihood with respect to $\boldsymbol{\beta}$ is nonsingular (see Section 4.2). The latter ensures asymptotic identifiability of the model. Conditions **A2** and **A3** can be replaced by assuming that the observations are on a regular grid with grid spacings converging to zero as $n \to \infty$. The assumption on ε_{ij}'s can be relaxed to that ε_{ij}'s are i.i.d. sub-Gaussian random variables.

4.2 Asymptotic identifiability

We present a detailed analysis of asymptotic identifiability. The identifiability condition (6) allows us to reparametrize $\boldsymbol{\beta}$ as follows:

$$\boldsymbol{\beta} = \boldsymbol{\beta}(\boldsymbol{\gamma}) := M^{-1}\mathbf{1}_M + \mathbf{C}\boldsymbol{\gamma}, \quad \boldsymbol{\gamma} \in \mathbb{R}^{M-1}, \tag{10}$$

where $\mathbf{C}$ is an $M \times (M-1)$ matrix satisfying $\mathbf{C}^T \mathbf{1}_M = \mathbf{0}$ and $rank(\mathbf{C}) = M-1$.

Due to the reparametrization of $\boldsymbol{\beta}$, given by (10), we express the likelihood function and its derivatives as a function of $\boldsymbol{\gamma}$. Let $\boldsymbol{\gamma}_0$ and $\boldsymbol{\theta}_i^*$ denote the true parameters, and a_i's denote the (true) initial conditions. In all of the following, we suppress the dependence of the trajectories on the initial conditions a_i's since these are treated as known.

Define the *negative penalized log H-likelihoood* for the i-th subject by

$$L_i^H(\boldsymbol{\theta}_i, \boldsymbol{\gamma}) \;=\; \frac{1}{2}\left(\sum_{j=1}^{m_i}(Y_{ij} - X(T_{ij}; \boldsymbol{\theta}_i, \boldsymbol{\beta}(\boldsymbol{\gamma})))^2 + \boldsymbol{\theta}_i^T \Psi^{-1} \boldsymbol{\theta}_i + \frac{\lambda}{n}\boldsymbol{\beta}(\boldsymbol{\gamma})^T \mathbf{H} \boldsymbol{\beta}(\boldsymbol{\gamma})\right) \quad (11)$$

where $\Psi = (1/\sigma_\varepsilon^2)\Sigma_\theta$. Note that, $L_i^H(\boldsymbol{\theta}_i, \boldsymbol{\gamma}) = \frac{1}{2}\left(\sum_{j=1}^{m_i} \ell_{ij}(\boldsymbol{\theta}_i, \boldsymbol{\beta}(\boldsymbol{\gamma})) + \frac{\lambda}{n}\boldsymbol{\beta}(\boldsymbol{\gamma})^T \mathbf{H} \boldsymbol{\beta}(\boldsymbol{\gamma})\right)$, where ℓ_{ij} is as in (7) with $\boldsymbol{\mu}_\theta = 0$.

Due to the lack of convexity of $L_i^H(\boldsymbol{\theta}_i, \boldsymbol{\gamma})$ with respect to $\boldsymbol{\theta}_i$, in general the value of $\boldsymbol{\theta}_i$ minimizing this function for a given $\boldsymbol{\gamma}$ is not unique. Therefore, throughout, we use the following definition of $\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\gamma}) \equiv \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta})$:

$$\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\gamma}) = \arg \min_{\boldsymbol{\theta}_i \in \mathcal{B}(\boldsymbol{\theta}_i^*, \rho_n)} L_i^H(\boldsymbol{\theta}_i, \boldsymbol{\gamma}), \quad (12)$$

where $\rho_n = O((\log n)^{-2})$ and $\mathcal{B}(\boldsymbol{\theta}_i^*, \rho_n) = \{\boldsymbol{\theta}_i \in \mathbb{R}^p : \| \boldsymbol{\theta}_i - \boldsymbol{\theta}_i^* \| \leq \rho_n\}$. Thus, $\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\gamma})$ is a local minimizer of $L_i^H(\boldsymbol{\theta}_i, \boldsymbol{\gamma})$ given $\boldsymbol{\gamma}$ that is also a global minimizer within a radius ρ_n of the true $\boldsymbol{\theta}_i^*$. See Remark 1 for obtaining an initial estimate of $\boldsymbol{\theta}_i$ that satisfies this. The "estimator" $\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\gamma})$ can be shown to be uniformly close to $\boldsymbol{\theta}_i^*$ when $\boldsymbol{\gamma}$ is within a suitably small distance from $\boldsymbol{\gamma}_0$ (see (A.8) for details). Define the *negative profile log H-likelihoood* for $\boldsymbol{\gamma}$ by

$$L^P(\boldsymbol{\gamma}) = \sum_{i=1}^{n} L_i^P(\boldsymbol{\gamma}), \quad \textit{where} \quad L_i^P(\boldsymbol{\gamma}) := L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\gamma}), \boldsymbol{\gamma}), \quad i = 1, \cdots, n,$$

and $\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\gamma})$ is as in (12). Henceforth, we treat L_i^P interchangeably as a function of $\boldsymbol{\gamma}$ or $\boldsymbol{\beta}$.

We now discuss the asymptotic identifiability of the model. Let $q(t) := (1, t, \ldots, t^{p-1})$ so that $Z(t, \boldsymbol{\theta}) = \boldsymbol{\theta}^T q(t)$. Define

$$\Xi^{11}(a, \boldsymbol{\theta}, \boldsymbol{\beta}) = \int_0^1 (X^{\boldsymbol{\beta}}(t; a, \boldsymbol{\theta}, \boldsymbol{\beta}))(X^{\boldsymbol{\beta}}(t; a, \boldsymbol{\theta}, \boldsymbol{\beta}))^T f_T(t) dt \tag{13}$$

$$\Xi^{12}(a, \boldsymbol{\theta}, \boldsymbol{\beta}) = \int_0^1 (X^{\boldsymbol{\beta}}(t; a, \boldsymbol{\theta}, \boldsymbol{\beta}))(X^{\theta}(t; a, \boldsymbol{\theta}, \boldsymbol{\beta}))^T f_T(t) dt \tag{14}$$

$$\Xi^{22}(a, \boldsymbol{\theta}, \boldsymbol{\beta}) = \int_0^1 (X^{\theta}(t; a, \boldsymbol{\theta}, \boldsymbol{\beta}))(X^{\theta}(t; a, \boldsymbol{\theta}, \boldsymbol{\beta}))^T f_T(t) dt, \tag{15}$$

where $X(t; a, \boldsymbol{\theta}, \boldsymbol{\beta})$ denotes the solution to the equation

$$x'(t) = e^{\boldsymbol{\theta}^T q(t)} g_{\boldsymbol{\beta}}(x(t)), \quad t \in [0, 1], \quad x(0) = a, \tag{16}$$

and $X^{\boldsymbol{\beta}} := \frac{\partial}{\partial \boldsymbol{\beta}} X$ and $X^{\theta} = \frac{\partial}{\partial \boldsymbol{\theta}} X$.

We make a (mild) additional assumption:

A4 For all $\boldsymbol{\beta}$ in a neighborhood $\mathcal{B}$ of $\boldsymbol{\beta}_0$, and all $(a, \boldsymbol{\theta}) \in \text{supp}(F_a \times F_\theta)$, the minimum eigenvalue of $\Xi^{22}(a, \boldsymbol{\theta}, \boldsymbol{\beta})$ is bounded below by some constant $c > 0$.

We define

$$\Xi^{1|2}(a, \boldsymbol{\theta}, \boldsymbol{\beta}) = \Xi^{11}(a, \boldsymbol{\theta}, \boldsymbol{\beta}) - \Xi^{12}(a, \boldsymbol{\theta}, \boldsymbol{\beta}) \Xi^{22}(a, \boldsymbol{\theta}, \boldsymbol{\beta})^{-1} \Xi^{12}(a, \boldsymbol{\theta}, \boldsymbol{\beta})^T; \tag{17}$$

$$\widetilde{G}(a, \boldsymbol{\theta}, \boldsymbol{\gamma}) = \mathbf{C}^T \Xi^{1|2}(a, \boldsymbol{\theta}, \boldsymbol{\beta}(\boldsymbol{\gamma})) \mathbf{C}. \tag{18}$$

We also make a slight refinement of **A0** and **A1**, stated as condition **F1** in Section A.1, which ensures that if $\boldsymbol{\beta}$ is in a neighborhood of $\boldsymbol{\beta}_0$, then the combined support of the basis functions $\{\phi_1, \ldots, \phi_M\}$ is covered by trajectories corresponding to suitably chosen pairs $(a_i, \boldsymbol{\theta}_i)$.

Based on these, we deduce that (details given in Section A.1) the matrix

$$\mathcal{G}(\boldsymbol{\gamma}) = \int \widetilde{G}(a, \boldsymbol{\theta}, \boldsymbol{\gamma}) dF_a(a) dF_\theta(\boldsymbol{\theta}) \tag{19}$$

is well-conditioned. Since $\mathcal{G}(\boldsymbol{\gamma})$ is the integrated Fisher information matrix of the profile H-likelihood $L^P(\boldsymbol{\gamma})$ for $\boldsymbol{\gamma}$, this fact is equivalent to the asymptotic identifiability of $\boldsymbol{\gamma}$. This is because, it can be shown that closeness in terms of the values of the objective function implies closeness in terms of the values of $\boldsymbol{\gamma}$, due to the well-conditioning of $\mathcal{G}(\boldsymbol{\gamma})$.

4.3 Consistency and Asymptotic Normality

We prove consistency of the estimator of $\boldsymbol{\beta}$ (equivalently, g) by showing that under the identifiability constraint (6), and assuming that the roughness penalty parameter $\lambda \equiv \lambda_\beta$ is sufficiently small, there is a sequence of local minimizers of the loss function (8) that converges in probability to the true $\boldsymbol{\beta}_0$ as $n \to \infty$. We also determine its rate of convergence and prove its asymptotic normality after appropriate centering and scaling. We introduce the notation $\widetilde{O}$ to mean that if $X_n = \widetilde{O}(c_n)$ then given any $C > 0$, there exists $C' > 0$ such that $|X_n|/c_n \leq C'$ with probability at least $1 - n^{-C}$ for all n.

We make the following assumption about the density of measurements.

A5 There is a $\delta_0 > 0$ such that $\liminf_{n\to\infty} \underline{m} n^{-\delta_0} > 0$ where $\underline{m} = \min_{1\leq i\leq n} m_i$.

This condition ensures that in a Taylor expansion of the loss function with respect to the parameters, terms beyond the quadratics can be ignored asymptotically.

Remark 1. *A good initial estimate for the parameters is helpful for convergence of the algorithm. We can obtain an initial estimate for each θ_i by a simple two-stage method consisting of first obtaining nonparametric smoothers for $X_i(\cdot)$ and $X_i'(\cdot)$ for each i; and then using the expansion*

$$\log \widehat{X}_i'(T_{ij}) = q(T_{ij})^T \boldsymbol{\theta}_i + \log g(\widehat{X}_i(T_{ij})) + \delta_{ij}, \qquad j = 1, \ldots, m_i, \tag{20}$$

where $q(t) = (1, t, \ldots, t^{p-1})^T$ *and* δ_{ij}*'s are approximation errors, to estimate* $\boldsymbol{\theta}_i$ *through regression, while treating* $\log g$ *as an arbitrary smooth function. It can be shown that, under the identifiability restriction and conditions* ***A1***–***A5*** *and* ***F1****, the estimators* $\widetilde{\boldsymbol{\theta}}_i$ *thus obtained satisfy* $\max_{1\le i\le n} \parallel \widetilde{\boldsymbol{\theta}}_i - \boldsymbol{\theta}_i^* \parallel = o_P(\rho_n)$*, where* ρ_n *is as in (12).*

Theorem 1. *Suppose that* **A0**–**A5** *and* **F1** *are satisfied and* $\lambda = o(\sqrt{n})$*. Then there exists a root* $\widehat{\boldsymbol{\gamma}}$ *of the equation* $\frac{d}{d\boldsymbol{\gamma}} L^P(\boldsymbol{\gamma}) = 0$ *that is a local minimizer of* $L^P(\boldsymbol{\gamma})$ *and satisfies*

$$\parallel \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 \parallel = O_P\left(\max\{\frac{1}{\sqrt{n\underline{m}}}, \frac{1}{\underline{m}}\}\right). \tag{21}$$

Clearly, (21) also implies that

$$\parallel \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0 \parallel = O_P\left(\max\{\frac{1}{\sqrt{n\underline{m}}}, \frac{1}{\underline{m}}\}\right). \tag{22}$$

The basic strategy of the proof is as follows. For a small $\alpha_n > 0$, we compare the value of $L^P(\boldsymbol{\gamma}_0 + \alpha_n \boldsymbol{\delta})$, where $\boldsymbol{\delta}$ is a unit vector in $\mathbb{R}^{M-1}$, with that of $L^P(\boldsymbol{\gamma}_0)$. Our goal is to show that, for an appropriately chosen sequence $\alpha_n \to 0$, we have that

$$\mathbb{P}(\inf_{\boldsymbol{\delta}\in\mathbb{R}^{M-1}} L^P(\boldsymbol{\gamma}_0 + \alpha_n\boldsymbol{\delta}) > L^P(\boldsymbol{\gamma}_0)) \to 1. \tag{23}$$

We choose the sequence α_n to be a constant multiple of $\max\{(\log n)^{1/2}(n\underline{m})^{-1/2}, 1/\underline{m}, \lambda/n\}$. This also establishes the existence of a root $\widehat{\boldsymbol{\gamma}}$ of $\frac{d}{d\boldsymbol{\gamma}} L^P(\boldsymbol{\gamma}) = 0$ within a radius of α_n around $\boldsymbol{\gamma}_0$. The result (21) is obtained by expanding $\frac{d}{d\boldsymbol{\gamma}} L^P(\widehat{\boldsymbol{\gamma}})$ around $\boldsymbol{\gamma}_0$ in a Taylor series and thereby obtaining an asymptotic representation of $\widehat{\boldsymbol{\gamma}}$. The details are given in the Appendix.

We also prove asymptotic normality of the estimator of $\boldsymbol{\gamma}$ by imposing the following stronger condition on the rate of growth of m_i's.

A5′ $\underline{m} = \min_{1\le i\le n} m_i \gg (\log n)^6 n^{1/2}$ as $n \to \infty$.

Theorem 2. *Suppose that* **A0**–**A4**, **A5′** *and* **F1** *are satisfied and* $\lambda = o(\sqrt{n})$. *Let* $N_n = \sum_{i=1}^{n} m_i$. *Then there exists a root* $\widehat{\boldsymbol{\gamma}}$ *of the equation* $\frac{d}{d\boldsymbol{\gamma}} L^P(\boldsymbol{\gamma}) = 0$ *that satisfies*

$$\sqrt{N_n}(\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 - \Gamma_n(\boldsymbol{\gamma}_0)^{-1}\mathfrak{b}_n(\boldsymbol{\gamma}_0)) \Longrightarrow N(0, \sigma_\varepsilon^2 \Gamma(\boldsymbol{\gamma}_0)^{-1}) \tag{24}$$

where $\mathfrak{b}_n(\boldsymbol{\gamma}_0)$ *is a stochastic bias term of the form*

$$\mathfrak{b}_n(\boldsymbol{\gamma}_0) = \frac{1}{N_n} \sum_{i=1}^{n} \sum_{\ell=1}^{4} \mathbf{C}^T f_\ell(a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\gamma}_0) \tag{25}$$

where $\{f_\ell(a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\gamma}_0)\}_{\ell=1}^4$,*are given by (S.1)–(S.4) in the Supplementary Material, and*

$$\Gamma_n(\boldsymbol{\gamma}_0) = \frac{1}{N_n} \sum_{i=1}^{n} m_i \widetilde{G}(a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\gamma}_0), \tag{26}$$

and $\Gamma(\boldsymbol{\gamma}_0) := \lim_{n\to\infty} \Gamma_n(\boldsymbol{\gamma}_0)$, *assuming the limit exists in probability. If this limit exists in probability only along a subsequence, then the limit in (24) holds along the same subsequence.*

The proof of Theorem 2 is in the Appendix.

Remark 2. *If* m_i*'s are i.i.d. following a distribution (indexed by* n*) supported on* $[\underline{m}, \overline{m}]$, *independent of* $(\boldsymbol{\theta}_i, \{\varepsilon_{ij}\})$, *and* **A5′** *holds, then* $\Gamma(\boldsymbol{\gamma}_0)$ *in (26) exists almost surely and equals to the integrated Fisher information* $\mathcal{G}(\boldsymbol{\gamma}_0)$ *(see equation (19)).*

Remark 3. *The term* $\Gamma_n(\boldsymbol{\gamma}_0)^{-1}\mathfrak{b}_n(\boldsymbol{\gamma}_0)$ *is a bias term that is of the order* $O_P(\underline{m}^{-1})$. *This term results from the nonlinearity of* $X(t; a_i, \boldsymbol{\theta}_i, \boldsymbol{\beta})$ *with respect to* $\boldsymbol{\theta}_i$. *If* $\underline{m} \gg n$, *i.e., the measurements per subject are sufficiently dense, then this term can be neglected in (24).*

Remark 4. *By Theorem 2, the asymptotic variance of* $\widehat{\boldsymbol{\gamma}}$ *is* $\sigma_\varepsilon^2 \Gamma_n(\boldsymbol{\gamma}_0)^{-1}$. *We estimate* σ_ε^2 *by*

$$\widehat{\sigma}_\varepsilon^2 = \frac{1}{df_n} \sum_{i=1}^{n} \sum_{j=1}^{m_i} (Y_{ij} - \widehat{X}_{ij}(\widehat{\boldsymbol{\beta}}))^2 \tag{27}$$

where $df_n := \sum_{i=1}^{n}(m_i - p) - M = N_n - np - M$ *is the degrees of freedom (since* $\boldsymbol{\theta}_i$ *is* p*-dimensional and* $\boldsymbol{\beta}$ *is* M*-dimensional). We can also estimate* $\Gamma_n(\boldsymbol{\gamma}_0)$ *and* $\mathfrak{b}_n(\boldsymbol{\gamma}_0)$ *from data, as is shown in Section S.1 of the Supplementary Material.*

5 Simulation Study

In this section, we investigate the finite sample performance of the proposed methodology in estimating the gradient functions and the trajectories. We also examine the implications of mis-specification of the subject-specific effect components $Z(\cdot;\boldsymbol{\theta}_i)$.

We refer to the true baseline gradient function g used here as the "two-peak" function according to its shape. The "two-peak" gradient function g and some random realizations of the trajectories following model (1) are depicted in Figure S.1 in the Supplementary Material. The chosen g is not exactly representable by a finite number of cubic B-spline functions with equally spaced knots, though the accuracy of approximation improves with more knots.

We consider two different sampling rates: (i) `sparse`: 3–8 measurements per trajectory/subject; and (ii) `dense`: 30–50 measurements per trajectory/subject. For each setting, we generate $n = 25$ sample trajectories. We set the subject-specific functions $Z(\cdot, \boldsymbol{\theta}_i)$ as linear functions in t (i.e., $p = 2$). For the `dense` case, the random parameters $\boldsymbol{\theta}_i$'s are i.i.d. Normal with $\boldsymbol{\mu}_\theta = (0, 2)$ and $\Sigma_\theta = \text{diag}(0.1^2, 0.2^2)$. For the `sparse` case, $\boldsymbol{\mu}_\theta = (0, 0)^T$ and $\boldsymbol{\mu}_\theta = (0, 2)^T$ the variance Σ_θ is the same as in the `dense` case. Measurement errors ε_{ij}s are i.i.d. $N(0, \sigma_\varepsilon^2)$. For the `sparse` case, we set $\sigma_\varepsilon = 0.01$. For the `dense` case, we consider $\sigma_\varepsilon = 0.01$ and $\sigma_\varepsilon = 0.02$. For each scenario, 500 independent replicates are generated.

In the fitting procedure, we use M cubic B-spline basis functions with equally spaced

knots on the combined range of the observed trajectories. We choose $M = 30$ to allow a high degree of flexibility in representing g. Twenty values of the roughness penalty parameter λ_β over an appropriate range are considered and the "optimal" value $\lambda_{\beta,opt}$ is selected by the approximate leave-one-subject-out cross validation score $\widetilde{CV}$. We use $\widehat{g}$ and $\{\hat{X}_i(\cdot)\}_{i=1}^n$ to denote the estimates of the baseline gradient function and sample trajectories, respectively, corresponding to $\lambda_{\beta,opt}$. We also use the true model with $p = 2$ for the subject-specific effect $Z(\cdot;\boldsymbol{\theta}_i)$ and allow up to 5000 iterations in the Levenberg-Marquardt algorithm.

For performance evaluation, we report summary statistics of integrated squared error (ISE) of gradient functions estimation and trajectories estimation, denoted by $ISE(\widehat{g})$ and $ISE(\widehat{X})$, respectively, across the 500 independent replicates. In order to simultaneously evaluate the estimated gradient functions across all subjects, we define

$$ISE(\widehat{g}) := \int_0^1 (e^{\hat{\mu}(t)}\widehat{g}(\widehat{\overline{X}}(t)) - e^{\mu(t)}g(\overline{X}(t)))^2 dt,$$

where $\hat{\mu}(t) = \sum_{k=1}^p \hat{\mu}_k t^{k-1}$ is the estimated mean subject-specific effect, $\overline{X}(t) = n^{-1}\sum_{i=1}^n X_i(t)$ is the mean of the true sample trajectories, and $\widehat{\overline{X}}(t) = n^{-1}\sum_{i=1}^n \widehat{X}_i(t)$ is the mean of the estimated sample trajectories. The reason to use the function $e^{\mu(t)}g(\overline{X}(t))$ as the benchmark is that, this way we can capture the variability in estimation of both g and $\boldsymbol{\theta}_i$'s simultaneously. This function also reflects the derivative of $X_i(t)$ averaged across subjects and thus can be seen as an overall measure of the rate of growth. Finally, $ISE(\widehat{X})$ is defined as $n^{-1}\sum_{i=1}^n \int_0^1 \left(\widehat{X}_i(t) - X_i(t)\right)^2 dt$.

Estimation accuracy of the gradient functions and the trajectories, based on 500 independent replicates for each scenario, is summarized in Table 1, where $\widetilde{CV}$ is used for selecting the penalty parameter λ_β. We report the mean, standard deviation (SD), median and me-

dian absolute deviation (MAD) of $ISE(\widehat{g})$ and $ISE(\widehat{X})$. These results show that there is a substantial improvement in performance when the sampling rate is increased from the `sparse` case to the `dense` case. One may compare these results with those in Table S.1 of the Supplementary Material. The latter reports the results under the "ideal" model selection criterion, i.e., based on the minimization of $ISE(\widehat{g})$. The median $ISE(\widehat{g})$ under $\widetilde{CV}$ is within a factor of two to that under "ideal" model selection and the difference is smaller in the `dense` case. In terms of trajectory estimation, there is little difference in terms of median $ISE(\widehat{X})$ between these two model selection criteria.

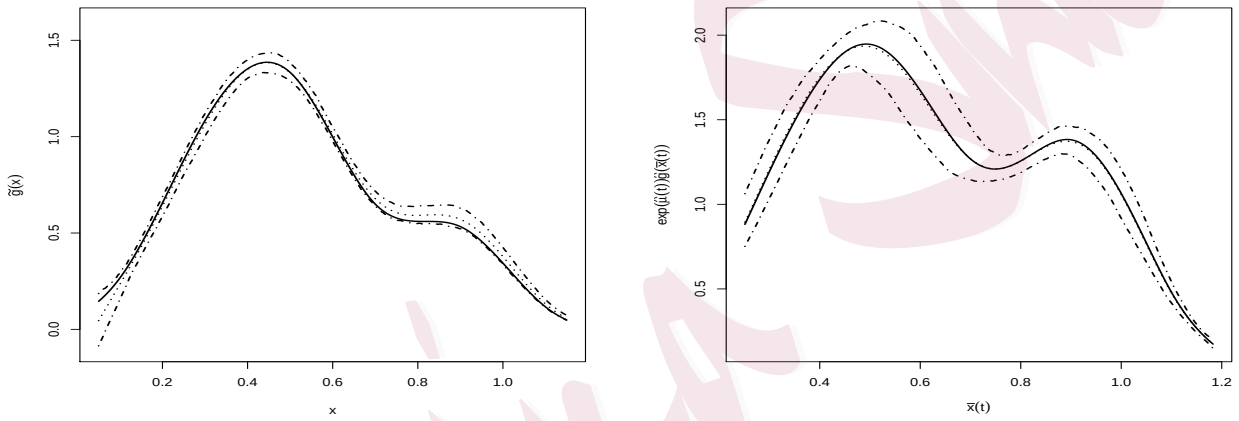
Table 1: Simulation with "Two-peak" baseline gradient function g and linear ($p = 2$) $Z(\cdot, \theta_i)$s. $M = 30$ cubic B-spline basis functions are used in fitting. $\widetilde{CV}$ is used for selection of $\lambda_{\boldsymbol{\beta}}$.

Sampling rate	μ_2	σ_ε	Mean(ISE)	SD(ISE)	Median(ISE)	MAD(ISE)
			$ISE(\widehat{g})$			
sparse	0	0.01	0.001801	0.001238	0.001558	0.001053
	2	0.01	0.004457	0.002762	0.003956	0.002453
dense	2	0.01	0.001500	0.001588	0.001024	0.000810
	2	0.02	0.002730	0.002371	0.001935	0.001319
			$1000 \times ISE(\widehat{X})$			
sparse	0	0.01	0.053307	0.034124	0.049418	0.014179
	2	0.01	0.084852	0.701658	0.050115	0.013050
dense	2	0.01	0.039038	0.700262	0.006779	0.001343
	2	0.02	0.022918	0.005040	0.022217	0.004702

In Figure 1, we depict the estimated gradient functions under the `dense` case with $\sigma_\varepsilon = 0.01$. On the left panel, we plot the point-wise 5% and 95% percentiles and mean of the function $\widetilde{g}(x) := s_\beta \widehat{g}(x)$, along with the true $g(x)$, against x, across 500 independent replicates. Here, $s_\beta := \sum_{j=1}^{M} \beta_j^{tr}$ and $(\beta_j^{tr})_{j=1}^{M}$ is the set of basis coefficients of the projected g onto the $M = 30$ B-spline basis functions with equally spaced knots used in the model fitting procedure. Rescaling by s_β is to account for the scaling in $\widehat{g}$ due to the constraint $\sum_{j=1}^{M} \beta_j = 1$. On the right panel, we plot the point-wise 5% and 95% percentiles and mean of the function $e^{\hat{\mu}(t)} \widehat{g}(\widehat{\overline{X}}(t))$, along with the point-wise mean of $e^{\mu(t)} g(\overline{X}(t))$, all treated as a function of the mean of $\overline{X}(t)$. These plots demonstrate that the gradient function estimation is quite accurate and captures the shape of the true gradient functions. The estimation under all the other cases are depicted in Figures S.2, S.3 and S.4 of the Supplementary Material. The summary for $ISE(\widetilde{g}) := \int (s_\beta \widehat{g}(x) - g(x))^2 dx$, treating $\widetilde{g}(x) = s_\beta \widehat{g}(x)$ as an estimator of $g(x)$, is reported in Table S.2 of the Supplementary Material.

We also investigate the impact of misspecification of the subject-specific effects $Z(\cdot; \boldsymbol{\theta}_i)$s on the estimation accuracy. We use p_{tr} to denote the true order of $Z(\cdot; \boldsymbol{\theta}_i)$s. We consider $p_{tr} = 1$ and $p_{tr} = 2$, and generate data under the `sparse` and `dense` sampling rates for each of these two models. We again use the "Two-peak" function as the true baseline gradient function and $n = 25$ subjects. The hyperparameters are $\boldsymbol{\mu}_\theta = 0$ and $\Sigma_\theta = 0.1^2$ when $p_{tr} = 1$, and $\boldsymbol{\mu}_\theta = (0, 0)^T$ (for `sparse`), or $\boldsymbol{\mu}_\theta = (0, 2)^T$ (for `dense`) and $\Sigma_\theta = \text{diag}(0.1^2, 0.2^2)$ when $p_{tr} = 2$. In all settings, the error variance is $\sigma_\varepsilon^2 = 0.01^2$. In the fitting procedure, we use $M = 30$ cubic B-spline basis functions with equally spaced knots and consider $p = 1$ and $p = 2$ for both models (i.e., $p_{tr} = 1$ and $p_{tr} = 2$).

Figure 1: Simulation with "Two-peak" gradient function g and linear $Z(\cdot, \boldsymbol{\theta}_i)$: `dense` case with $\boldsymbol{\mu}_\theta = (0, 2)^T$ and $\sigma_\varepsilon = 0.01$. **Left panel:** X-axis: x; Solid line: true $g(x)$; Dotted line: point-wise mean of $\widetilde{g}(x) = s_\beta \widehat{g}(x)$; Dash-dotted line : point-wise 5% and 95% percentiles of $\widetilde{g}(x)$. **Right panel:** X-axis: mean of $\overline{X}(t)$; Solid line: point-wise mean of $e^{\mu(t)} g(\overline{X}(t))$; Dotted line: point-wise mean of $e^{\hat{\mu}(t)} \widehat{g}(\widehat{\overline{X}}(t))$; Dash-dotted line : point-wise 5% and 95% percentiles of $e^{\hat{\mu}(t)} \widehat{g}(\widehat{\overline{X}}(t))$.



We report $ISE(\widehat{g})$, while using $p = p_{tr}$ and the "ideal" model selection for λ_β, and $ISE(\widehat{g})$, while using $\widetilde{CV}$ for the selection of both λ_β and p, in Tables S.3 and S.4, respectively, in the Supplementary Material. $\widetilde{CV}$ tends to select larger p, as is evidenced by the fact that in at least 95% of the cases the larger model ($p = 2$) is selected regardless of p_{tr}. However, when $p_{tr} = 2$, the model with $p = 1$ produces significantly biased estimation and inflated ISE (results not reported). On the other hand, when $p_{tr} = 1$, using the model with $p = 2$ results in reasonably good fit. Specifically, the median ISE under $\widetilde{CV}$ is within a factor of three of the median $ISE(\widehat{g})$ under the "ideal" case, i.e., where p_{tr} is used in model fitting.

Moreover, the difference in $ISE(\widehat{g})$ between $\widetilde{CV}$ and the "ideal" case becomes smaller when the sampling rate is increased. These results show that when the models are nested, although $\widetilde{CV}$ tends to select larger models, the proposed method is still reliable in terms of estimating the gradient function, especially for relatively dense observations. Additional model selection results are reported in Section S.8 of the Supplementary Material.

6 Application to Berkeley growth data

We apply the proposed method to the *Berkeley Growth Data* (Tuddenham and Snyder, 1954). This data set consists of measurements of heights (in centimeters) of 54 female and 39 male subjects, measured at 31 time points (same for all subjects) from the age of 1 to 18 years. Our aim to estimate the population level common growth dynamics as well as individual dynamics using the proposed methodology.

Many statistical analyses have been done to describe the features of human growth. A popular approach is curve alignment based on estimation of time-warping functions, including landmark registration (Gasser et al., 1991), continuous monotone registration (Ramsay and Li, 1998), "self-modeling" registration (Gervini and Gasser, 2004). Many parametric models have been proposed for describing postnatal growth in humans, for example, Jenss and Bayley (1937) and Count (1943) for early childhood growth, and Preece and Baines (1978), Bock et al. (1973) and Hauspie et al. (1980) for adolescent growth by logistic and the Gompertz functions. Several models have been proposed to fit individual trajectories at different age intervals (see Hauspie et al. (2004) for an overview). Growth velocities at the level of individual subjects, for various age groups, have also been analyzed nonparametrically by

Gasser et al. (1984) and Gasser et al. (1985).

One important difference of these works from the proposed method is that they primarily focus on fitting individual growth trajectories rather than estimating the common growth dynamics. Our method has the advantage of providing a description of the dynamics at a population level while isolating subject-specific phase and amplitude variations. Thus our approach contributes to enhancing the understanding of common patterns and variations of human growth in a population.

We first carry out a preliminary study to understand the nature of the growth dynamics. We plot the empirical derivatives, denoted by $Y'(t)$, computed by taking successive divided differences, against the observed heights $Y(t)$, across all individuals (Figure S.5 in the Supplementary Material). The empirical gradient shows that the growth rate decreases rapidly at an early age. Around a mean height of 145cm, the female growth rate peaks again, while the male growth rate reaches a peak at a mean height of around 160cm, before slowly decreasing to nearly zero at about 160cm for female and 180cm for male subjects.

We make several small modifications in the fitting procedure to improve its stability and accuracy. First, to improve the fit at an early age and thereby reduce the boundary effect, we linearly extrapolate each trajectory for age below one year. Also, since the rate of growth nearly vanishes beyond a certain height, we force the baseline gradient g to be close to zero for large x by adding a tail penalty of the form

$$\lambda_R \int_A^{200} g(x)^2 dx = \lambda_R \boldsymbol{\beta}^T \left[\int_A^{200} \phi(x)\phi(x)^T dx \right] \boldsymbol{\beta},$$

where $\lambda_R > 0$ and $0 < A < 200$ are selected by $\widetilde{CV}$. Moreover, to improve convergence of the algorithm, we fix $\mu_k = \mathbb{E}(\theta_{ik})$ at zero for $k > 1$.

We use $M = 15$ cubic B-splines with equally spaced knots. We consider three models with constant ($p = 1$) (autonomous model), linear ($p = 2$) and quadratic ($p = 3$) random effects $Z(\cdot, \boldsymbol{\theta_i})$. Since the estimates are not very sensitive to the specification of the variance of the random effects Σ_θ, after preliminary studies, we set Σ_θ to be diagonal with diagonal elements $(\sigma^2_{\theta_k})_{k=1}^p$, with each $\sigma_{\theta_k} = 5$. We also set the error variance σ^2_ε at 0.25. Initial conditions a_i's are treated as known and equal to the value of each (extrapolated) trajectory at time zero. We use $\widetilde{CV}$ to choose various model parameters including $\lambda_\beta, p, A, \lambda_R$. The selected models have $p = 3$ for both genders and have $(A, \lambda_R) = (175, 1000)$ for the female group and $(A, \lambda_R) = (190, 500)$ for the male group.

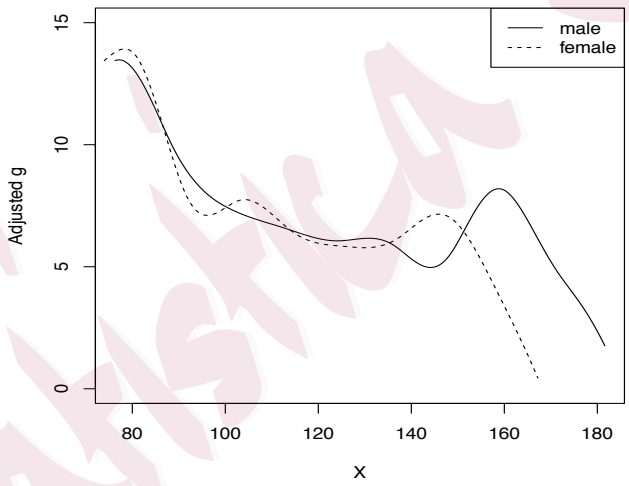


Figure 2: Berkeley growth data. Fitted gradient $e^{Z(t,\widehat{\boldsymbol{\mu}}_\theta)} g(\overline{X}(t))$ against $\overline{X}(t)$ under quadratic subject-specific effects ($p = 3$) for female group (dashed line) and male group (solid line).

Observed and fitted growth trajectories (under the selected model with $p = 3$) are plotted in Figure S.7 of the Supplementary Material which shows good fits for the selected models.

In addition, the MISEs between the fitted and observed trajectories with quadratic subject-specific effects ($p = 3$) improve upon those with the linear subject-specific effect ($p = 2$) by 57.55% (female) and 38.69% (male), respectively, and improve upon those with constant subject-specific effect ($p = 1$, autonomous model) by 75.74% (female) and 61.67% (male), respectively. Residual plots of the trajectory fits against time, under $p = 1, 2, 3$, given in Figure S.8 of the Supplementary Material, show improvements in trajectory fitting with model larger order (i.e., larger p). We also considered the model with $p = 4$. Although it tends to have even smaller MISE and slightly less spread out residuals, since it yields very similar estimate of the gradient function as under $p = 3$, we place the detailed results under this setting in Section S.8 of the Supplementary Material.

Plots of fitted gradient function $e^{Z(t,\widehat{\boldsymbol{\mu}}_\theta)}\widehat{g}(\widehat{\overline{X}}(t))$ against the mean observed trajectory $\overline{X}(t)$ for $p = 3$ are shown in Figure 2. Notably, both females and males have more prominent growth spurts than suggested by the empirical growth dynamics in Figure S.5. Moreover, fitted individual growth rates $\widehat{X}_i'(t)$ versus t and versus the fitted trajectory $\widehat{X}_i(t)$, respectively, depicted in Figure S.6 in the Supplementary Material, clearly show both phase and amplitude variations.

7 Discussion

We propose a flexible approach to modeling a collection of trajectories through ordinary differential equations with subject-specific effects. Our model has a time varying multiplicative random effect component capturing phase and amplitude variations in the trajectories, and a fixed baseline gradient function reflecting population level common dynamics. We imple-

ment an estimation procedure through the hierarchical likelihood framework and provide a detailed asymptotic theory. The proposed method can be used to extract both phase and amplitude variations in the dynamics in an interpretable manner, as is shown by the application to the Berkeley growth data. A nontrivial extension of the theory will be to the setting where the baseline gradient function is treated nonparametrically. The method may also be extended to model dynamics of multivariate trajectories and data involving covariates.

Supplementary Materials

A Supplementary Material, referred as such in the manuscript, contains details of proofs, additional figures and tables, and further simulation and real data analysis results.

Acknowledgement: We thank the Editor, the Associate Editor and the Referees for their constructive suggestions. This research is partially supported by NSF grants DMS-10-07583, DMS-11-06690, DMS-14-07530 and NIH grant 1R01EB021707.

Appendix

A.1 Nonsingularity of observed Fisher information

We make the following assumption which is a slight refinement of **A0** and **A1**. In order to state it, we define $\mathrm{supp}(\Phi)$ to be the set $\bigcup_{j=1}^{M} \mathrm{supp}(\phi_j)$.

- **F1** There exists an integer $K \geq 1$ (not depending on n and m_i's), nonoverlapping intervals $A_1, \ldots, A_K \subset \mathrm{supp}(F_a)$, and a set $\Theta_0 \subset \mathrm{supp}(F_\theta)$ such that,

1. $\mathbb{P}(a \in A_k) > 0$ for all k, and $\mathbb{P}(\boldsymbol{\theta} \in \Theta_0) > 0$;

2. if $p > 1$, then for all $\boldsymbol{\theta} \in \Theta_0$, $\theta_j \neq 0$ for some $j \in \{2, \ldots, p\}$, where θ_j denotes the j-th coordinate of $\boldsymbol{\theta}$;

3. if $(a_{i_k}, \boldsymbol{\theta}_{i_k}) \in A_k \times \Theta_0$ for $k = 1, \ldots, K$, and $\boldsymbol{\beta}$ is in a fixed neighborhood of $\boldsymbol{\beta}_0$, then the ranges of $X(t; a_{i_k}, \boldsymbol{\theta}_{i_k}, \boldsymbol{\beta})$ for successive k's intersect, and one of the following holds:

   (a) $\cup_{k=1}^{K} \{X(t; a_{i_k}, \boldsymbol{\theta}_{i_k}, \boldsymbol{\beta}) : t \in [0,1]\} \supseteq \text{supp}(\Phi)$;

   (b) $\cup_{k=1}^{K} \{X(t; a_{i_k}, \boldsymbol{\theta}_{i_k}, \boldsymbol{\beta}) : t \in [0,1]\} \supseteq B_\Phi$ where the interval $B_\Phi \subseteq \text{supp}(\Phi)$ is such that $\int_{B_\Phi} \phi_j(x)dx = 1$ for all $j = 1, \ldots, M$.

Note that condition 3(b) in **F1** is easily satisfied through proper choice of B_Φ and appropriate renormalization of $\{\phi_j\}_{j=1}^{M}$.

We write $\mathbf{i} \in \mathcal{I}_0$ to indicate that the set of indices $\mathbf{i} = (i_1, \ldots, i_K)$ is such that $a_{i_k} \in A_k$ and $\boldsymbol{\theta}_{i_k} \in \Theta_0$ for all k. Define, for a set of indices $\mathbf{i} \subset \{1, \ldots, n\}$,

$$\widetilde{G}_{\mathbf{i}}(\boldsymbol{\gamma}) = \sum_{k=1}^{K} \widetilde{G}(a_{i_k}, \boldsymbol{\theta}_{i_k}, \boldsymbol{\gamma}).$$

We first show that for any $\mathbf{i} \in \mathcal{I}_0$, the matrix $\widetilde{G}_{\mathbf{i}}(\boldsymbol{\gamma})$ is nonsingular, whenever $\boldsymbol{\gamma}$ (correspondingly, $\boldsymbol{\beta}$) lies in an appropriate fixed neighborhood of $\boldsymbol{\gamma}_0$ (correspondingly, $\boldsymbol{\beta}_0$) and satisfies conditions **A0**, **A1**, **A4** and **F1**. Formally, in the following, we establish that the smallest eigenvalue of $\widetilde{G}_{\mathbf{i}}$, for $\mathbf{i} \in \mathcal{I}_0$, is a positive valued random variable.

The matrix $\widetilde{G}_{\mathbf{i}}(\boldsymbol{\gamma})$ is singular if and only if there exists a nonzero vector $\mathbf{h}$, such that

$\mathbf{h}^T \widetilde{G}_{\mathbf{i}}(\boldsymbol{\gamma})\mathbf{h} = 0$. This condition can be expressed as

$$\begin{aligned} 0 &= \sum_{k=1}^{K} \mathbf{h}^T \mathbf{C}^T G(a_{i_k}, \boldsymbol{\theta}_{i_k}, \boldsymbol{\beta}(\boldsymbol{\gamma}))\mathbf{C}\mathbf{h} \\ &= \sum_{k=1}^{K} \Big[\int_0^1 (\mathbf{h}^T \mathbf{u}_k(t))^2 f_T(t)dt \\ &\quad - (\int_0^1 \mathbf{h}^T \mathbf{u}_k(t)\mathbf{v}_k(t) f_T(t)dt)(\int_0^1 \mathbf{v}_k(t)\mathbf{v}_k(t)^T f_T(t)dt)^{-1}(\int_0^1 \mathbf{h}^T \mathbf{u}_k(t))\mathbf{v}_k(t) f_T(t)dt)^T \Big] \end{aligned}$$

where $\mathbf{u}_k(t) = \mathbf{C}^T X^{\boldsymbol{\beta}}(t; a_{i_k}, \boldsymbol{\theta}_{i_k}, \boldsymbol{\beta}(\boldsymbol{\gamma}))$ and $\mathbf{v}_k(t) = X^{\boldsymbol{\theta}}(t; a_{i_k}, \boldsymbol{\theta}_{i_k}, \boldsymbol{\beta}(\boldsymbol{\gamma}))$. Therefore, using a standard argument from multivariate linear regression, by treating $\mathbf{u}_k(t)$ as response, $\mathbf{v}_k(t)$ as predictor and using the least squares principle, we conclude that there exist $p \times 1$ vectors $\{\mathbf{d}_{i_k}\}_{k=1}^K$ such that

$$\sum_{k=1}^{K} \int_0^1 \Big[\mathbf{h}^T \mathbf{C}^T X_i^{\boldsymbol{\beta}}(t; a_{i_k}, \boldsymbol{\theta}_{i_k}, \boldsymbol{\beta}(\boldsymbol{\gamma})) - \mathbf{d}_{i_k}^T X_i^{\boldsymbol{\theta}}(t; a_{i_k}, \boldsymbol{\theta}_{i_k}, \boldsymbol{\beta}(\boldsymbol{\gamma})) \Big]^2 f_T(t)dt = 0 \tag{A.1}$$

which is equivalent to the following condition: for all $k = 1, \ldots, K$,

$$\mathbf{h}^T X_i^{\boldsymbol{\gamma}}(t; a_{i_k}, \boldsymbol{\theta}_{i_k}, \boldsymbol{\beta}(\boldsymbol{\gamma})) = \mathbf{h}^T \mathbf{C}^T X_i^{\boldsymbol{\beta}}(t; a_{i_k}, \boldsymbol{\theta}_{i_k}, \boldsymbol{\beta}(\boldsymbol{\gamma})) = \mathbf{d}_{i_k}^T X_i^{\boldsymbol{\theta}}(t; a_{i_k}, \boldsymbol{\theta}_{i_k}, \boldsymbol{\beta}(\boldsymbol{\gamma})). \tag{A.2}$$

Under the reparametrization (10), using (16), we have

$$\begin{aligned} g_{\boldsymbol{\beta}(\boldsymbol{\gamma})}(x) &= \boldsymbol{\beta}(\boldsymbol{\gamma})^T \Phi(x) = \boldsymbol{\gamma}^T \mathbf{C}^T \Phi(x) + M^{-1}\mathbf{1}_M^T \Phi(x) =: \tilde{g}_{\boldsymbol{\gamma}}(x) \\ X_i'(t) &= e^{\boldsymbol{\theta}_i^T q(t)} \tilde{g}_{\boldsymbol{\gamma}}(X_i(t)) = e^{\boldsymbol{\theta}_i^T q(t)} \left(\boldsymbol{\gamma}^T \mathbf{C}^T \Phi(X_i(t)) + M^{-1}\mathbf{1}_M^T \Phi(X_i(t)) \right) \\ \frac{d}{dt}(X_i^{\boldsymbol{\theta}}(t)) &= e^{\boldsymbol{\theta}_i^T q(t)} \tilde{g}_{\boldsymbol{\gamma}}(X_i(t)) \frac{\partial X_i(t)}{\partial \boldsymbol{\theta}} + e^{\boldsymbol{\theta}_i^T q(t)} \tilde{g}_{\boldsymbol{\gamma}}(X_i(t)) q(t) \qquad \text{(A.3)} \\ \frac{d}{dt}(X_i^{\boldsymbol{\gamma}}(t)) &= e^{\boldsymbol{\theta}_i^T q(t)} \tilde{g}_{\boldsymbol{\gamma}}(X_i(t)) \frac{\partial X_i(t)}{\partial \boldsymbol{\gamma}} + e^{\boldsymbol{\theta}_i^T q(t)} \mathbf{C}^T \Phi(X_i(t)). \qquad \text{(A.4)} \end{aligned}$$

Differentiating (A.2) with respect to t, and combining with (A.3) and (A.4),

$$\mathbf{h}^T \mathbf{C}^T \Phi(X_{i_k}(t)) = g_{\boldsymbol{\beta}(\boldsymbol{\gamma})}(X_{i_k}(t)) \mathbf{d}_{i_k}^T q(t), \qquad k = 1, \ldots, K, \quad t \in [0, 1]. \tag{A.5}$$

Note that $q(t)$ is a vector of monomials of t of dimension p, $\mathbf{d}_{i_k} \in \mathbb{R}^p$, and the functions ϕ_j's are piecewise polynomials. From this, it can be checked that (A.5) holds only if $\mathbf{d}_{i_k}^T q(t) = c_k$ for all k, where c_k's do not depend on t (see Section S.2 of the Supplementary Material for detail). By **F1**, c_k's must all be equal since the sets $\{X_{i_k}(t)t \in [0,1]\}$ overlap for successive k. Without loss of generality, $c_1 = \cdots = c_K = 1$. Again by **F1**, we conclude that either, $(\mathbf{C}(\mathbf{h}-\boldsymbol{\gamma}) - M^{-1}\mathbf{1}_M)^T\Phi(x) = 0$ for all $x \in \text{supp}(\Phi)$ (under **F1**.3(a)), or $(\mathbf{C}(\mathbf{h}-\boldsymbol{\gamma}) - M^{-1}\mathbf{1}_M)^T\Phi(x) = 0$ for all $x \in B_\Phi$ (under **F1**.3(b)). Consequently, under both scenarios,

$$\mathbf{1}_M^T\mathbf{C}(\mathbf{h}-\boldsymbol{\gamma}) = M^{-1}\mathbf{1}_M^T\mathbf{1}_M = 1.$$

The left hand side of the above equation is 0 since $\mathbf{1}_M^T\mathbf{C} = 0$ and the right hand side is 1. This contradiction proves that there does not exist a vector $\mathbf{h}$ such that $\mathbf{h}^T\widetilde{G}_{\mathbf{i}}(\boldsymbol{\gamma})\mathbf{h} = 0$, which implies that $\widetilde{G}_{\mathbf{i}}(\boldsymbol{\gamma})$ is non-singular and hence its smallest eigenvalue is strictly positive.

A.2 Proof of Theorem 1

Some of the proof details are deferred to the Supplementary Material. Throughout, we use $\boldsymbol{\theta}_i^*$ to denote the true value of $\boldsymbol{\theta}_i$ and drop the reference to the initial value a_i, which is assumed to be known. We use $\boldsymbol{\beta}(\boldsymbol{\gamma})$ to mean $\boldsymbol{\beta}$, where $\boldsymbol{\beta}(\boldsymbol{\gamma}) = \mathbf{C}\boldsymbol{\gamma} + M^{-1}\mathbf{1}_M$. Also, we use $\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta})$ and $\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\gamma})$ interchangeably.

The basic building block is an asymptotic expansion of the restricted optimizer $\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\gamma})$ (see (A.12)). Let $\mathbb{E}_T$ denote the expectation with respect to $T \sim f_T$. Define the following:

$$\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\gamma}|\boldsymbol{\theta}_i^*) = \mathbb{E}_T\left(X(T;\boldsymbol{\theta},\boldsymbol{\beta}) - X(T;\boldsymbol{\theta}_i^*,\boldsymbol{\beta}_0)\right)^2,$$

$$\text{and} \qquad \bar{\boldsymbol{\theta}}_i(\boldsymbol{\beta}) \equiv \bar{\boldsymbol{\theta}}_i(\boldsymbol{\gamma}) = \arg\min_{\boldsymbol{\theta}\in\mathcal{B}(\boldsymbol{\theta}_i^*,\rho_n)} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\gamma}|\boldsymbol{\theta}_i^*),$$

where ρ_n and $\mathcal{B}(\boldsymbol{\theta}_i^*, \rho_n)$ are as in the definition of $\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\gamma})$ (equation (12)). Using the fact that $\mathcal{L}(\boldsymbol{\theta}_i^*, \boldsymbol{\gamma}_0|\boldsymbol{\theta}_i^*) = 0$ and $\boldsymbol{\theta}_i^*$ is a global minimum of $\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\gamma}_0|\boldsymbol{\theta}_i^*)$, it follows through a Taylor expansion that uniformly on $\mathcal{B}(\theta_i^*, \rho_n)$,

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\gamma}|\boldsymbol{\theta}_i^*) &= \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_i^*)^T \nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} \mathcal{L}(\boldsymbol{\theta}_i^*, \boldsymbol{\gamma}_0|\boldsymbol{\theta}_i^*)(\boldsymbol{\theta} - \boldsymbol{\theta}_i^*) \\ &\quad + O(\| \boldsymbol{\theta} - \boldsymbol{\theta}_i^* \|^3) + O(\| \boldsymbol{\theta} - \boldsymbol{\theta}_i^* \| \| \boldsymbol{\gamma} - \boldsymbol{\gamma}_0 \|) + O(\| \boldsymbol{\gamma} - \boldsymbol{\gamma}_0 \|^2). \quad \text{(A.6)}\end{aligned}$$

From (A.6), we can easily deduce that $\| \bar{\boldsymbol{\theta}}_i(\boldsymbol{\gamma}) - \boldsymbol{\theta}_i^* \| = O(\| \boldsymbol{\gamma} - \boldsymbol{\gamma}_0 \|)$ and that $\bar{\boldsymbol{\theta}}_i(\boldsymbol{\gamma})$ satisfies the equation $\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}, \boldsymbol{\gamma}|\boldsymbol{\theta}_i^*)|_{\boldsymbol{\theta} = \bar{\boldsymbol{\theta}}_i(\boldsymbol{\gamma})} = 0$. Using standard arguments, we can now show that under the conditions of Theorem 1, as long as $\| \boldsymbol{\gamma} - \boldsymbol{\gamma}_0 \| = O(\alpha_n)$, $\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\gamma})$ satisfies

$$\nabla_{\boldsymbol{\theta}} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\gamma}), \boldsymbol{\gamma}) = 0 \quad \text{(A.7)}$$

for all i, with probability tending to 1. Moreover, by Implicit Function Theorem, $\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\gamma})$ is a smooth function of $\boldsymbol{\gamma}$ in this neighborhood of $\boldsymbol{\gamma}_0$. Furthermore, we can establish that

$$\max_{1 \le i \le n} \sup_{\boldsymbol{\gamma}: \|\boldsymbol{\gamma} - \boldsymbol{\gamma}_0\| \le \alpha_n} \| \widehat{\boldsymbol{\theta}}_i(\boldsymbol{\gamma}) - \boldsymbol{\theta}_i^* \| = \widetilde{O}(\max\{\sqrt{\log n} \underline{m}^{-1/2}, \alpha_n\}). \quad \text{(A.8)}$$

We then proceed to prove (23). By a Taylor series expansion of $L^P(\boldsymbol{\gamma})$,

$$\begin{aligned}L^P(\boldsymbol{\gamma}_0 + \alpha_n \boldsymbol{\delta}) - L^P(\boldsymbol{\gamma}_0) &\approx \alpha_n \boldsymbol{\delta}^T (\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\gamma}})^T \frac{dL^P(\boldsymbol{\beta}_0)}{d\boldsymbol{\beta}} + \frac{\alpha_n^2}{2} \boldsymbol{\delta}^T (\frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\gamma}})^T \frac{d^2 L^P(\boldsymbol{\beta}_0)}{d\boldsymbol{\beta} d\boldsymbol{\beta}^T} \frac{\partial \boldsymbol{\beta}}{\partial \boldsymbol{\gamma}} \boldsymbol{\delta} \\ &= \alpha_n \boldsymbol{\delta}^T C^T \frac{dL^P(\boldsymbol{\beta}_0)}{d\boldsymbol{\beta}} + \frac{\alpha_n^2}{2} \boldsymbol{\delta}^T \mathbf{C}^T \frac{d^2 L^P(\boldsymbol{\beta}_0)}{d\boldsymbol{\beta} d\boldsymbol{\beta}^T} \mathbf{C} \boldsymbol{\delta}, \quad \text{(A.9)}\end{aligned}$$

where the sign $\approx$ means that the difference between the expression on the left hand side and right hand side of (A.9) is of a smaller order than the right hand side of (A.9) and hence can be ignored asymptotically. The approximation error in the above expansion can be controlled uniformly in $\boldsymbol{\delta}$. To justify this, and similar approximations, throughout we use Hoeffding's

and Bernstein's inequalities, without explicitly referring to them for brevity and ease of presentation. The key fact we need is that, the expected value of $\mathbf{C}^T(d^2L^P(\boldsymbol{\beta})/d\boldsymbol{\beta}d\boldsymbol{\beta}^T)\mathbf{C}$ at $\boldsymbol{\beta}_0$ is a positive definite matrix and is well-conditioned. This allows us to prove (23) and obtain a rate of convergence for $\widehat{\boldsymbol{\beta}}$.

Based on the derivations in Section S.3.1 of the Supplementary Material, the following expressions are valid for $i = 1, \ldots, n$:

$$\frac{d}{d\boldsymbol{\beta}} L_i^P(\boldsymbol{\beta}) = \nabla_{\boldsymbol{\beta}} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}), \boldsymbol{\beta}); \tag{A.10}$$

$$\begin{aligned}\frac{d^2}{d\boldsymbol{\beta}d\boldsymbol{\beta}^T} L_i^P(\boldsymbol{\beta}) = \quad & \nabla_{\boldsymbol{\beta}\boldsymbol{\beta}^T} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}), \boldsymbol{\beta}) \\ & -\nabla_{\boldsymbol{\beta}\boldsymbol{\theta}^T} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}), \boldsymbol{\beta}) \left[\nabla_{\boldsymbol{\theta}\boldsymbol{\theta}^T} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}), \boldsymbol{\beta})\right]^{-1} \nabla_{\boldsymbol{\theta}\boldsymbol{\beta}^T} L_i^H(\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}), \boldsymbol{\beta}).\end{aligned} \tag{A.11}$$

Based on the derivation in Section S.4.2 of the Supplementary Material, we have

$$\widehat{\boldsymbol{\theta}}_i(\boldsymbol{\beta}_0) - \boldsymbol{\theta}_i^* = W_i^*(\boldsymbol{\beta}_0)^{-1} p_{i,\boldsymbol{\theta}} - W_i^*(\boldsymbol{\beta}_0)^{-1}\Psi^{-1}\boldsymbol{\theta}_i^* + W_i^*(\boldsymbol{\beta}_0)^{-1} P_{i,\boldsymbol{\theta}\boldsymbol{\theta}^T} W_i^*(\boldsymbol{\beta}_0)^{-1} p_{i,\boldsymbol{\theta}} + r_{2,i}, \tag{A.12}$$

where $r_{2,i}$ is negligible and the terms $W_i^*(\boldsymbol{\beta}_0)$, $p_{i,\boldsymbol{\theta}}$ and $P_{i,\boldsymbol{\theta}\boldsymbol{\theta}^T}$ are defined in Section S.4.5.

Next, from the derivation in Section S.4.3, we have the expansion

$$\frac{d}{d\boldsymbol{\beta}} L_i^P(\boldsymbol{\beta}_0) = V_i^{(1)}(\boldsymbol{\beta}_0) + V_i^{(2)}(\boldsymbol{\beta}_0) + r_{4,i}, \tag{A.13}$$

where $V_i^{(1)}$ and $V_i^{(2)}$, defined in (S.44) and (S.45), contribute primarily to the asymptotic variance and asymptotic bias, respectively, and $r_{4,i}$ is a negligible remainder term. Further calculations, detailed in Section S.4.5, allow us to conclude that

$$\frac{1}{m_i}\frac{d^2}{d\boldsymbol{\beta}d\boldsymbol{\beta}^T} L_i^P(\boldsymbol{\beta}_0) \quad = \quad \Xi^{1|2}(a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\beta}_0) + \widetilde{O}\left(\max\left\{\frac{\lambda}{n}, \sqrt{\frac{\log n}{\underline{m}}}\right\}\right), \tag{A.14}$$

By the definition of $\widetilde{G}(a_i, \boldsymbol{\theta}_i, \boldsymbol{\beta})$ in (18), the empirical Fisher information matrix satisfies

$$F_n(\boldsymbol{\gamma}_0) := \frac{1}{N_n}\sum_{i=1}^{n} \mathbf{C}^T \frac{d^2}{d\boldsymbol{\beta} d\boldsymbol{\beta}^T} L_i^P(\boldsymbol{\beta}_0)\mathbf{C} = \frac{1}{N_n}\sum_{i=1}^{n} m_i \widetilde{G}(a_i, \boldsymbol{\theta}_i^*, \boldsymbol{\gamma}_0) + \widetilde{O}\left(\max\{\frac{\lambda}{n}, \sqrt{\frac{\log n}{\underline{m}}}\}\right). \tag{A.15}$$

By (A.15), and the fact that $\mathcal{G}(\boldsymbol{\gamma})$, defined in (19), is positive definite at $\boldsymbol{\gamma}_0$, we conclude that for a given $c > 0$, with probability at least $1 - n^{-c}$, the inverse of $F_n(\boldsymbol{\gamma}_0)$ exists and the maximum eigenvalue of the inverse is bounded.

Combining (A.15) with (A.13) and (A.9), and using the fact that $\lambda = o(\sqrt{n})$, we obtain (23) with $\alpha_n = c\max\{(\log n)^{1/2}(n\underline{m})^{-1/2}, \underline{m}^{-1}\}$ for some suitable positive constant c. This establishes that there exists a root $\widehat{\boldsymbol{\gamma}}$ of $\frac{d}{d\boldsymbol{\gamma}} L^P(\boldsymbol{\gamma}) = 0$ satisfying $\| \widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 \| = O_P(\alpha_n)$. Finally, we expand the equation $\frac{d}{d\boldsymbol{\gamma}} L^P(\widehat{\boldsymbol{\gamma}}) = 0$ around $\boldsymbol{\gamma}_0$ to obtain

$$\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0 = -\left(\frac{1}{N_n}\sum_{i=1}^{n} \mathbf{C}^T \frac{d^2}{d\boldsymbol{\beta} d\boldsymbol{\beta}^T} L_i^P(\boldsymbol{\gamma}_0)\mathbf{C}\right)^{-1} \left(\frac{1}{N_n}\sum_{i=1}^{n} \mathbf{C}^T \frac{d}{d\boldsymbol{\beta}} L_i^P(\boldsymbol{\gamma}_0)\right) + O_P(\alpha_n^2 \log n), \tag{A.16}$$

which concludes the proof of Theorem 1 once we isolate the leading terms in $\frac{d}{d\boldsymbol{\beta}} L_i^P(\boldsymbol{\gamma}_0)$ (namely, (S.44) and (S.45)) and use the fact that $(a_i, \boldsymbol{\theta}_i^*)$ are i.i.d.

A.3 Proof of Theorem 2

The proof follows from a careful treatment of (A.16). We move the contribution of the bias term $V_i^{(2)}(\boldsymbol{\beta}_0)$ in (S.45) to the side of $\widehat{\boldsymbol{\gamma}} - \boldsymbol{\gamma}_0$ in (A.16), and use the representations (S.37), (S.38), (S.39) and (S.42), and take appropriate conditional expectations to derive the form of the asymptotic bias $\mathfrak{b}_n(\boldsymbol{\gamma}_0)$. Next we use the representation (A.13) of $\frac{d}{d\boldsymbol{\beta}} L_i^P(\boldsymbol{\beta}_0)$, and the expression of the variance term $V_i^{(1)}(\boldsymbol{\beta}_0)$ in (S.44), and finally a version of martingale central limit theorem, through the independence of $(a_i, \boldsymbol{\theta}_i^*, (\varepsilon_{ij})_{j=1}^{m_i})$ across i, to conclude that

$V_i^{(1)}(\boldsymbol{\beta}_0)$ has an asymptotic Gaussian limit. The condition **A5**' ensures that the remainder terms are $o_P(1)$. The nonsingularity of $\mathcal{G}(\boldsymbol{\gamma}_0)$ combined with (A.15) concludes the derivation.

References

Bock, R. D., Wainer, H., Petersen, A., Thissen, D., Murray, J., and Roche, A. (1973). A parameterization for individual human growth curves. *Human Biology*, 45:63–80.

Count, E. W. (1943). Growth patterns of the human physique: an approach to kinetic anthropometry: Part I. *Human Biology*, 15(1):1–32.

Gasser, T., Kneip, A., Binding, A., Prader, A., and Molinari, L. (1991). The dynamics of linear growth in distance, velocity and acceleration. *Annals of Human Biology*, 18(3):187–205.

Gasser, T., Muller, H.-G., Kohler, W., Molinari, L., and Prader, A. (1984). Nonparametric regression analysis of growth curves. *The Annals of Statistics*, pages 210–229.

Gasser, T., Müller, H.-G., Köhler, W., Prader, A., Largo, R., and Molinari, L. (1985). Velocity and acceleration of height growth using kernel estimation. *Annals of Human Biology*, 12:129–138.

Gervini, D. and Gasser, T. (2004). Self-modelling warping functions. *Journal of the Royal Statistical Society: Series B*, 66(4):959–971.

Guedj, J., Thiébaut, R., and Commenges, D. (2007). Maximum likelihood estimation in dynamical models for HIV. *Biometrics*, 63:1198–1206.

Hauspie, R. C., Cameron, N., and Molinari, L. (2004). *Methods in Human Growth Research*, volume 39. Cambridge University Press.

Hauspie, R. C., Wachholder, A., Baron, G., Cantraine, F., Susanne, C., and Graffar, M. (1980). A comparative study of the fit of four different functions to longitudinal data of growth in height of Belgian girls. *Annals of Human Biology*, 7(4):347–358.

Huang, Y., Liu, D., and Wu, H. (2006). Hierachical Bayesian methods for estimation of parameters in a longitudinal HIV dynamic system. *Biometrics*, 62:413–423.

Huang, Y. and Lu, T. (2008). Modeling long-term longitudinal HIV dynamics with application to an AIDS clinical study. *The Annals of Applied Statistics*, 2:1384–1408.

Jenss, R. M. and Bayley, N. (1937). A mathematical method for studying the growth of a child. *Human Biology*, 9(4):556–563.

Jiang, J. (2007). *Linear and Generalized Linear Mixed Models and Their Applications*. Springer.

Lee, Y., Nelder, J. A., and Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*. CRC Press.

Li, L., Brown, M. B., Lee, K.-H., and Gupta, S. (2002). Estimation and inference for a splineenhanced population pharmacokinetic model. *Biometrics*, 58:601–611.

Nocedal, J. and Wright, S. (2006). *Numerical Optimization*. Springer, New York.

Paul, D., Peng, J., and Burman, P. (2011). Semiparametric modeling of autonomous nonlinear dynamical systems with application to plant growth. *The Annals of Applied Statistics*, 5(3):2078–2108.

Paul, D., Peng, J., and Burman, P. (2016). Nonparametric estimation of dynamics of monotone trajectories. *The Annals of Statistics*, 44:2401–2432.

Preece, M. A. and Baines, M. J. (1978). A new family of mathematical models describing the human growth curve. *Annals of Human Biology*, 5(1):1–24.

Ramsay, J. O. and Li, X. (1998). Curve registration. *Journal of the Royal Statistical Society: Series B*, 60(2):351–363.

Ruppert, D. (2002). Selecting the number of knots for penalized splines. *Journal of Computational and Graphical Statistics*, 11:735–757.

Tuddenham, R. D. and Snyder, M. M. (1954). Physical growth of california boys and girls from birth to eighteen years. *Publications in child development. University of California, Berkeley*, 1(2):183.

Wang, L., Cao, J., Ramsay, J. O., Burger, D. M., Laporte, C. J. L., and Rockstroh, J. K. (2014). Estimating mixed-effects differential equation models. *Statistics and Computing*, 24:111–121.

Yu, Y. and Ruppert, D. (2002). Penalized spline estimation for partially linear single-index models. *Journal of the American Statistical Association*, 97:1042–1054.