Semiparametric modeling of autonomous nonlinear dynamical systems with applications (Technical Report)

Debashis Paul*, Jie Peng* & Prabir Burman

Department of Statistics, University of California, Davis

Abstract

In this paper, we propose a semi-parametric model for autonomous nonlinear dynamical systems and devise an estimation procedure for model fitting. This model incorporates subject-specific effects and can be viewed as a nonlinear semi-parametric mixed effects model. We also propose a computationally efficient model selection procedure. We prove consistency of the proposed estimator under suitable regularity conditions. We show by simulation studies that the proposed estimation as well as model selection procedures can efficiently handle sparse and noisy measurements. Finally, we apply the proposed method to a plant growth data used to study growth displacement rates within meristems of maize roots under two different experimental conditions.

Key words: *autonomous dynamical systems; nonlinear optimization; Levenberg-Marquardt method; leave-one-curve-out cross-validation; plant growth*

1 Introduction

Continuous time dynamical systems arise, among other places, in modeling certain biological processes. For example, in plant science, the spatial distribution of growth is an active area of research (Basu *et al.*, 2007; Schurr, Walter and Rascher, 2006; van der Weele *et al.*, 2003; Walter *et al.*, 2002). One particular region of interest is the root apex, which is characterized by cell division, rapid cell expansion and cell differentiation. A single cell can be followed over time, and thus it is relatively easy to measure its cell division rate. However, in a meristem[1], there is a changing population of dividing cells. Thus the cell division rate, which is defined as the local rate of formation of cells, is not directly observable. If one observes root development from an origin attached to the apex, tissue elements appear to flow through, giving an analogy between primary growth in plant root and fluid flow (Silk, 1994). Thus in Sacks, Silk and Burman (1997), the authors propose to estimate the cell division rates by a continuity equation that is based on the principle of conservation of mass. Specifically, if we assume a steady growth, then the cell division rate is estimated as the gradient (with respect to distance) of cell flux – the rate at which cells are

*equal contributors

[1]meristem is the tissue in plants consisting of undifferentiated cells and found in zones of the plant where growth can take place.

moving past a spatial point. Cell flux is the product of cell number density and growth velocity field. The former can be found by counting the number of cells per small unit file. The latter is the rate of displacement of a particle placed along the root and thus it is a function of distance from the root apex. Hereafter we refer to it as the growth displacement rate. Note that, growth displacement rate is not to be confused with "growth rate" which usually refers to the derivative of the growth trajectory with respect to time. For more details, see Sacks *et al.* (1997). The growth displacement rate is also needed for understanding some important physiological processes such as biosynthesis (Silk and Erickson, 1979; Schurr *et al.*, 2006). Moreover, a useful growth descriptor called the "relative elemental growth rate" (REGR) can be calculated as the gradient of the growth displacement rate (with respect to distance), which shows quantitatively the magnitude of growth at each location within the organ.

There are a lot of research aiming to understand the effect of environmental conditions on the growth in plant. For example, root growth is highly sensitive to environmental factors such as temperature, water deficit or nutrients (Schurr *et al.*, 2006; Walter *et al.*, 2002). For example, in Sharp, Silk and Hsiao (1988), the authors study the effect of water potential on the root elongation in maize primary roots. Root elongation has considerable physiological advantages in drying soil, and therefore knowledge of the locations and magnitudes of growth response to water potential facilitates the quantitative understanding of the underlying regulatory process. In Sacks *et al.* (1997), an experiment is conducted to study the effect of water stress on cortical cell division rates through growth displacement rate within the meristem of the primary root of maize seedlings. In this study, for each plant, measurements are taken on the displacement, measured as the distance in millimeters from the root cap junction (root apex), of a number of markers on the root over a period of 12 hours (Fig. 1: right panel). The plants are divided into two groups - a control group under normal water availability; and a treatment group under a water stress. In Fig. 2, the growth (displacement) trajectories of one plant with 28 markers in the control group, and another plant with 26 markers in the treatment group are depicted. The meristem region of the root, where the measurements are taken, is shown in Fig. 1 (left panel). Note that, by definition, the growth displacement rate characterizes the relationship between the growth trajectory and its derivative (with respect to time). Thus it is simply the gradient function in the corresponding dynamical system. (See Section 2 for more details).

Motivated by this study, in this paper, we focus on modeling and fitting the underlying dynamical system based on data measured over time (referred as *sample curves* or *sample paths*) for a group of subjects. In particular, we are interested in the case where there are multiple replicates corresponding to different initial conditions for each subject. Moreover, for a given initial condition, instead of observing the whole sample path, measurements are taken only at a sparse set of time points together with (possible) measurement noise. In the plant data application, each plant is a subject. And the positions of the markers which are located at different distances at time zero from the root cap junction correspond to different initial conditions. There are in total 19 plants and 445 sample curves in this study. The number of replicates (i.e. markers) for each plant varies between 10 and 31. Moreover, smoothness of the growth trajectories indicates low observational noise levels and an absence of extraneous shocks in the system. Hence, in this paper, we model the growth trajectories through deterministic differential equations with plant-specific effects. We refer to the (common) gradient function of these differential equations as the baseline growth displacement rate.

We first give a brief overview of the existing literature on fitting smooth dynamical systems in continuous time. A large number of physical, chemical or biological processes are modeled through systems of parametric differential equations (cf. Ljung and Glad, 1994, Perthame, 2007, Strogatz,

Figure 1: Left Panel: image of root tip with meristem*: 1 - meristem; 4 - root cap; 5 - elongation zone; Right Panel: an illustration of the root tip with the displacements of three markers indicated at times t_0, t_1, t_2, t_3. (*From wikipedia)

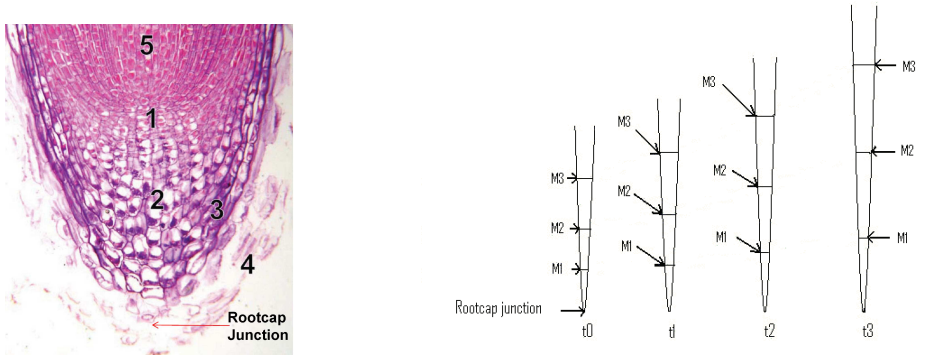
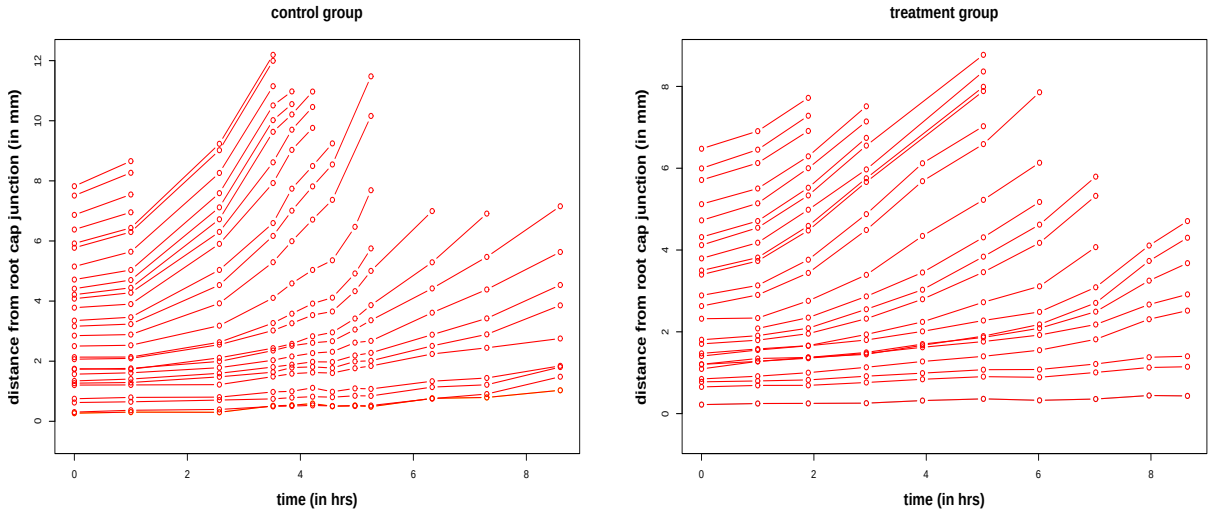


Figure 2: Growth trajectories for plant data. Left panel : a plant in control group; Right panel : a plant in treatment group



2001). Ramsay, Hooker, Campbell and Cao (2007) consider modeling a continuously stirred tank reactor. Zhu and Wu (2007) adopt a state space approach for estimating the dynamics of cell-virus interactions in an AIDS clinical trial. Poyton *et al.* (2006) use the principal differential analysis approach to fit dynamical systems. Recently Chen and Wu (2008a, 2008b) propose to estimate differential equations with known functional forms and nonparametric time-dependent coefficients. Wu and Ding (1999) and Wu, Ding and DeGruttola (1998) propose using nonlinear least squares procedure for fitting differential equations that take into account subject-specific effects. In a recent work, Cao, Fussmann and Ramsay (2008) model a nonlinear dynamical system using splines with predetermined knots for describing the gradient function. Most of the existing approaches assume known functional forms of the dynamical system; and many of them require data measured on a dense grid (e.g., Varah, 1982; Zhu and Wu, 2007).

For the problems that we are interested in this paper, measurements are taken on a sparse set of points for each sample curve. Thus numerical procedures for solving differential equations can become unstable if we treat each sample curve separately. Moreover, we are more interested in estimating the baseline dynamics than the individual dynamics of each subject. For example, in the plant study described above, we are interested in comparing the growth displacement rates (as a function of distance from the root cap junction) under two different experimental conditions. On the other hand, we are not so interested in the displacement rate corresponding to each plant. Another important aspect in modeling data with multiple subjects is that adequate measures need to be taken to model possible subject-specific effects, otherwise the estimates of model parameters can have inflated variability. Thus in this paper, we incorporate subject-specific effects into the model while combining information across different subjects. In addition, because of insufficient knowledge of the problem as is the case for the plant growth study, in practice one often has to resort to modeling the dynamical system nonparametrically. For example, there is controversy among plant scientists about whether there is a growth bump in the middle of the meristem. There are also some natural boundary constraints of the growth displacement rate, making it hard to specify a simple and interpretable parametric system. (See more discussions in Section 3). Therefore, in this paper, we propose to model the baseline dynamics nonparametrrically through a basis representation approach. We use an estimation procedure that combines nonlinear optimization techniques with a numerical ODE solver to estimate the unknown parameters. In addition, we derive a computationally efficient approximation of the leave-one-curve-out cross validation score for model selection. We prove consistency of the proposed estimators under appropriate regularity conditions. Our asymptotic scenario involves keeping the number of subjects fixed and allowing the number of measurements per subject to grow to infinity. The analysis differs from the usual nonparametric regression problems due to the structures imposed by the differential equations model. We show by simulation studies that the proposed approach can efficiently estimate the baseline dynamics under the setting of multiple replicates per subject with sparse noisy measurements. Moreover, the proposed model selection procedure is effective in maintaining a balance between fidelity to the data and to the underlying model. Finally, we apply the proposed method to the plant data described earlier and compare the estimated growth displacement rates under the two experimental conditions.

The rest of paper is organized as follows. In Section 2, we describe the proposed model. In Sections 3 and 4, we discuss the model fitting and model selection procedures, respectively. In Section 5, we prove consistency of the proposed estimator. In Section 6, we conduct simulation studies to illustrate finite sample performance of the proposed method. Section 7 is the application of this method to the plant data. Technical details are in the appendices. An `R` package `dynamics`

for fitting the model described in this paper is available upon request.

2 Model

In this section, we describe a class of autonomous dynamical systems that is suitable for modeling the problems exemplified by the plant data (Section 1). An autonomous dynamical system has the following general form:

$$X'(t) = f(X(t)), \quad t \in [T_0, T_1].$$

Without loss of generality, henceforth $T_0 = 0$ and $T_1 = 1$. Note that, the above equation means that $X(t) = a + \int_0^t f(X(u))du$, where $a = X(0)$ is the initial condition. Thus in an autonomous system, the dynamics (which is characterized by f) depends on time t only through $X(t)$. This type of systems arises in various scientific studies such as modelling prey-predator dynamics, virus dynamics, or epidemiology (cf. Perthame, 2007). Many studies in plant science such as Silk (1994), Sacks *et al.* (1997), Fraser, Silk and Rost (1990) all suggest reasonably steady growth velocity across the meristem under both normal and water-stress conditions at an early developmental stage. Moreover, exploratory regression analysis based on empirical derivatives and empirical fits of the growth trajectories indicates that time is not a significant predictor and thus an autonomous model is reasonable. This assumption is equivalent to the assertion that the growth displacement rate depends only on the distance from the root cap junction. It means that time zero does not play a role in terms of estimating the dynamical system and there is also no additional variation associated with individual markers.

Figure 3 shows the scatter plot of empirical derivatives versus empirical fits in the treatment group. It indicates that there is an increase in the growth displacement rate starting from a zero rate at the root cap junction, then followed by a nearly constant rate beyond a certain location. This means that growth stops beyond this point and the observed displacements are due to growth in the part of the meristem closer to the root cap junction. Where and how growth stops is of great scientific interest. The scatter plot also indicates excess variability towards the end which is probably caused by plant-specific scaling effects.

Some of the features described above motivate us to consider the following class of autonomous dynamical systems:

$$X'_{il}(t) = g_i(X_{il}(t)), \quad l = 1, \cdots, N_i; i = 1, \ldots, n, \tag{1}$$

where $\{X_{il}(t) : t \in [0,1], l = 1, \cdots, N_i; i = 1, \ldots, n\}$ is a collection of smooth curves corresponding to n subjects, and there are N_i curves associated with the i-th subject. For example, in the plant study, each plant is a subject and each marker corresponds to one growth curve. We assume that, all the curves associated with the same subject follow the same dynamics, and these are described by the functions $\{g_i(\cdot)\}_{i=1}^n$. We also assume that only a snapshot of each curve $X_{il}(\cdot)$ is observed. That is, the observations are given by

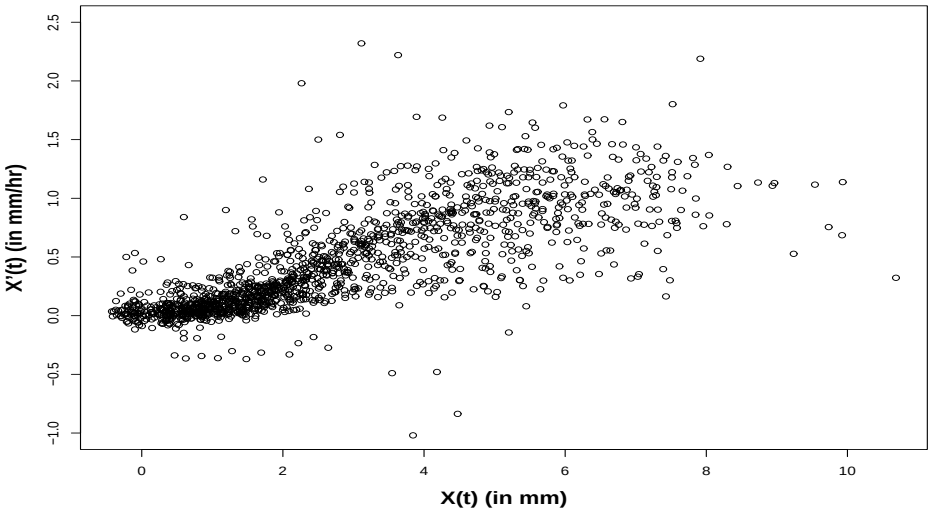
$$Y_{ilj} = X_{il}(t_{ilj}) + \varepsilon_{ilj}, \quad j = 1, \ldots, m_{il}, \tag{2}$$

where $0 \le t_{il1} < \cdots < t_{ilm_{il}} \le 1$ are the observation times for the l^{th} curve of the i^{th} subject, and $\{\varepsilon_{ilj}\}$ are independently and identically distributed noise with mean zero and variance $\sigma_\varepsilon^2 > 0$. In this paper, we model $\{g_i(\cdot)\}_{i=1}^n$ as:

$$g_i(\cdot) = e^{\theta_i} g(\cdot), \quad i = 1, \ldots, n, \tag{3}$$

where

Figure 3: Empirical derivatives (divided differences) $\widehat{X}'(t)$ against empirical fits (averaged measurements) $\widehat{X}(t)$ for treatment group.



(1) the function $g(\cdot)$ reflects the common underlying mechanism regulating all these dynamical systems. It is assumed to be a smooth function and is referred as the *gradient function*. For the plant study, it represents the baseline growth displacement rate for all plants within a given group (i.e., control vs. water-stress).

(2) $\theta_i's$ reflect subject-specific effects in these systems. The mean of θ_i's is assumed to be zero to impose identifiability. In the plant study, $\theta_i's$ represent plant-specific scaling effects in the growth displacement rates for individual plants.

The simplicity and generality of this model make it appealing for modeling a wide class of dynamical systems. First, the gradient function $g(\cdot)$ can be an arbitrary smooth function. If g is nonnegative, and the initial conditions $X_{il}(0)'s$ are also nonnegative, then the sample trajectories are increasing functions, which encompasses growth models that are autonomous. Secondly, the scale parameter e^{θ_i} provides a subject-specific tuning of the dynamics, which is flexible in capturing variations of the dynamics in a population. In this paper, our primary goal is to estimate the gradient function g nonparametrically. For the plant data, the form of g is not known to the biologists, only its behavior at root cap junction and at some later stage of growth are known (Silk, 1994). The fact that the growth displacement rate increases from zero at root cap junction before becoming a constant at a certain (unknown) distance away from the root tip implies that a linear ODE model is apparently not appropriate. Moreover, popular parametric models such as the Michaelis-Menten type either do not satisfy the boundary constraints, and/or have parameters without clear interpretations in the current context. On the other hand, nonparametric modeling provides flexibility and is able to capture features of the dynamical system which are not known to us *a priori* (Section 7). In addition, the nonparametric fit can be used for diagnostics for lack of fit, if realistic parametric models can be proposed.

The gradient function g being smooth means that it can be well approximated by a basis representation approach:

$$g(x) = \sum_{k=1}^{M} \beta_k \phi_{k,M}(x) \tag{4}$$

where $\phi_{1,M}(\cdot), \ldots, \phi_{M,M}(\cdot)$ are linearly independent basis functions, chosen so that their combined support covers the range of the observed trajectories. For example, we can use cubic splines with a suitable set of knots. Thus, for a given choice of the basis functions, the unknown parameters in the model are the basis coefficients $\boldsymbol{\beta} := (\beta_1, \ldots, \beta_M)^T$, the scale parameters $\boldsymbol{\theta} := \{\theta_i\}_{i=1}^n$, and possibly the initial conditions $\boldsymbol{a} := \{a_{il} := X_{il}(0) : l = 1, \cdots, N_i\}_{i=1}^n$. Also, various model parameters, such as the number of basis functions M and the knot sequence, need to be selected based on the data. Therefore, in essence, this is a nonlinear, semi-parametric, mixed effects model.

In the plant data, g is nonnegative and thus a modeling scheme imposing this constraint may be more advantageous. However, the markers are all placed at a certain distance from the root cap junction, where the growth displacement rate is already positive, and the total number of measurements per plant is moderately large. These mean that explicitly imposing nonnegativity is not crucial for the plant data. Indeed, with the imposition of the boundary constraints, the estimate of g turns out to be nonnegative over the entire domain of the measurements (Section 7). In general, if g is strictly positive over the domain of interest, then we can model the logarithm of g by basis representation. Also, in this case, the dynamical system is stable in the sense that there is no bifurcation phenomenon (Strogatz, 2001).

3 Model Fitting

In this section, we propose an iterative estimation procedure that imposes regularization on the estimate of $\boldsymbol{\theta}$ and possibly $\boldsymbol{a}$. One way to achieve this is to treat them as unknown random parameters from some parametric distributions. Specifically, we use the following set of working assumptions: (i) a_{il}'s are independent and identically distributed as $N(\alpha, \sigma_a^2)$ and θ_i's are independent and identically distributed as $N(0, \sigma_\theta^2)$, for some $\alpha \in \mathbb{R}$ and $\sigma_a^2 > 0, \sigma_\theta^2 > 0$; (ii) the noise ε_{ilj}'s are independent and identically distributed as $N(0, \sigma_\varepsilon^2)$ for $\sigma_\varepsilon^2 > 0$; (iii) the three random vectors $\boldsymbol{a}$, $\boldsymbol{\theta}$, $\boldsymbol{\varepsilon} := \{\varepsilon_{ilj}\}$ are independent. Under these assumptions, the negative joint log-likelihood of the observed data $Y := \{Y_{ilj}\}$, the scale parameters $\boldsymbol{\theta}$ and the initial conditions $\boldsymbol{a}$ is, up to an additive constant and a positive scale constant,

$$\sum_{i=1}^{n} \sum_{l=1}^{N_i} \sum_{j=1}^{m_{il}} [Y_{ilj} - \widetilde{X}_{il}(t_{ilj}; a_{il}, \theta_i, \beta)]^2 + \lambda_1 \sum_{i=1}^{n} \sum_{l=1}^{N_i} (a_{il} - \alpha)^2 + \lambda_2 \sum_{i=1}^{n} \theta_i^2, \tag{5}$$

where $\lambda_1 = \sigma_\varepsilon^2/\sigma_a^2$, $\lambda_2 = \sigma_\varepsilon^2/\sigma_\theta^2$, and $\widetilde{X}_{il}(\cdot)$ is the trajectory determined by a_{il}, θ_i, and $\boldsymbol{\beta}$. This can be viewed as a hierarchical maximum likelihood approach (Lee, Nelder and Pawitan, 2006), which is considered to be a convenient alternative to the full (restricted) maximum likelihood approach. Define

$$\ell_{ilj}(a_{il}, \theta_i, \boldsymbol{\beta}) := [Y_{ilj} - \widetilde{X}_{il}(t_{ilj}; a_{il}, \theta_i, \boldsymbol{\beta})]^2 + \lambda_1 (a_{il} - \alpha)^2 / m_{il} + \lambda_2 \theta_i^2 / \sum_{l=1}^{N_i} m_{il} \ .$$

Then the loss function in (5) equals $\sum_{i=1}^{n} \sum_{l=1}^{N_i} \sum_{j=1}^{m_{il}} \ell_{ilj}(a_{il}, \theta_i, \boldsymbol{\beta})$. Note that the above distributional assumptions are simply working assumptions. The expression in (5) can also be viewed as

a regularized ℓ_2 loss with penalties on the variability of $\boldsymbol{\theta}$ and $\boldsymbol{a}$. For the plant data, the initial conditions (markers) are chosen according to some fixed experimental design, thus it is natural to treat them as fixed effects. Moreover, it does not seem appropriate to shrink the estimates toward some common value in this case. Thus in Section 7, we set $\lambda_1 = 0$ when estimating $\boldsymbol{a}$. For certain other problems, treating the initial conditions as random effects may be more suitable. For example, Huang, Liu and Wu (2006) study a problem of HIV dynamics where the initial conditions are subject-specific and unobserved.

In many situations, there are boundary constraints on the gradient function g. For example, according to plant science, both the growth displacement rate and its derivative at the root cap junction should be zero. Moreover, it should become a constant at a certain (unknown) distance from the root cap junction. Thus for the plant data, it is reasonable to assume that, $g(0) = 0 = g'(0)$ and $g'(x) = 0$ for $x \geq A$ for a given $A > 0$. The former can be implemented by an appropriate choice of the basis functions. For the latter, we consider constraints of the form: $\boldsymbol{\beta}^T \mathbf{B} \boldsymbol{\beta}$ for an $M \times M$ positive semi-definite matrix $\mathbf{B}$, which can be thought of as an ℓ_2-type constraint on some derivative of g. (See Section 7 for the specification of $\mathbf{B}$). Consequently, the modified objective function becomes

$$L(a, \boldsymbol{\theta}, \boldsymbol{\beta}) := \sum_{i=1}^{n} \sum_{l=1}^{N_i} \sum_{j=1}^{m_{il}} \ell_{ilj}(a_{il}, \theta_i, \boldsymbol{\beta}) + \boldsymbol{\beta}^T \mathbf{B} \boldsymbol{\beta}. \tag{6}$$

The proposed estimator is then the minimizer of the objective function:

$$(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}}) := \arg\min_{\boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\beta}} L(\boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\beta}). \tag{7}$$

Note that, here our main interest is the gradient function g. Thus estimating the parameters of the dynamical system together with the sample trajectories and their derivatives simultaneously is most efficient. In contrast, if the trajectories and their derivatives are first obtained via pre-smoothing (as is done for example in Chen and Wu (2008a, 2008b), Varah (1982)), and then used in a nonparametric regression framework to obtain g, it will be inefficient in estimating g. This is because, errors introduced in the pre-smoothing step cause loss of information which is not retrievable later on, and also information regarding g is not efficiently combined across curves.

In the following, we propose a numerical procedure for solving (7) that has two main ingredients:

- Given $(\boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\beta})$, reconstruct the trajectories $\{\widetilde{X}_{il}(\cdot) : l = 1, \cdots, N_i\}_{i=1}^n$ and their derivatives. This step can be carried out using a numerical ODE solver, such as the 4^{th} order Runge-Kutta method (cf. Tenenbaum and Pollard, 1985).

- Minimize (6) with respect to $(\boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\beta})$. This amounts to a nonlinear least squares problem (Bates and Watts, 1988). It can be carried out using either a nonlinear least squares solver, like the Levenberg-Marquardt method; or a general optimization procedure, such as the Newton-Raphson algorithm.

The above procedure bears some similarity to the local, or gradient-based, methods discussed in Miao *et al.* (2008).

We now briefly describe an optimization procedure based on the *Levenberg-Marquardt method* (cf. Nocedal and Wright, 2006). For notational convenience, denote the current estimates by $\boldsymbol{a}^* := \{a_{il}^*\}$, $\boldsymbol{\theta}^* := \{\theta_i^*\}$ and $\boldsymbol{\beta}^*$, and define the current residuals as: $\tilde{\varepsilon}_{ilj} = Y_{ilj} - \widetilde{X}_{il}(t_{ilj}; a_{il}^*, \theta_i^*, \boldsymbol{\beta}^*)$. For each $i = 1, \cdots, n$ and $l = 1, \cdots, N_i$, define the $m_{il} \times 1$ column vectors

$$\mathbf{J}_{il, a_{il}^*} := \left(\frac{\partial}{\partial a_{il}} \widetilde{X}_{il}(t_{ilj}; a_{il}^*, \theta_i^*, \boldsymbol{\beta}^*) \right)_{j=1}^{m_{il}}, \quad \widetilde{\boldsymbol{\varepsilon}}_{il} = (\widetilde{\varepsilon}_{ilj})_{j=1}^{m_{il}}.$$

For each $i = 1, \cdots, n$, define the $m_{i\cdot} \times 1$ column vectors

$$\mathbf{J}_{i,\theta_i^*} = \left(\frac{\partial}{\partial \theta_i} \widetilde{X}_{il}(t_{ilj}; a_{il}^*, \theta_i^*, \boldsymbol{\beta}^*) \right)_{j=1,l=1}^{m_{il},N_i} ; \quad \widetilde{\boldsymbol{\varepsilon}}_i = (\widetilde{\varepsilon}_{ilj})_{j=1,l=1}^{m_{il},N_i},$$

where $m_{i\cdot} := \sum_{l=1}^{N_i} m_{il}$ is the total number of measurements of the i^{th} cluster. Finally, for each $k = 1, \cdots, M$, define the $m_{\cdot\cdot} \times 1$ column vectors:

$$\mathbf{J}_{\beta_k^*} = \left(\frac{\partial}{\partial \beta_k} \widetilde{X}_{il}(t_{ilj}; a_{il}^*, \theta_i^*, \boldsymbol{\beta}^*) \right)_{j=1,l=1,i=1}^{m_{il},N_i,n} ; \quad \widetilde{\boldsymbol{\varepsilon}} = (\widetilde{\varepsilon}_{ilj})_{j=1,l=1,i=1}^{m_{il},N_i,n},$$

where $m_{\cdot\cdot} := \sum_{i=1}^{n} \sum_{l=1}^{N_i} m_{il}$ is the total number of measurements. Note that, given $\boldsymbol{a}^*$,$\boldsymbol{\theta}^*$ and $\boldsymbol{\beta}^*$, the trajectories $\{\widetilde{X}_{il}\}'s$ and their gradients (as well as Hessians) can be easily evaluated on a fine grid by using numerical ODE solvers such as the 4^{th} order Runge-Kutta method as mentioned above (see Appendix A).

We break the updating step into three parts corresponding to the three different sets of parameters. For each set of parameters, we first derive a first order Taylor expansion of the curves $\{\widetilde{X}_{il}\}$ around the current values of these parameters and then update them by a least squares fitting, while keeping the other two sets of parameters fixed at the current values. The equation for updating $\boldsymbol{\beta}$, while keeping $\boldsymbol{a}^*$ and $\boldsymbol{\theta}^*$ fixed, is

$$\left[J_{\boldsymbol{\beta}^*}^T J_{\boldsymbol{\beta}^*} + \lambda_3 \, \mathrm{diag}(J_{\boldsymbol{\beta}^*}^T J_{\boldsymbol{\beta}^*}) + \mathbf{B} \right] (\boldsymbol{\beta} - \boldsymbol{\beta}^*) = J_{\boldsymbol{\beta}^*}^T \widetilde{\boldsymbol{\varepsilon}} - \mathbf{B}\boldsymbol{\beta}^*,$$

where $J_{\boldsymbol{\beta}^*} := (J_{\beta_1^*} : \cdots : J_{\beta_M^*})$ is an $m_{\cdot\cdot} \times M$ matrix. Here λ_3 is a sequence of positive constants converging to zero as the number of iterations increases. They are used to avoid possible singularities in the system of equations. The normal equation for updating θ_i is

$$(\mathbf{J}_{i,\theta_i^*}^T \mathbf{J}_{i,\theta_i^*} + \lambda_2)(\theta_i - \theta_i^*) = \mathbf{J}_{i,\theta_i^*}^T \widetilde{\boldsymbol{\varepsilon}}_i - \lambda_2 \theta_i^*, \quad i = 1, \ldots, n. \tag{8}$$

The equation for updating a_{il} is derived similarly, while keeping θ_i and $\boldsymbol{\beta}$ fixed at θ_i^*, $\boldsymbol{\beta}^*$:

$$(\mathbf{J}_{il,a_{il}^*}^T \mathbf{J}_{il,a_{il}^*} + \lambda_1)(a_{il} - a_{il}^*) = \mathbf{J}_{il,a_{il}^*}^T \widetilde{\boldsymbol{\varepsilon}}_{il} + \lambda_1 \alpha_{il}^*, \quad l = 1, \cdots, N_i, \quad i = 1, \cdots, n, \tag{9}$$

where $\alpha^* = \sum_{i=1}^{n} \sum_{l=1}^{N_i} a_{il}^* / N_\cdot$, $\alpha_{il}^* = \alpha^* - a_{il}^*$ with $N_\cdot := \sum_{i=1}^{n} N_i$ being the total number of sample curves.

In summary, this procedure begins by taking initial estimates and then iterates by cycling through the updating steps for $\boldsymbol{\beta}$, $\boldsymbol{\theta}$ and $\boldsymbol{a}$ until convergence. The initial estimates can be conveniently chosen. For example, $a_{il}^{ini} = Y_{il1}$, $\theta_i^{ini} \equiv 0$; or $a_{il}^{ini} \equiv \frac{1}{N_\cdot} \sum_{i=1}^{n} \sum_{l=1}^{N_i} Y_{i1l}$. Even though the model is identifiable, in practice, for small n, there can be drift in the estimates of θ_i and g due to flatness of the objective function in some regions. To avoid this and increase stability, we also impose the condition that $\sum_{i=1}^{n} \theta_i^* = 0$. This can be easily achieved by subtracting $\bar{\theta}^* := \frac{1}{n} \sum_{i=1}^{n} \theta_i^*$ from θ_i^* at each iteration after updating $\{\theta_i\}$.

All three updating steps described above are based on the general principle of Levenberg-Marquardt algorithm by the linearization of the curves $\{\widetilde{X}_{il}\}$ (see Appendix B). However, the tuning parameter λ_3 plays a different role than the penalty parameters λ_1 and λ_2. The parameter λ_3 is used to stabilize the updates of $\boldsymbol{\beta}$ and thereby facilitate convergence. Thus it needs to decrease to zero with increasing iterations in order to avoid introducing bias in the estimate. There are ways

of implementing this adaptively (see e.g. Nocedal and Wright, 2006, Ch. 10). In this paper, we use a simple non-adaptive method: $\lambda_{3j} = \lambda_3^0/j$ for the j-th iteration, for some pre-specified $\lambda_3^0 > 0$. On the other hand, λ_1 and λ_2 are parts of the penalized loss function (6). Their main role is to control the bias-variance trade-off of the estimators, even though they also help in regularizing the optimization procedure. From the likelihood view point, λ_1, λ_2 are determined by the variances σ_ε^2, σ_a^2 and σ_θ^2. After each loop over all the parameter updates, we can estimate these variances from the current residuals and current values of $\boldsymbol{a}$ and $\boldsymbol{\theta}$. By assuming that $m_{il} > 2$ for each pair (i, l),

$$\begin{aligned}
\widehat{\sigma}_\varepsilon^2 &= \frac{1}{m_{\cdot\cdot} - N_{\cdot} - n - M} \sum_{i=1}^{n} \sum_{l=1}^{N_i} \sum_{j=1}^{m_{il}} \widetilde{\varepsilon}_{ilj}^2, \\
\widehat{\sigma}_a^2 &= \frac{1}{N_{\cdot} - 1} \sum_{i=1}^{n} \sum_{l=1}^{N_i} (a_{il}^* - \alpha^*)^2, \quad \widehat{\sigma}_\theta^2 = \frac{1}{n-1} \sum_{i=1}^{n} (\theta_i^*)^2.
\end{aligned}$$

We can then plug in the estimates $\widehat{\sigma}_\varepsilon^2$, $\widehat{\sigma}_a^2$ and $\widehat{\sigma}_\theta^2$ to get new values of λ_1 and λ_2 for the next iteration. On the other hand, if we take the penalized loss function view point, we can simply treat λ_1, λ_2 as fixed regularization parameters, and then use a model selection approach to select their values based on data. In the following sections, we refer the method as `adaptive` if λ_1, λ_2 are updated after each iteration; and refer the method as `non-adaptive` if they are kept fixed throughout the optimization.

The Levenberg-Marquardt method is quite stable and robust to the initial estimates. However, it converges slowly in the neighborhood of the minima of the objective function. On the other hand, the Newton-Raphson algorithm has a very fast convergence rate when starting from estimates that are already near the minima. Thus, in practice the we first use the Levenberg-Marquardt approach to obtain a reasonable estimate, and then use the Newton-Raphson algorithm to expedite the search of the minima. The implementation of the Newton-Raphson algorithm of the current problem is standard and is outlined in Appendix C.

4 Model Selection

After specifying a scheme for the basis functions $\{\phi_{k,M}(\cdot)\}$, we still need to determine various model parameters such as the number of basis functions M, the knot sequence, etc. In the literature AIC/BIC/AICc criteria have been proposed for model selection while estimating dynamical systems with nonparametric time-dependent components (e.g. Miao *et al.*, 2008). Here, we propose an approximate leave-one-curve-out cross-validation score for model selection. Under the current context, the leave-one-curve-out CV score is defined as

$$CV := \sum_{i=1}^{n} \sum_{l=1}^{N_i} \sum_{j=1}^{m_{il}} \ell_{ilj}^{cv}(\widehat{a}_{il}^{(-il)}, \widehat{\theta}_i^{(-il)}, \widehat{\boldsymbol{\beta}}^{(-il)}) \tag{10}$$

where $\widehat{\theta}_i^{(-il)}$ and $\widehat{\boldsymbol{\beta}}^{(-il)}$ are estimates of θ_i and $\boldsymbol{\beta}$, respectively, based on the data after dropping the l^{th} curve in the i^{th} cluster; and $\widehat{a}_{il}^{(-il)}$ is the minimizer of $\sum_{j=1}^{m_{il}} \ell_{ilj}(a_{il}, \widehat{\theta}_i^{(-il)}, \widehat{\boldsymbol{\beta}}^{(-il)})$ with respect to a_{il}. The function ℓ_{ilj}^{cv} is a suitable criterion function for cross validation. Here, we use the prediction error loss:

$$\ell_{ilj}^{cv}(a_{il}, \theta_i, \boldsymbol{\beta}) := \left(Y_{ilj} - \widetilde{X}_{il}(t_{ilj}; a_{il}, \theta_i, \boldsymbol{\beta})\right)^2.$$

Calculating CV score (10) is computationally very demanding. Therefore, we propose to approximate $\widehat{\theta}_i^{(-il)}$ and $\widehat{\boldsymbol{\beta}}^{(-il)}$ by a first order Taylor expansion around the estimates $\widehat{\theta}_i, \widehat{\boldsymbol{\beta}}$ based on the full data. Consequently we derive an approximate CV score which is computationally inexpensive. A similar approach is taken in Peng and Paul (2009) under the context of functional principal component analysis. Observe that, when evaluated at the estimate $\widehat{\boldsymbol{a}}$, $\widehat{\boldsymbol{\theta}}$ and $\widehat{\boldsymbol{\beta}}$ based on the full data,

$$\frac{\partial}{\partial \theta_i}\left(\sum_{l,j} \ell_{ilj}^{cv}\right) + 2\lambda_2\theta_i = 0, \quad i = 1, \cdots, n; \quad \frac{\partial}{\partial \boldsymbol{\beta}}\left(\sum_{i,l,j} \ell_{ilj}^{cv}\right) + 2\mathbf{B}\boldsymbol{\beta} = 0. \tag{11}$$

Whereas, when evaluated at the drop (i,l)-estimates: $\widehat{a}_{il}^{(-il)}, \widehat{\theta}_i^{(-il)}, \widehat{\boldsymbol{\beta}}^{(-il)}$,

$$\frac{\partial}{\partial \theta_i}\left(\sum_{l^*, j: l^* \neq l} \ell_{il^*j}^{cv}\right) + 2\lambda_2\theta_i = 0; \quad \frac{\partial}{\partial \boldsymbol{\beta}}\left(\sum_{i^*, l^*, j: (i^*, l^*) \neq (i,l)} \ell_{i^*l^*j}^{cv}\right) + 2\mathbf{B}\boldsymbol{\beta} = 0. \tag{12}$$

Expanding the left hand side of (12) around $\widehat{\boldsymbol{\beta}}$, we obtain

$$\begin{aligned} 0 &\approx \sum_{i^*, l^*, j: (i^*, l^*) \neq (i,l)} \frac{\partial}{\partial \boldsymbol{\beta}} \ell_{i^*l^*j}^{cv} \Big|_{\widehat{\boldsymbol{\beta}}} + 2\mathbf{B}\widehat{\boldsymbol{\beta}} + \left[\sum_{i^*, l^*, j: (i^*, l^*) \neq (i,l)} \frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} \ell_{i^*l^*j}^{cv} \Big|_{\widehat{\boldsymbol{\beta}}} + 2\mathbf{B}\right](\widehat{\boldsymbol{\beta}}^{(-il)} - \widehat{\boldsymbol{\beta}}) \\ &\approx -\sum_{j=1}^{m_{il}} \frac{\partial \ell_{ilj}^{cv}}{\partial \boldsymbol{\beta}} \Big|_{(\widehat{a}_{il}, \widehat{\theta}_i, \widehat{\boldsymbol{\beta}})} + \left[\sum_{i^*, l^*, j: (i^*, l^*) \neq (i,l)} \frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} \ell_{i^*l^*j}^{cv} \Big|_{(\widehat{a}_{i^*l^*}, \widehat{\theta}_{i^*}, \widehat{\boldsymbol{\beta}})} + 2\mathbf{B}\right](\widehat{\boldsymbol{\beta}}^{(-il)} - \widehat{\boldsymbol{\beta}}), \end{aligned}$$

where in the second step we invoked (11) and approximated $\{\widehat{a}_{il}^{(-il)}\}, \{\widehat{\theta}_i^{(-il)}\}$ by $\{\widehat{a}_{il}\}$, $\{\widehat{\theta}_i\}$, respectively. Similar calculations are carried out for $\widehat{\theta}_i^{(-il)}$. Thus we obtain the following first order approximations:

$$\begin{aligned} \widehat{\theta}_i^{(-il)} &\approx \widetilde{\theta}_i^{(-il)} := \widehat{\theta}_i + \left[\sum_{l'=1}^{N_i} \sum_{j'=1}^{m_{il'}} \frac{\partial^2 \ell_{il'j'}^{cv}}{\partial \theta_i^2} + 2\lambda_2\right]^{-1} \sum_{j=1}^{m_{il}} \left(\frac{\partial \ell_{ilj}^{cv}}{\partial \theta_i}\right) \\ \widehat{\boldsymbol{\beta}}^{(-il)} &\approx \widetilde{\boldsymbol{\beta}}^{(-il)} := \widehat{\boldsymbol{\beta}} + \left[\sum_{i'=1}^{n} \sum_{l'=1}^{N_{i'}} \sum_{j'=1}^{m_{i'l'}} \frac{\partial^2 \ell_{i'l'j'}^{cv}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} + 2\mathbf{B}\right]^{-1} \left(\sum_{j=1}^{m_{il}} \frac{\partial \ell_{ilj}^{cv}}{\partial \boldsymbol{\beta}}\right). \end{aligned} \tag{13}$$

These gradients and Hessians are all evaluated at $(\widehat{\boldsymbol{a}}, \widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}})$, and thus they have already been computed (on a fine grid) in the course of obtaining these estimates. Thus, there is almost no additional computational cost to obtain these approximations. Now for $i = 1, \ldots, n; l = 1, \cdots, N_i$, define

$$\widetilde{a}_{il}^{(-il)} = \arg\min_a \sum_{j=1}^{m_{il}} (Y_{ilj} - \widetilde{X}_{il}(t_{ilj}; a, \widetilde{\theta}_i^{(-il)}, \widetilde{\boldsymbol{\beta}}^{(-il)}))^2 + \lambda_1(a - \widehat{\alpha})^2, \tag{14}$$

where $\widehat{\alpha}$ is the estimator of α obtained from the full data. Finally, the approximate leave-one-curve-out cross-validation score is

$$\widetilde{CV} := \sum_{i=1}^{n} \sum_{l=1}^{N_i} \sum_{j=1}^{m_{il}} \ell_{ilj}^{cv}(\widetilde{a}_{il}^{(-il)}, \widetilde{\theta}_i^{(-il)}, \widetilde{\boldsymbol{\beta}}^{(-il)}). \tag{15}$$

5 Asymptotic Theory

In this section, we present a result on the consistency of the proposed estimator of g under suitable technical conditions. We assume that the number of subjects n is fixed; and the number of measurements per curve m_{il}, and number of curves N_i per subject, increase to infinity together. When n is fixed, the asymptotic analysis is similar irrespective of whether θ_i's are viewed as fixed effects or random effects. Hence, for simplicity, we treat θ_i's as fixed effects and impose the identifiability constraint $\theta_1 = 0$. Due to this restriction, we modify the loss function (5) slightly by replacing the penalty $\lambda_2 \sum_{i=1}^n \theta_i^2$ with $\lambda_2 \sum_{i=2}^n (\theta_i - \overline{\theta})^2$ where $\overline{\theta} = \sum_{i=2}^n \theta_i/(n-1)$. Moreover, since n is finite, in practice we can relabel the subjects so that the curves corresponding to subject 1 has the highest rate of growth, and hence $\theta_i \leq 0$ for all $i > 1$. This relabeling is not necessary but simplifies the arguments considerably.

Moreover, to be consistent with the setting of the plant data, we focus on the case where the time points for the different curves corresponding to the same subject are the same, so that, in particular, $m_{il} \equiv m_i$. We assume that the time points come from a common continuous distribution F_T. We also assume that the gradient function $g(x)$ is positive for $x > 0$ and is defined on a domain $D = [x_0, x_1] \subset \mathbb{R}^+$; and the initial conditions $\{a_{il} := X_{il}(0)\}'s$ are observed (and hence $\lambda_1 = 0$) and are randomly chosen from a common continuous distribution F_a with support $[x_0, x_2]$ where $x_2 < x_1$.

Before we state the regularity conditions required for proving the consistency result, we highlight two aspects of the asymptotic analysis. Note that, the current problem differs from standard semiparametric nonlinear mixed effects models. First, the estimation of g is an inverse problem, since it implicitly requires knowledge of the derivatives of the trajectories of the ODE which are not directly observed. The degree of ill-posedness is quantified by studying the behavior of the expected Jacobian matrix of the sample trajectory with respect to $\boldsymbol{\beta}$. This matrix would be well-conditioned under a standard nonparametric function estimation context. However, in the current case, its condition number goes to infinity with the dimension of the model space M. Secondly, unlike in standard nonparametric function estimation problems where the effect of the estimation error is localized, the estimation error propagates throughout the entire domain of g through the dynamical system. Therefore, sufficient knowledge of the behavior of g at the boundaries is imperative.

We assume the following:

A1 $g \in C^p(D)$ for some integer $p \geq 4$, where $D = [x_0, x_1] \subset \mathbb{R}^+$.

A2 θ_i's are fixed parameters with $\theta_1 = 0$.

A3 The collection of basis functions $\Phi_M := \{\phi_{1,M}, \ldots, \phi_{M,M}\}$ satisfies: (i) $\phi_{k,M} \in C^2(D)$ for all k; (ii) $\sup_{x \in D} \sum_{k=1}^M |\phi_{k,M}^{(j)}(x)|^2 = O(M^{1+2j})$, for $j = 0, 1, 2$; (iii) for every k, the length of the support of $\phi_{k,M}$ is $O(M^{-1})$; (iv) for every M, there is a $\boldsymbol{\beta}^* \in \mathbb{R}^M$ such that $\| g - \sum_{k=1}^M \beta_k^* \phi_{k,M} \|_{L^\infty(D)} = O(M^{-2p})$; $\| g' - \sum_{k=1}^M \beta_k^* \phi'_{k,M} \|_{L^\infty(D)} = O(M^{-c})$, for some $c > 0$; $\sum_{k=1}^M \beta_k^* \phi''_{k,M}$ is Lipschitz with Lipschitz constant $O(M)$; and $\| \sum_{k=1}^M \beta_k^* \phi''_{k,M} \|_{L^\infty(D)} = O(1)$.

A4 $X_{il}(0)$'s are i.i.d. from a continuous distribution F_a. Denote $\text{supp}(F_a) = [x_0, x_2]$ and let Θ be a fixed, open interval containing the true θ_i's, denoted by θ_i^*. Then there exists a $\tau > 0$ such that for all $a \in F_a$ and for all $\theta \in \Theta$, the initial value problem

$$x'(t) = e^\theta f(x(t)), \quad x(0) = a \tag{16}$$

has a solution $x(t) := x(t; a, \theta, f)$ on $[0,1]$ for all $f \in \mathcal{M}(g,\tau)$, where

$$\mathcal{M}(g,\tau) := \{f \in C^1(D) :\| f - g \|_{1,D} \leq \tau\}.$$

Moreover, the range of $x(\cdot;\cdot,\cdot,f)$ (as a mapping from $[0,1] \times \text{supp}(F_a) \times \Theta$) is contained in $D \pm \epsilon(\tau)$ for some $\epsilon(\tau) > 0$ (with $\lim_{\tau \to 0} \epsilon(\tau) = 0$) for all $f \in \mathcal{M}(g,\tau)$. Here, $\| \cdot \|_{1,D}$ is the seminorm defined by $\| f \|_{1,D} = \| f \|_{L^\infty(D)} + \| f' \|_{L^\infty(D)}$. Furthermore, the range of $x(\cdot;\cdot,0,g)$ contains D.

A5 For each $i = 1, \ldots, n$, for all $l = 1, \ldots, N_i$, the time points t_{ilj} $(j = 1, \ldots, m_i)$ belong the set $\{T_{i,j'} : 1 \leq j' \leq m_i\}$. And $\{T_{i,j'}\}$ are i.i.d. from the continuous distribution F_T supported on $[0,1]$ with a density f_T satisfying $c_1 \leq f_T \leq c_2$ for some $0 < c_1 \leq c_2 < \infty$. Moreover, $\overline{m} := \sum_{i=1}^n m_i/n \to \infty$ as $\overline{N} := \sum_{i=1}^n N_i/n \to \infty$. Also, both N_i's and m_i's increase to infinity uniformly meaning that $\max_i N_i / \min_i N_i$ and $\max_i m_i / \min_i m_i$ remain bounded.

A6 Define $X_{il}(\cdot; X_{il}(0), \theta_i, \boldsymbol{\beta})$ to be the solution of the initial value problem

$$x'(t) = e^{\theta_i} \sum_{k=1}^M \beta_k \phi_{k,M}(x(t)), \quad t \in [0,1], \quad x(0) = X_{il}(0). \tag{17}$$

Let $X_{il}^{\theta_i}(\cdot;\theta_i,\boldsymbol{\beta})$ and $X_{il}^{\boldsymbol{\beta}}(\cdot;\theta_i,\boldsymbol{\beta})$ be its partial derivatives with respect to parameters θ_i and $\boldsymbol{\beta}$. And let $\boldsymbol{\beta}^* \in \mathbb{R}^M$ be as in **A3**. Define $G^i_{*,\theta\theta} := \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\beta}^*}(X^{\theta_i}_{i1}(T_{i,1};\theta_i^*,\boldsymbol{\beta}^*))^2$, $G^i_{*,\beta\theta} := \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\beta}^*}\left(X^{\theta_i}_{i1}(T_{i,1};\theta_i^*,\boldsymbol{\beta}^*)X^{\boldsymbol{\beta}}_{i1}(T_{i,1};\theta_i^*,\boldsymbol{\beta}^*)\right)$ and $G^i_{*,\beta\beta} := \mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\beta}^*}\left(X^{\boldsymbol{\beta}}_{i1}(T_{i,1};\theta_i^*,\boldsymbol{\beta}^*)(X^{\boldsymbol{\beta}}_{i1}(T_{i,1};\theta_i^*,\boldsymbol{\beta}^*))^T\right)$, where $\mathbb{E}_{\boldsymbol{\theta}^*,\boldsymbol{\beta}^*}$ denotes the expectation over the joint distribution of $(X_{i1}(0), T_{i,1})$ evaluated at $\theta_i = \theta_i^*$ and $\beta = \beta^*$. Define $G_{*,\theta\theta} = diag(G^i_{*,\theta\theta})_{i=2}^n$, $G_{*,\beta\theta} = \left[G^2_{*,\beta\theta} : \cdots : G^n_{*,\beta\theta}\right]$, and $G_{*,\beta\beta} = \sum_{i=1}^n G^i_{*,\beta\beta}$. Then, there exists a function κ_M and a constant $c_3 \in (0,\infty)$, such that,

$$\| (G_{*,\beta\beta})^{-1} \| \leq \kappa_M \quad \text{and} \quad \| (G_{*,\theta\theta})^{-1} \| \leq c_3. \tag{18}$$

A7 The noise ε_{ilj}'s are i.i.d. $N(0, \sigma_\varepsilon^2)$ with σ_ε^2 bounded above.

Before stating the main result, we give a brief explanation of these assumptions. **A1** ensures enough smoothness of the solution paths of the differential equation (16). It also ensures that the approximation error, when g is approximated in the basis Φ_M, is of an appropriate order. Condition **A3** is satisfied when we approximate g using the $(p-1)$-th order B-splines with equally spaced knots on the interval D which are normalized so that $\int_D \phi_{k,M}(x)^2 dx = 1$ for all k. Note that $g_{\boldsymbol{\beta}^*}$ can be viewed as an optimal approximation of g in the space generated by Φ_M. Condition **A4** ensures that a solution of (16) exist for all f of the form $g_{\boldsymbol{\beta}}$ with $\boldsymbol{\beta}$ sufficiently close to $\boldsymbol{\beta}^*$. This implies that we can apply the perturbation theory of differential equations to bound the fluctuations of the sample paths due to a perturbation of the parameters. Condition **A5** ensures that the time-points $\{T_{i,j}\}$ cover the domain D randomly and densely, and that there is a minimum amount of information per sample curve in the data. Condition **A6** is about the estimability of a parameter (in this case g) in a semiparametric problem in the presence of nuisance parameters (in this case $\{\theta_i\}$). Indeed, the matrix $G_{*,\beta\beta} - G_{*,\beta\theta}(G_{*,\theta\theta})^{-1}G_{*,\theta\beta}$ plays the role of the information matrix for $\boldsymbol{\beta}$ at $(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*)$. Equation (18) essentially quantifies the degree of ill-conditionedness of the information matrix for $\boldsymbol{\beta}$. Note that **A4** together with **A6** implicitly imposes a restriction on the magnitude of $\| g' \|_{L_\infty(D)}$. Condition **A6** has further implications. Unlike in parametric problems, where the

information matrix is typically well-conditioned, we have $\kappa_M \to \infty$ in our setting (see Theorem 2 and Proposition 1 below). Note that in situations when $g \geq 0$ and the initial conditions are nonnegative, one can simplify **A6** considerably, since then we can obtain explicit formulas for the derivatives of the sample paths (see Appendix A). And then one can easily verify the second part of equation (18).

Theorem 1: *Assume that the data follow the model described by equations (1), (2) and (3) with* $\theta_1 = 0$. *Suppose that the true gradient function* g, *the distributions* F_a *and* F_T, *and the collection of basis functions* Φ_M *satisfy* **A1**-**A7**. *Suppose further that* g *is strictly positive over* $D = [x_0, x_1]$. *Suppose that* $\{X_{il}(0)\}$ *are known (so that* $\lambda_1 = 0$*),* $\lambda_2 = o(\alpha_N \overline{N}\overline{m}\kappa_M^{-1})$ *and the sequence* $M = M(\overline{N}, \overline{m})$ *is such that* $\min\{\overline{N}, \overline{m}\} \gg \kappa_M M \log(\overline{N}\overline{m})$, $\kappa_M M^{-(p-1)} \to 0$, *and* $\alpha_N \max\{\kappa_M M^{1/2}, \kappa_M^{1/2} M^{3/2}\} \to 0$ *as* $\overline{N}, \overline{m} \to \infty$, *where* $\alpha_N \geq C \max\{\sigma_\varepsilon \kappa_M^{1/2} M^{1/2} (\overline{N}\overline{m})^{-1/2}, \kappa_M^{1/2} M^{-p}\}$ *for some sufficiently large constant* $C > 0$. *Then there exists a minimizer* $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}})$ *of the objective function (5) such that if* $\widehat{g} := \sum_{k=1}^{M} \widehat{\beta}_k \phi_{k,M}$, *then the following holds with probability tending to 1:*

$$\int_D |\widehat{g}(x) - g(x)|^2 dx \leq \alpha_N^2 + O(M^{-2p}), \qquad \sum_{i=2}^{n} |\widehat{\theta}_i - \theta_i^*|^2 \leq \alpha_N^2. \tag{19}$$

As explained earlier, κ_M is related to the inverse of the smallest eigenvalue of the matrix

$$\begin{bmatrix} G_{*,\boldsymbol{\beta}\boldsymbol{\beta}} & G_{*,\boldsymbol{\beta}\boldsymbol{\theta}} \\ G_{*,\boldsymbol{\theta}\boldsymbol{\beta}} & G_{*,\boldsymbol{\theta}\boldsymbol{\theta}} \end{bmatrix}$$

In order to show that our method leads to a consistent estimator of g, we need to know the behavior of κ_M as $M \to \infty$. The following result quantifies the behavior when we choose a B-spline basis with equally spaced knots inside the domain D.

Theorem 2: *Suppose that* $supp(F_a) = [x_0, x_2] \subset \mathbb{R}^+$ *and* g *is strictly positive over the domain* $D = [x_0, x_1]$. *Suppose also that the (normalized) B-splines of order* ≥ 2 *are used as basis functions* $\{\phi_{k,M}\}$ *where the knots are equally spaced on the interval* $[x_0 + \delta, x_1 - \delta]$, *for some small constant* $\delta > 0$. *Then* $\kappa_M = O(M^2)$.

The condition that the knots are in the interior of the domain D is justified if the function g is completely known on the set $[x_0, x_0 + \delta] \cup [x_1 - \delta, x_1]$. Then this information can be used to modulate the B-splines near the boundaries so that all the properties listed in **A3** still hold and we have the appropriate order of the approximations. We conjecture that the same result ($\kappa_M = O(M^2)$) still holds even if g is known only up to a parametric form near the boundaries, and a combination of the parametric form and B-splines with equally spaced knots is used to represent it. If instead the distribution F_a is such that near the end points (x_0 and x_2) of the support of F_a, the density behaves like $(x - x_0)^{-1+\gamma}$ and $(x_2 - x)^{-1+\gamma}$, for some $\gamma \in (0, 1]$, then it can be shown that (Proposition 1) $\kappa_M = O(M^{2+2\gamma})$. Thus, in the worst case scenario, we can only guarantee that $\kappa_M = O(M^4)$. In that case g needs to have a higher order of smoothness ($g \in C^{6+\epsilon}(D)$, for some $\epsilon > 1/2$), and higher-order (at least seventh order) B-splines are needed to ensure consistency.

It can be shown that under mild conditions κ_M should be at least $O(M^2)$. Thus, the condition $\alpha_N \max\{\kappa_M M^{1/2}, \kappa_M^{1/2} M^{3/2}\} = o(1)$ can be simplified to $\kappa_M \alpha_N M^{1/2} = o(1)$. When $\kappa_M \asymp M^2$, Theorem 1 holds with $p = 4$, so that $g \in C^4$ and cubic B-splines can be used. Moreover, under that setting as long as $\overline{m}/\overline{N}$ is bounded both above and below and σ_ε is bounded below, then

$\min\{\overline{N}, \overline{m}\} \gg \kappa_M M \log(\overline{N}\overline{m})$. The following proposition states the dependence of κ_M on the behavior of the density of the distribution F_a.

Proposition 1: *Assume that the density of F_a behaves like $(x - x_0)^{-1+\gamma}$ and $(x_2 - x)^{-1+\gamma}$, near the endpoints x_0 and x_2, for some $\gamma \in (0, 1]$, and is bounded away from zero in the interior. Then $\kappa_M = O(M^{2+2\gamma})$.*

The proof of Theorem 1 involves a second order Taylor expansion of loss function around the *optimal parameter* $(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*)$. We apply results on perturbation of differential equations (cf. Deuflhard and Bornemann, 2002, Ch. 3) to bound the *bias terms* $|X_{il}(t_{ilj}; a_{il}, \theta_i, f) - X_{il}(t_{ilj}; a_{il}, \theta_i, g)|$ for arbitrary θ_i and functions f, g. The same approach also allows us to provide bounds for various terms involving partial derivatives of the sample paths with respect to the parameters in the aforementioned Taylor expansion. Proof of Theorem 2 involves an inequality (*Halerpin-Pitt inequality*) on bounding the square integral of a function by the square integrals of its derivatives (Mitrinovic, Pecaric and Fink, 1991, p. 8). The detailed proofs are given in Appendix E.

6 Simulation

In this section, we conduct a simulation study to demonstrate the effectiveness of the proposed estimation and model selection procedures. In the simulation, the true gradient function g is represented by $M_* = 4$ cubic B-spline basis functions with knots at $(0.35, 0.6, 0.85, 1.1)$ and basis coefficients $\boldsymbol{\beta} = (0.1, 1.2, 1.6, 0.4)^T$. It is depicted by the solid curve in Figure 4. We consider two different settings for the number of measurements per curve: `moderate` case – m_{il}'s are independently and identically distributed as Uniform$[5, 20]$; `sparse` case – m_{il}'s are independently and identically distributed as Uniform$[3, 8]$. Measurement times $\{t_{ilj}\}$ are independently and identically distributed as Uniform$[0, 1]$. The scale parameters θ_i's are randomly sampled from $N(0, \sigma_\theta^2)$ with $\sigma_\theta = 0.1$; and the initial conditions a_{il}'s are randomly sampled from a $c_a \chi^2_{k_a}$ distribution, with $c_a, k_a > 0$ chosen such that $\alpha = 0.25, \sigma_a = 0.05$. Finally, the residuals ε_{ilj}'s are randomly sampled from $N(0, \sigma_\varepsilon^2)$ with $\sigma_\varepsilon = 0.01$. Throughout the simulation, we set the number of subjects $n = 10$ and the number of curves per subject $N_i \equiv N = 20$. Observations $\{Y_{ilj}\}$ are generated using the model specified by equations (1) - (4) in Section 2. For all the settings, 50 independent data sets are used to evaluate the performance of the proposed procedure.

In the estimation procedure, we consider cubic B-spline basis functions with knots at points $0.1 + (1 : M)/M$ to model g, where M varies from 2 to 6. The Levenberg-Marqardt step is chosen to be `non-adaptive`, and the Newton-Raphson step is chosen to be `adaptive` (see Section 3 for the definition of `adaptive` and `non-adaptive`). We examine three different sets of initial values for λ_1 and λ_2: (i) $\lambda_1 = \sigma_\varepsilon^2/\sigma_a^2 = 0.04, \lambda_2 = \sigma_\varepsilon^2/\sigma_\theta^2 = 0.01$ ("true" values); (ii) $\lambda_1 = 0.01, \lambda_2 = 0.0025$ ("deflated" values); (iii) $\lambda_1 = 0.16, \lambda_2 = 0.04$ ("inflated" values). It turns out that the estimation and model selection procedures are quite robust to the initial choice of (λ_1, λ_2), thereby demonstrating the effectiveness of the `adaptive` method used in the Newton-Raphson step. Thus in the following, we only report the results when the "true" values are used.

We also compare results when (i) the initial conditions $\boldsymbol{a}$ are known, and hence not estimated; and (ii) when $\boldsymbol{a}$ are estimated. As can be seen from Table 1, the estimation procedure converges well and the true model ($M_* = 4$) is selected most of the times for all the cases. Mean integrated squared error (MISE) and Mean squared prediction error (MSPE) and the corresponding standard deviations, SD(ISE) and SD(SPE), based on 50 independent data sets, are used for measuring the estimation accuracy of $\widehat{g}$ and $\widehat{\boldsymbol{\theta}}$, respectively. Since the true model is selected most of the times, we

Table 1: Convergence and model selection based on 50 independent replicates.

		a known					**a** estimated				
	Model	2	3	**4**	5	6	2	3	**4**	5	6
`moderate`	Number converged	50	50	50	50	50	50	7	50	50	46
	Number selected	0	0	46	1	3	0	0	49	1	0
`sparse`	Number converged	50	50	50	50	50	50	5	49	44	38
	Number selected	0	0	45	0	5	1	0	47	1	1

Table 2: Estimation accuracy under the true model*

		MISE($\widehat{g}$)	SD(ISE)	MSPE($\widehat{\theta}$)	SD(SPE)
a known	`moderate`	0.069	0.072	0.085	0.095
	`sparse`	0.072	0.073	0.085	0.095
a estimated	`moderate`	0.088	0.079	0.086	0.095
	`sparse`	0.146	0.129	0.087	0.094

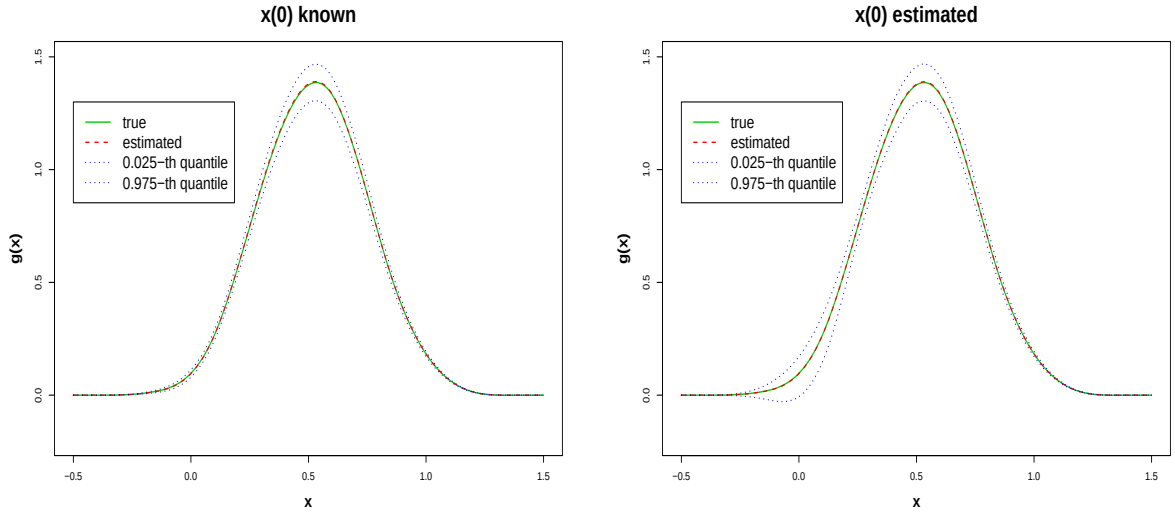
* All numbers are multiplied by 100

only report results under the true model in Table 2. As can be seen from this table, when the initial conditions $\boldsymbol{a}$ are known, there is not much difference of the performance between the `moderate case` and the `sparse` case. On the other hand, when $\boldsymbol{a}$ are not known, the advantages of having more measurements become much more prominent. In Figure 4, we have a visual comparison of the fits when the initial conditions $\boldsymbol{a}$ are known versus when they are estimated in the `sparse` case. In the `moderate` case, there is very little visual difference under these two settings. We plot the true g (solid green curve), the pointwise mean of $\widehat{g}$ (broken red curve), and 2.5% and 97.5% pointwise quantiles (dotted blue curves) under the true model. These plots show that both fits are almost unbiased. Also, when $\boldsymbol{a}$ are estimated, there is greater variability in the estimated g at smaller values of x, partly due to scarcity of data in that region. Overall, as can be seen from these tables and figures, the proposed estimation and model selection procedures perform effectively. Moreover, with sufficient information, explicitly imposing nonnegativity in the model does not seem to be crucial: for the `moderate` and/or "$\boldsymbol{a}$ known" cases the resulting estimators of g are always nonnegative.

7 Application: Plant Growth Data

In this section, we apply the proposed method to the plant growth data from Sacks *et al.* (1997) described in the earlier Sections. The data consist of measurements on ten plants from a control group and nine plants from a treatment group where the plants are under water stress. The primary roots had grown for approximately 18 hours in the normal and stressed conditions before the measurements were taken. The roots were marked at different places using a water-soluble marker and high-resolution photographs were used to measure the displacements of the marked places. The measurements were in terms of distances from the root cap junction (in millimeters) and were taken for each of these marked places, hereafter markers, over an approximate 12-hour

Figure 4: True and fitted gradient functions for the `sparse` case. Left panel: $\boldsymbol{a}$ known; Right panel: $\boldsymbol{a}$ estimated.



period while the plants were growing. Note that, measurements were only taken in the meristem. Thus whenever a marker moved outside of the meristem, its displacement would not be recorded at later times anymore. This, together with possible technical failures (in taking measurements), is the reason why in Figure 2 some growth trajectories were cut short. A similar, but more sophisticated, data acquisition technique is described in Walter et al. (2002), who study the diurnal pattern of root growth in maize. Van der Weele et al. (2003) describe a more advanced data acquisition technique for measuring the expansion profile of a growing root at a high spatial and temporal resolution. They also propose computational methods for estimating the growth velocity from this dense image data. Basu et al. (2007) develop a a new image-analysis technique to study spatio-temporal patterns of growth and curvature of roots that tracks the displacement of particles on the root over space and time. These methods, while providing plant scientists with valuable information, are limited in that, they do not provide an inferential framework and they require very dense measurements. Our method, even though designed to handle sparse data, is potentially applicable to these data as well.

Consider the model described in Section 2. For the control group, we have the number of curves per subject N_i varying in between 10 and 29; and for the water stress group, we have $12 \leq N_i \leq 31$. The observed growth displacement measurements $\{Y_{ilj} : j = 1, \ldots, m_{il}, l = 1, \ldots, N_i\}_{i=1}^{n}$ are assumed to follow model (2), where m_{il} is the number of measurements taken for the i^{th} plant at its l^{th} marker, which varies between 2 and 17; and $\{t_{ilj} : j = 1, \cdots, m_{il}\}$ are the times of measurements, which are in between $[0, 12]$ hours. Altogether, for the control group there are 228 curves with a total of 1486 measurements and for the treatment group there are 217 curves with 1712 measurements in total. We are interested in comparing the baseline growth displacement rate between the treatment and control groups.

As discussed earlier, there are natural constraints for the plant growth dynamics. Theoretically, $g(0) = 0 = g'(0)$ and $g'(x) = 0$ for $x \geq A$ for some constant $A > 0$. For the former constraint, we

can simply omit the constant and linear terms in the spline basis. And for the latter constraint, in the objective function (6) we use

$$\boldsymbol{\beta}^T \mathbf{B} \boldsymbol{\beta} := \lambda_R \int_A^{2A} (g'(x))^2 dx = \lambda_R \boldsymbol{\beta}^T [\int_A^{2A} \phi'(x)(\phi'(x))^T dx] \boldsymbol{\beta}$$

where $\phi = (\phi_{1,M}, \ldots, \phi_{M,M})^T$ and λ_R is a large positive number quantifying the severity of this constraint; and $A > 0$ determines where the growth displacement rate becomes a constant. A and λ_R are both adaptively determined by the model selection scheme discussed in Section 4. Moreover, as discussed earlier, since it is not appropriate to shrink the initial conditions $\{a_{il}\}$ towards a fixed number, we set $\lambda_1 = 0$ in the loss function (6).

We first describe a simple regression-based method for getting a crude initial estimate of the function $g(\cdot)$, as well as selecting a candidate set of knots. This involves (i) computing the re-scaled empirical derivatives $e^{-\widehat{\theta}_i^{(0)}} \widehat{X}'_{ilj}$ of the sample curves from the data, where the empirical derivatives are defined by taking divided differences: $\widehat{X}'_{ilj} := (Y_{il(j+1)} - Y_{ilj})/(t_{il(j+1)} - t_{ilj})$, and $\widehat{\theta}_i^{(0)}$ is a preliminary estimate of θ_i; and (ii) regressing the re-scaled empirical derivatives onto a set of basis functions evaluated at the corresponding sample averages: $\widehat{X}_{ilj} := (Y_{il(j+1)} + Y_{ilj})/2$. In this paper, we use the basis $\{x^2, x^3, (x - x_k)_+^3\}_{k=1}^K$ with a pre-specified, dense set of knots $\{x_k\}_{k=1}^K$. Then, a model selection procedure, like the stepwise regression, with either AIC or BIC criterion, can be used to select a set of candidate knots. In the following, we shall refer this method as `stepwise-regression`. The resulting estimate of g and the selected knots can then act as a starting point for the proposed procedure. We expect this simple method to work reasonably well only when the number of measurements per curve is at least moderately large. Comparisons given later (Figure 7) demonstrate a clear superiority of the proposed method over this simple approach.

Next, we fit the model to the control group and the treatment group separately. For the control group, we first fit models with g represented in cubic B-splines with equally spaced knot sequence $1 + 11.5(1 : M)/M$ for $M = 2, 3, 4, \cdots, 12$. At this stage, we set $\boldsymbol{\beta}^{ini} = \mathbf{1}_M$, $\boldsymbol{\theta}^{ini} = \mathbf{0}_n$, $\boldsymbol{a}^{ini} = (X_{il}(t_{il1}) : l = 1, \ldots, N_i)_{i=1}^n$. For Levenberg-Marquardt step, we fix $\lambda_1 = 0$ and $\lambda_2 = 0.0025$; and we update λ_1, λ_2 adaptively in the Newton-Raphson step. The criterion based on the approximate CV score (15) selects the model with $M = 9$ basis functions (see Appendix D). This is not surprising since when equally spaced knots are used, usually a large number of basis functions are needed to fit the data adequately. In order to get a more parsimonious model, we consider the `stepwise-regression` method to obtain an initial estimate of g as well as finding a candidate set of knots. We use 28 equally spaced candidate knots on the interval $[0.5, 14]$ and use the fitted values $\{\widehat{\theta}_i\}_{i=1}^{10}$ from the previous fit. The AIC criterion selects 11 knots. We then consider various submodels with knots selected from this set of 11 knots and fit the corresponding models again using the procedure described in Section 3. Specifically, we first apply the Levenberg-Marquardt procedure with λ_1, λ_2 fixed at $\lambda_1 = 0$ and $\lambda_2 = (\widehat{\sigma}_\varepsilon^{ini})^2/(\widehat{\sigma}_\theta^{ini})^2 = 0.042$, respectively, where $\widehat{\sigma}_\varepsilon^{ini}$ and $\widehat{\sigma}_\theta^{ini}$ are obtained from the `stepwise-regression` fit. Then, after convergence of $\boldsymbol{\beta}$ up to a desired precision (threshold of 0.005 for $\| \boldsymbol{\beta}^{old} - \boldsymbol{\beta}^{new} \|$), we apply the Newton-Raphson procedure with λ_1 fixed at zero, but λ_2 adaptively updated from the data. The approximate CV scores for various submodels are reported in Table 3. The parameters A and λ_R are also varied and selected by the approximate CV score. Based on the approximate CV score, the model with knot sequence $(3.0, 4.0, 6.0, 9.0, 9.5)$ and $(A, \lambda_R) = (9, 10^5)$ is selected. A similar procedure is applied to the treatment group. It turns out that the model with knot sequence $(3.0, 3.5, 7.5)$ performs considerably better than other candidate models, and hence we only report the approximate CV

Table 3: Model selection for real data. Control group: approximate CV scores for four *submodels* of the model selected by the AIC criterion in the `stepwise-regression` step. `M1`: knots = $(3.0, 4.0, 5.0, 6.0, 9.0, 9.5)$; `M2`: knots = $(3.0, 4.0, 5.5, 6.0, 9.0, 9.5)$; `M3`: knots = $(3.0, 4.0, 6.0, 9.0, 9.5)$; `M4`: knots = $(3.0, 4.5, 6.0, 9.0, 9.5)$. Treatment group: approximate CV scores for the model `M`: knots = $(3.0, 3.5, 7.5)$.

		$\lambda_R = 10^3$			$\lambda_R = 10^5$		
Control	Model	$A = 8.5$	$A = 9$	$A = 9.5$	$A = 8.5$	$A = 9$	$A = 9.5$
	`M1`	53.0924	53.0877	53.1299	54.6422	53.0803	53.1307
	`M2`	53.0942	53.0898	53.1374	54.5190	53.0835	53.1375
	`M3`	53.0300	53.0355	53.0729	53.8769	**53.0063**	53.0729
	`M4`	53.0420	53.0409	53.0723	54.0538	53.0198	53.0722
Treatment	Model	$A = 7$	$A = 7.5$	$A = 8$	$A = 7$	$A = 7.5$	$A = 8$
	`M`	**64.9707**	64.9835	64.9843	65.5798*	64.9817	64.9817

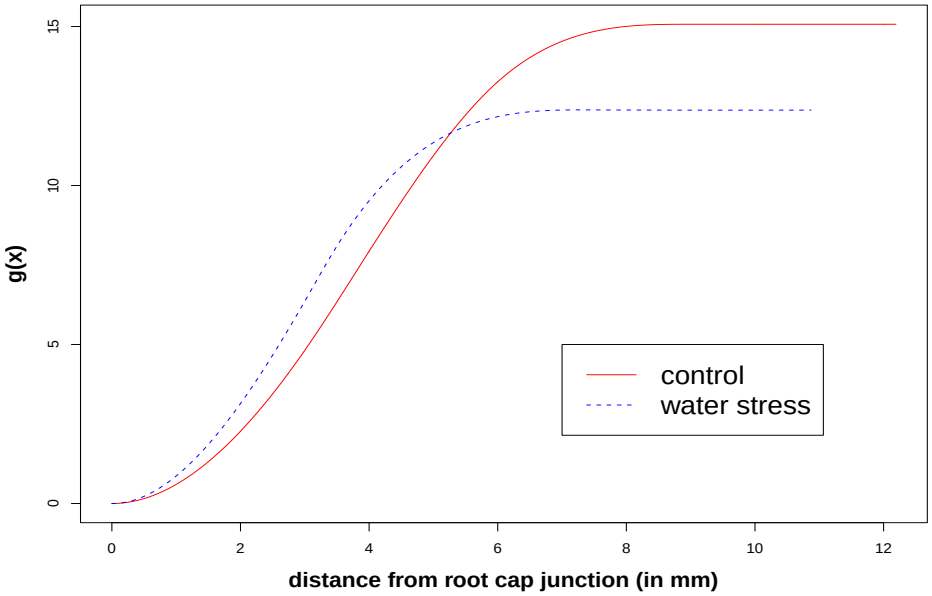
* no convergence

scores under this model in Table 3 with various choices of (A, λ_R). It can be seen that, $(A, \lambda_R) = (7, 10^3)$ has the smallest approximate CV score.

Figure 5 shows the estimated gradient functions $\widehat{g}$ under the selected models for the control and treatment groups, respectively. First of all, there is no growth bump observed for either group. This plot also indicates that different dynamics are at play for the two groups. In the part of the meristem closer to the root cap junction (distance within $\sim$ 5.5mm), the growth displacement rate for the treatment group is higher than that for the control group. This is probably due to the greater cell elongation rate under water stress condition in this part of the meristem so that the root can reach deeper in the soil to get enough water. This is a known phenomenon in plant science. The growth displacement rate for the treatment group flattens out beyond a distance of about 6 mm from the root cap junction. The same phenomenon happens for the control group, however at a further distance of about 8 mm from the root cap junction. Also, the final constant growth displacement rate of the control group is higher than that of the treatment group. This is due to the stunting effect of water stress on these plants, which results in an earlier stop of growth and a slower cell division rate. Figure 6 shows the estimated relative elemental growth rates (i.e., $\widehat{g}'$) for these two groups. Relative elemental growth rate (REGR) relates the magnitude of growth directly to the location along the meristem. For both groups, the growth is fastest in the middle part of the meristem ($\sim$ 3.8 mm for control group and $\sim$ 3.1 for treatment group), and then growth dies down pretty sharply and eventually stops. Again, we observe a faster growth in the part of the meristem closer to the root cap junction for the water stress group and the growth dies down more quickly compared to the control group. The shape of the estimated g may suggest that it might be modeled by a logistic function with suitably chosen location and scale parameters, even though the scientific meaning of these parameters is unclear and the boundary constraints are not satisfied exactly. As discussed earlier, there is insufficient knowledge from plant science to suggest a functional form beforehand. This points to one major purpose of nonparametric modeling, which is to provide insight and to suggest candidate parametric models for further study.

Figure 7 shows the residual versus time plot for the treatment group. The plot for the control group is similar and thus is omitted. This plot shows that the procedure based on minimizing the

Figure 5: Fitted gradient functions under the selected models for control and treatment (water-stress) groups, respectively.



objective function (6) has much smaller and more evenly spread residuals (SSE = 64.50) than the fit by `stepwise-regression` (SSE = 147.57), indicating a clear benefit of the more sophisticated approach. Overall, by considering the residual plots and CV scores, the estimation and model selection procedures give reasonable fits under both experimental conditions. Note that, for the first six hours, the residuals (right panel of Figure 7) show some time-dependent pattern, which is not present for later times. Since throughout the whole 12 hour period, the residuals remain small compared to the scale of the measurements, the autonomous system approximation seems to be adequate for practical purposes. Modeling growth dynamics through nonautonomous systems may enable scientists to determine the stages of growth that are not steady across a region of the root. This is a topic of future research.

Acknowledgement

Peng and Paul are partially supported by NSF-DMS grant 0806128. The authors would like to thank Professor Wendy Silk of the Department of Land, Air and Water Resources, University of California, Davis, for providing the data used in the paper and for helpful discussions on the scientific aspects of the problem.

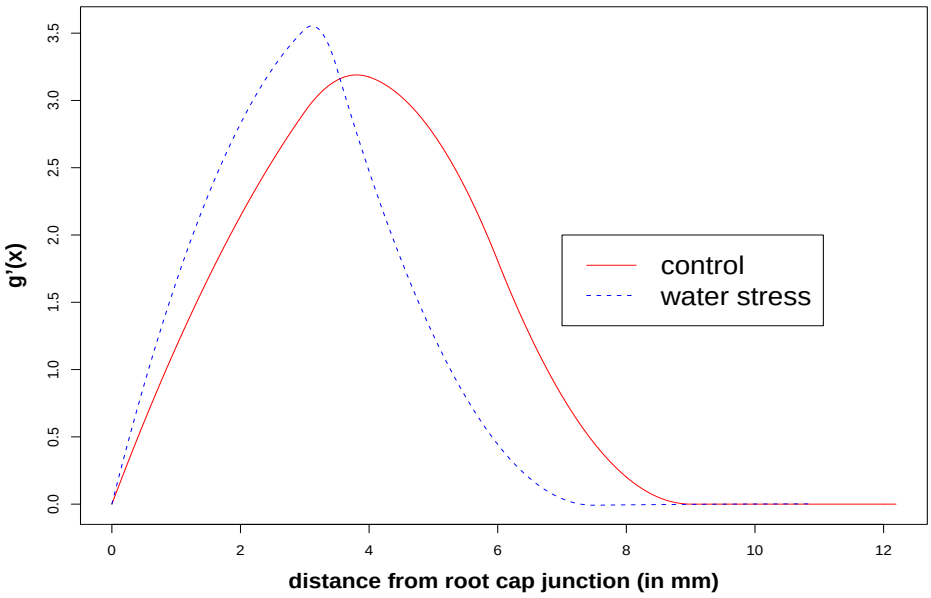
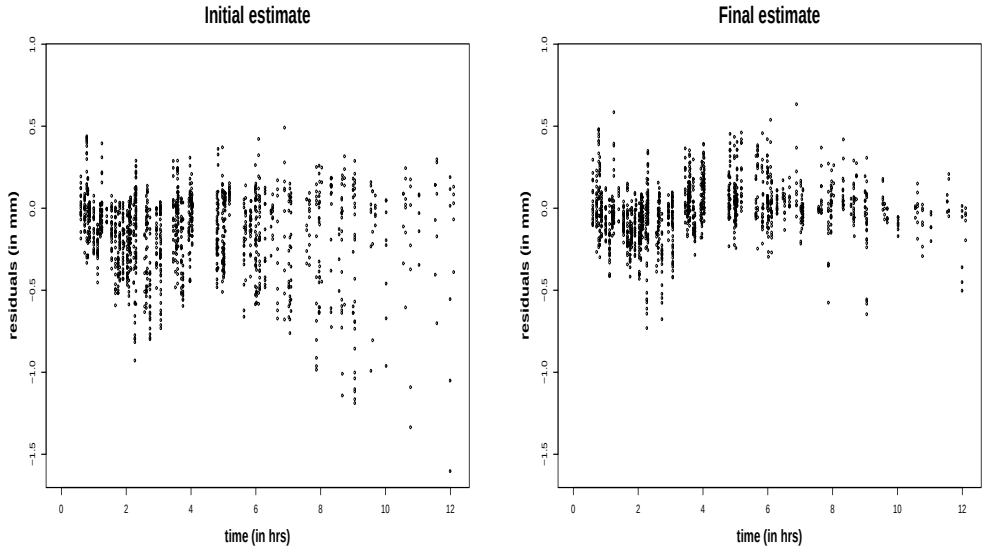


Figure 6: Fitted relative elemental growth rate (REGR) under the selected models for control and treatment groups, respectively.

Figure 7: Residual versus time plots for the treatment group. Left panel: fit by `stepwise-regression`; Right panel: fit by the proposed method.



References

1. Bates, D. M. and Watts, D. G. (1988). *Nonlinear Regression and Its Applications.* Wiley, New York.

2. Basu, P., Pal, A., Lynch, J. P. and Brown, K. M. (2007). A novel image-analysis technique for kinematic study of growth and curvature. *Plant Physiology* **145**, 305-316.

3. Cao, J., Fussmann, G. F. and Ramsay, J. O. (2008). Estimating a predator-prey dynamical model with the parameter cascades method. *Biometrics* **64**, 959-967.

4. Chen, J. and Wu, H. (2008a). Estimation of time-varying parameters in deterministic dynamic models with application to HIV infections. To appear in *Statistica Sinica* **18**, 987-1006.

5. Chen, J. and Wu, H. (2008b). Efficient local estimation for time-varying coefficients in deterministic dynamic models with applications to HIV-1 dynamics. *Journal of American Statistical Association* **103**, 369-384.

6. Chicone, C. (2006). *Ordinary Differential Equations with Applications.* Springer.

7. de Boor, C. (1978). *A Practical Guide to Splines.* SpringerVerlag.

8. Deuflhard, P. and Bornemann, F. (2002). *Scientific Computing with Ordinary Differential Equations.* Springer.

9. Fraser, T. K., Silk, W. K. and Rost, T. L. (1990). Effects of low water potential on cortical cell length in growing regions of maize roots. *Plant Physiology* **93**, 648-651.

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

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

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

13. Ljung, L. and Glad, T. (1994). *Modeling of Dynamical Systems.* Prentice Hall.

14. Miao, H., Dykes, C., Demeter, L. M. and Wu, H. (2008). Differential equation modeling of HIV viral fitness experiments : model identification, model selection, and multimodel inference. *Biometrics* (to appear).

15. Mitrinovic, D. S., Pecaric, J. E. and Fink, A. M. (1991). *Inequalities Involving Functions and Their Integrals and Derivatives.* Kluwer Academic Publishers.

16. Nocedal, J. and Wright, S. J. (2006). *Numerical Optimization, 2nd Ed.* Springer.

17. Peng, J. and Paul, D. (2009). A geometric approach to maximum likelihood estimation of the functional principal components from sparse longitudinal data. To appear in *Journal of Computational and Graphical Statistics.* (`arXiv:0710.5343v1`). Also available at

    `http://anson.ucdavis.edu/~jie/pd-cov-likelihood-technical.pdf`

18. Perthame, B. (2007). *Transport Equations in Biology.* Birkhäuser.

19. Poyton, A. A., Varziri, M. S., McAuley, K. B., McLellan, P. J. and Ramsay, J. O. (2006). Parameter estimation in continuous dynamic models using principal differential analysis. *Computers & Chemical Engineering* **30**, 698-708.

20. Ramsay, J. O., Hooker, G., Campbell, D. and Cao, J. (2007). Parameter estimation for differential equations: a generalized smoothing approach. *Journal of the Royal Statistical Society, Series B* **69**, 741-796.

21. Sacks, M. M., Silk, W. K. and Burman, P. (1997). Effect of water stress on cortical cell division rates within the apical meristem of primary roots of maize. *Plant Physiology* **114**, 519-527.

22. Schurr, U., Walter, A. and Rascher, U. (2006). Functional dynamics of plant growth and photosynthesis – from steady-state to dynamics – from homogeneity to heterogeneity. *Plant, Cell and Environment* **29**, 340-352.

23. Sharp, R. E., Silk, W. K. and Hsiao, T. C. (1988). Growth of the maize primary root at low water potentials. *Plant Physiology* **87**, 50-57.

24. Silk, W. K., and Erickson, R. O. (1979). Kinametics of plant growth. Journal of Theoretical Biology **76**, 481-501.

25. Silk, W. K. (1994). Kinametics and dynamics of primary growth. *Biomimectics* **2**(3), 199-213.

26. Strogatz, S. H. (2001). *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering.* Perseus Books Group.

27. Tenenbaum, M., and Pollard, H. (1985). *Ordinary Differential Equations.* Dover.

28. Van der Weele, C. M., Jiang, H. S., Krishnan K. P., Ivanov, V. B., Palaniappan, K., and Baskin, T. I. (2003). A new algorithm for computational image analysis of deformable motion at high spatial and temporal resolution applied to root growth. roughly uniform elongation in the meristem and also, after an abrupt acceleration, in the elongation zone. *Plant Physiology* **132**, 1138-1148.

29. Varah, J. M. (1982). A spline least squares method for numerical parameter estimation in differential equations. *SIAM Journal on Scientific Computing* **3**, 28-46.

30. Walter, A., Spies, H., Terjung, S., Küsters, R., Kirchgebner, N. and Schurr, U. (2002). Spatio-temporal dynamics of expansion growth in roots: automatic quantification of diurnal course and temperature response by digital image sequence processing. *Journal of Experimental Botany* **53**, 689-698.

31. Wu, H., Ding, A. and DeGruttola, V. (1998). Estimation of HIV dynamic parameters. *Statistics in Medicine* **17**, 2463-2485.

32. Wu, H. and Ding, A. (1999). Population HIV-1 dynamics in vivo : applicable models and inferential tools for virological data from AIDS clinical trials. *Biometrics* **55**, 410-418.

33. Zhu, H. and Wu, H. (2007). Estimating the smooth time-varying parameters in state space models. *Journal of Computational and Graphical Statistics* **20**, 813-832.

Appendix A : Reconstruction of $X_{il}(\cdot)$ and its derivatives

In this section, we describe how to evaluate the (i,l)-th sample trajectory $X_{il}(\cdot)$ and its derivatives given $\boldsymbol{\beta}, \theta_i$ and a_{il} on a fine grid. For notational simplicity, we omit the dependence of the trajectories $X_{il}(\cdot)$ on the parameters $(\boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\beta})$, and drop the subscript M from $\phi_{k,M}$.

Note that, $X_{il}(\cdot)$ satisfies the first order ODE

$$\frac{d}{dt}X_{il}(t) = e^{\theta_i}\sum_{k=1}^{M}\beta_k\phi_k(X_{il}(t)), \quad X_{il}(0) = a_{il}, \quad t \in [0,1]. \tag{20}$$

Or equivalently

$$X_{il}(t) = a_{il} + \int_0^t e^{\theta_i}\sum_{k=1}^{M}\beta_k\phi_k(X_{il}(s))ds, \quad t \in [0,1]. \tag{21}$$

We first describe a numerical procedure (4^{th} order Runge-Kutta method) for constructing the sample trajectories $\widetilde{X}_{il}(t)$ and their derivatives (with respect to the parameters) on a pre-specified fine grid.

Runge-Kutta method: the general procedure

Suppose that a family of first order ODE is described in terms of the parameters generically denoted by $\boldsymbol{\eta} = (\eta_1, \eta_2)$, where η_1 denotes the initial condition and η_2 can be vector-valued:

$$\frac{d}{dt}f(t) = G(t, f(t), \eta_2), \quad f(0) = \eta_1, \quad t \in [0,1]. \tag{22}$$

where $G(t, x, \eta_2)$ is a smooth function. Denote the solution for this family of ODE as $f(t, \boldsymbol{\eta})$. Given the function G and the parameter $\boldsymbol{\eta}$, $f(t, \boldsymbol{\eta})$ can be solved numerically by an ODE solver. One of the commonly used approaches to solve such an initial value problem is the 4^{th} order Runge-Kutta method. For a pre-specified small value $h > 0$, the 4^{th} order Runge-Kutta method proceeds as follows:

1. Initial step: define $y_0 = \eta_1$ and $t_0 = 0$;
2. Iterative step: in the $m+1$ step (for $0 \le m < [1/h]$), define $y_{m+1} = y_m + \frac{h}{6}(k_1 + 2k_2 + 2k_3 + k_4)$, and $t_{m+1} = t_m + h$, where

$$\begin{aligned} k_1 &= G(t_m, y_m, \eta_2) \\ k_2 &= G\left(t_m + \frac{h}{2}, y_m + \frac{h}{2}k_1, \eta_2\right) \\ k_3 &= G\left(t_m + \frac{h}{2}, y_m + \frac{h}{2}k_2, \eta_2\right) \\ k_4 &= G(t_m + h, y_m + hk_3, \eta_2). \end{aligned}$$

3. Final step: set $f(t_m, \boldsymbol{\eta}) = y_m$ for $m = 0, \cdots, [1/h]$.

Thus, at the end we obtain an evaluation (approximation) of $f(\cdot, \boldsymbol{\eta})$ on the grid points $\{0, h, 2h, \cdots, \}$.

Note that $f(t,\boldsymbol{\eta})$ satisfies,

$$f(t,\boldsymbol{\eta}) = \eta_1 + \int_0^t G(s, f(s,\boldsymbol{\eta}), \eta_2)ds, \quad t \geq 0. \tag{23}$$

Partially differentiating $f(t,\boldsymbol{\eta})$ with respect to $\boldsymbol{\eta}$ and taking derivatives inside the integral, we obtain

$$\begin{aligned}
\frac{\partial}{\partial\eta_1} f(t,\boldsymbol{\eta}) &= 1 + \int_0^t \frac{\partial}{\partial\eta_1} f(s,\boldsymbol{\eta}) G_f(s, f(s,\boldsymbol{\eta}), \eta_2)ds, &(24)\\
\frac{\partial}{\partial\eta_2} f(t,\boldsymbol{\eta}) &= \int_0^t \left[\frac{\partial}{\partial\eta_2} f(s,\boldsymbol{\eta}) G_f(s, f(s,\boldsymbol{\eta}), \eta_2) + G_\eta(s, f(s,\boldsymbol{\eta}), \eta_2)\right] ds, &(25)
\end{aligned}$$

where G_f and G_η denote the partial derivatives of G with respect to its second and third arguments, respectively. In equations (24) and (25), if we view the $f(\cdot,\boldsymbol{\eta})$ inside G_f, G_η as known , $\frac{\partial}{\partial\eta_1} f(t,\boldsymbol{\eta})$ is the solution of the first order ODE

$$\frac{d}{dt}p(t) = H(t, p(t), \eta_2), \quad p(0) = 1, \quad t \in [0,1],$$

where $H(t,x,\eta_2) = xG_f(t, f(t,\boldsymbol{\eta}), \eta_2)$. Similarly, $\frac{\partial}{\partial\eta_2} f(t,\boldsymbol{\eta})$ is the solution of the first order ODE with $p(0) = 0$ and $H(t,x,\eta_2) = xG_f(t, f(t,\boldsymbol{\eta}), \eta_2) + G_\eta(t, f(t,\boldsymbol{\eta}), \eta_2)$. Thus, given the function G and the parameter $\boldsymbol{\eta}$, a general strategy for numerically computing $f(\cdot,\boldsymbol{\eta})$ and its gradient $\frac{\partial}{\partial\boldsymbol{\eta}} f(\cdot,\boldsymbol{\eta})$ on a fine grid is to first use the Runge-Kutta method to approximate the solution to (23), and then using that approximate solution in place of $f(\cdot,\boldsymbol{\eta})$ in equations (24) and (25) to compute the gradients by another application of the Runge-Kutta method. Note that, if we evaluate $f(\cdot,\boldsymbol{\eta})$ on the grid points $\{0, h, 2h, \cdots\}$, by the above procedure, we will obtain the gradients $\frac{\partial}{\partial\boldsymbol{\eta}} f(\cdot,\boldsymbol{\eta})$ on a rougher grid: $\{0, 2h, 4h, \cdots\}$.

Derivatives of the sample paths $\{X_{il}(\cdot)\}$ with respect to $(\boldsymbol{a}, \boldsymbol{\theta}, \boldsymbol{\beta})$

Differentiating (20) with respect to the parameters, we have

$$\begin{aligned}
X_{il}^{a_{il}}(t) := \frac{\partial X_{il}(t)}{\partial a_{il}} &= 1 + \int_0^t \frac{\partial X_{il}(s)}{\partial a_{il}} e^{\theta_i} \sum_{k=1}^M \beta_k \phi_k'(X_{il}(s))ds &(26)\\
X_{il}^{\theta_i}(t) := \frac{\partial X_{il}(t)}{\partial \theta_i} &= \int_0^t \left[\frac{\partial X_{il}(s)}{\partial \theta_i} e^{\theta_i} \sum_{k=1}^M \beta_k \phi_k'(X_{il}(s)) + e^{\theta_i} \sum_{k=1}^M \beta_k \phi_k(X_{il}(s))\right] ds &(27)\\
X_{il}^{\beta_r}(t) := \frac{\partial X_{il}(t)}{\partial \beta_r} &= \int_0^t \left[\frac{\partial X_{il}(s)}{\partial \beta_r} e^{\theta_i} \sum_{k=1}^M \beta_k \phi_k'(X_{il}(s)) + e^{\theta_i} \phi_r(X_{il}(s))\right] ds, &(28)
\end{aligned}$$

for $i = 1, \cdots, n; l = 1, \cdots, N_i; r = 1, \cdots, M$. In another word, these functions satisfy the differential equations:

$$\begin{aligned}
\frac{d}{dt} X_{il}^{a_{il}}(t) &= X_{il}^{a_{il}}(t) e^{\theta_i} \sum_{k=1}^M \beta_k \phi_k'(X_i(t)), \quad X_{il}^{a_{il}}(0) = 1, &(29)\\
\frac{d}{dt} X_{il}^{\theta_i}(t) &= X_{il}^{\theta_i}(t) e^{\theta_i} \sum_{k=1}^M \beta_k \phi_k'(X_{il}(t)) + e^{\theta_i} \sum_{k=1}^M \beta_k \phi_k(X_{il}(t)), \quad X_{il}^{\theta_i}(0) = 0, &(30)\\
\frac{d}{dt} X_{il}^{\beta_r}(t) &= X_{il}^{\beta_r}(t) e^{\theta_i} \sum_{k=1}^M \beta_k \phi_k'(X_{il}(t)) + e^{\theta_i} \phi_r(X_{il}(t)), \quad X_{il}^{\beta_r}(0) = 0. &(31)
\end{aligned}$$

Using similar arguments, it follows that the Hessian of $X_{il}(\cdot)$ with respect to $\boldsymbol{\beta}$, given by the matrix $(X_{il}^{\beta_r,\beta_{r'}})_{r,r'=1}^{M}$, where $X_{il}^{\beta_r,\beta_{r'}}(t) := \frac{\partial^2}{\partial\beta_r\partial\beta_{r'}}X_{il}(t)$, satisfies the system of ODEs, for $r, r' = 1, \cdots, M$:

$$\begin{aligned}\frac{d}{dt}X_{il}^{\beta_r,\beta_{r'}}(t) &= e^{\theta_i}\left[X_{il}^{\beta_r,\beta_{r'}}(t)\sum_{k=1}^{M}\beta_k\phi_k'(X_{il}(t)) + X_{il}^{\beta_r}(t)\phi_{r'}'(X_{il}(t)) + X_{il}^{\beta_{r'}}(t)\phi_r'(X_{il}(t))\right.\\ &+ \left. X_{il}^{\beta_r}(t)X_{il}^{\beta_{r'}}(t)\sum_{k=1}^{M}\beta_k\phi_k''(X_{il}(t))\right], \quad X_{il}^{\beta_r,\beta_{r'}}(0) = 0. \end{aligned} \tag{32}$$

The Hessian of $X_{il}(\cdot)$ with respect to θ_i, given by $X_{il}^{\theta_i,\theta_i}$, satisfies the ODE

$$\begin{aligned}\frac{d}{dt}X_{il}^{\theta_i,\theta_i}(t) &= e^{\theta_i}\left[\sum_{k=1}^{M}\beta_k\phi_k(X_{il}(t)) + (X_{il}^{\theta_i,\theta_i}(t) + 2X_{il}^{\theta_i}(t))\sum_{k=1}^{M}\beta_k\phi_k'(X_{il}(t))\right.\\ &+ \left.(X_{il}^{\theta_i}(t))^2\sum_{k=1}^{M}\beta_k\phi_k''(X_{il}(t))\right], \quad X_{il}^{\theta_i,\theta_i}(0) = 0. \end{aligned} \tag{33}$$

The Hessian of $X_{il}(\cdot)$ with respect to a_{il}, given by $X_{il}^{a_{il},a_{il}}$, satisfies the ODE

$$\begin{aligned}\frac{d}{dt}X_{il}^{a_{il},a_{il}}(t) &= e^{\theta_i}\left[X_{il}^{a_{il},a_{il}}(t)\sum_{k=1}^{M}\beta_k\phi_k'(X_{il}(t))\right.\\ &+ \left.(X_{il}^{a_{il}}(t))^2\sum_{k=1}^{M}\beta_k\phi_k''(X_i(t))\right], \quad X_{il}^{a_{il},a_{il}}(0) = 0. \end{aligned} \tag{34}$$

Also, for future reference (even though it is not used in the proposed algorithm), we calculate the mixed partial derivative of $X_{il}(\cdot)$ with respect to θ_i and β_r as $X_{il}^{\theta_i,\beta_r}(t) := \frac{\partial^2 X_{il}(t)}{\partial\theta_i\partial\beta_r}$ which satisfies the ODE

$$\begin{aligned}\frac{d}{dt}X_{il}^{\theta_i,\beta_r}(t) &= X_{il}^{\theta_i,\beta_r}(t)e^{\theta_i}\sum_{k=1}^{M}\beta_k\phi_k'(X_{il}(t)) + e^{\theta_i}\left[X_{il}^{\beta_r}(t)\sum_{k=1}^{M}\beta_k\phi_k'(X_{il}(t)) + \phi_r(X_{il}(t))\right.\\ &+ \left. X_{il}^{\theta_i}(t)\phi_r'(X_{il}(t)) + X_{il}^{\theta_i}(t)X_{il}^{\beta_r}(t)\sum_{k=1}^{M}\beta_k\phi_k''(X_{il}(t))\right], \; X_{il}^{\theta_i,\beta_r}(0) = 0. \end{aligned} \tag{35}$$

Thus, the approach described above shows that as long as we have evaluated (approximated) the function $X_{il}(\cdot)$ at the grid points $\{0 + mh/2 : m = 0, 1, \ldots, 2/h\}$, we shall be able to approximate the gradients $X_{il}^{a_{il}}(\cdot)$, $X_{il}^{\theta_i}(\cdot)$ and $\{X_{il}^{\beta_r}\}_{r=1}^{M}$ at the grid points $\{0 + mh : m = 0, 1, \ldots, 1/h\}$, and the Hessians $X_{il}^{a_{il},a_{il}}$, $X_{il}^{\theta_i,\theta_i}$ and $(X_{il}^{\beta_r,\beta_{r'}})_{r,r'=1}^{M}$at the grid points $\{0 + 2mh : m = 0, 1, \ldots, 1/(2h)\}$, by successively applying the 4^{th} order Runge-Kutta method.

Expression when g is positive

Note that (29), (30) and (31) are linear differential equations. For the *growth model* we have g positive and the initial conditions a_{il} also can be taken to be positive. If the function $g_{\boldsymbol{\beta}} := \sum_{k=1}^{M}\beta_k\phi_k$ is also positive on the domain of $\{a_{il}\}$'s, then the trajectories $X_{il}(t)$ are nondecreasing in t (in fact strictly increasing if $g_{\boldsymbol{\beta}}$ is strictly positive). In this case, and more generally, whenever

the solutions exist on a time interval $[0,1]$ and $g_{\boldsymbol{\beta}}$ is twice continuously differentiable (so that the solution paths for $X_{il}^{a_{il}}$, $X_{il}^{\theta_{il}}$, $X_{il}^{\beta_r}$, $X_{il}^{a_{il},a_{il}}$, $X_{il}^{\theta_i,\theta_i}$ and $X_{il}^{\beta_r,\beta_{r'}}$ are C^1 functions on $[0,1]$) the gradients of the trajectories can be solved explicitly:

$$
\begin{aligned}
X_{il}^{a_{il}}(t) &= \frac{g_{\boldsymbol{\beta}}(X_{il}(t))}{g_{\boldsymbol{\beta}}(X_{il}(0))}; && (36)\\
X_{il}^{\theta_i}(t) &= e^{\theta_i} t g_{\boldsymbol{\beta}}(X_{il}(t)); && (37)\\
X_{il}^{\beta_r}(t) &= g_{\boldsymbol{\beta}}(X_{il}(t)) \int_{X_{il}(0)}^{X_{il}(t)} \frac{\phi_r(x)}{(g_{\boldsymbol{\beta}}(x))^2} dx. && (38)
\end{aligned}
$$

In the following, we verify equation(38). The proofs for others are similar and thus omitted. We can express

$$
\begin{aligned}
X_{il}^{\beta_r}(t) &= e^{\theta_i} \int_0^t \phi_r(X_{il}(s)) \exp\left(e^{\theta_i} \int_s^t g'_{\boldsymbol{\beta}}(X_{il}(u)) du \right) ds \\
&= e^{\theta_i} \int_0^t \phi_r(X_{il}(s)) \exp\left(\int_s^t \frac{g'_{\boldsymbol{\beta}}(X_{il}(u))}{g_{\boldsymbol{\beta}}(X_{il}(u))} X'_{il}(u) du \right) ds \quad (\text{using } X'_{il}(u) = e^{\theta_i} g_{\boldsymbol{\beta}}(X_{il}(u)) \ ) \\
&= e^{\theta_i} \int_0^t \phi_r(X_{il}(s)) \exp(\log g_{\boldsymbol{\beta}}(X_{il}(t)) - \log g_{\boldsymbol{\beta}}(X_{il}(s))) ds \\
&= g_{\boldsymbol{\beta}}(X_{il}(t)) \int_0^t \frac{\phi_r(X_{il}(s))}{(g_{\boldsymbol{\beta}}(X_{il}(s)))^2} X'_{il}(s) ds \\
&= g_{\boldsymbol{\beta}}(X_{il}(t)) \int_{X_{il}(0)}^{X_{il}(t)} \frac{\phi_r(x)}{(g_{\boldsymbol{\beta}}(x))^2} dx.
\end{aligned}
$$

Using analogous calculations, we can obtain the Hessians in closed form as well. Thus, solutions to (34), (33) and (32) become

$$
\begin{aligned}
X_{il}^{a_{il},a_{il}}(t) &= \frac{g_{\boldsymbol{\beta}}(X_{il}(t))}{(g_{\boldsymbol{\beta}}(X_{il}(0)))^2} [g'_{\boldsymbol{\beta}}(X_{il}(t)) - g'_{\boldsymbol{\beta}}(X_{il}(0))]; && (39)\\
X_{il}^{\theta_i,\theta_i}(t) &= e^{\theta_i} g_{\boldsymbol{\beta}}(X_{il}(t)) [t + e^{\theta_i} t^2 g'_{\boldsymbol{\beta}}(X_{il}(t))]; && (40)
\end{aligned}
$$

$$
\begin{aligned}
& X_{il}^{\beta_r,\beta_{r'}}(t) \\
= \ & g_{\boldsymbol{\beta}}(X_{il}(t)) \int_{X_{il}(0)}^{X_{il}(t)} \frac{1}{g_{\boldsymbol{\beta}}(x)} \left[\phi'_r(x)(F_{r'}(x) - F_{r'}(X_{il}(0))) + \phi'_{r'}(x)(F_r(x) - F_r(X_{il}(0))) \right] dx \\
& + (F_r(X_{il}(t)) - F_r(X_{il}(0)))(F_{r'}(X_{il}(t)) - F_{r'}(X_{il}(0))) g_{\boldsymbol{\beta}}(X_{il}(t)) g'_{\boldsymbol{\beta}}(X_{il}(t)) \\
& - e^{-\theta_i} g_{\boldsymbol{\beta}}(X_{il}(t)) \int_{X_{il}(0)}^{X_{il}(t)} \frac{g'_{\boldsymbol{\beta}}(x)}{(g_{\boldsymbol{\beta}}(x))^3} \left[\phi_r(x)(F_{r'}(x) - F_{r'}(X_{il}(0))) + \phi_{r'}(x)(F_r(x) - F_r(X_{il}(0))) \right] dx,
\end{aligned} \tag{41}
$$

where, for $x_1 < x_2$,

$$
F_r(x_2) - F_r(x_1) = \int_{x_1}^{x_2} \frac{\phi_r(y)}{(g_{\boldsymbol{\beta}}(y))^2} dy, \quad 1 \le r \le M.
$$

We can express $X_{il}^{\beta_r,\beta_{r'}}(t)$ alternatively as

$$\begin{aligned} X_{il}^{\beta_r,\beta_{r'}}(t) &= e^{\theta_i} g_{\boldsymbol{\beta}}(X_{il}(t)) \int_0^t \frac{1}{g_{\boldsymbol{\beta}}(X_{il}(s))} \left[X_{il}^{\beta_r}(s)\phi'_{r'}(X_{il}(s)) + \phi'_r(X_{il}(s)) X_{il}^{\beta_{r'}}(s) \right] dt \\ &\quad + e^{\theta_i} g_{\boldsymbol{\beta}}(X_{il}(t)) \int_0^t \frac{1}{g_{\boldsymbol{\beta}}(X_{il}(s))} X_{il}^{\beta_r}(s) X_{il}^{\beta_{r'}}(s) g''_{\boldsymbol{\beta}}(X_{il}(s)) ds. \end{aligned} \tag{42}$$

Similarly, we have the representation

$$\begin{aligned} X_{il}^{\theta_i,\beta_r}(t) &= e^{\theta_i} g_{\boldsymbol{\beta}}(X_{il}(t)) \int_0^t \frac{1}{g_{\boldsymbol{\beta}}(X_{il}(s))} X_{il}^{\theta_i}(s) \phi'_r(X_{il}(s)) ds \\ &\quad + e^{\theta_i} g_{\boldsymbol{\beta}}(X_{il}(t)) \int_0^t \frac{1}{g_{\boldsymbol{\beta}}(X_{il}(s))} \left[X_{il}^{\beta_r}(s) g'_{\boldsymbol{\beta}}(X_{il}(s)) + \phi_r(X_{il}(s)) + X_{il}^{\theta_i}(s) X_{il}^{\beta_r}(s) g''_{\boldsymbol{\beta}}(X_{il}(s)) \right] ds. \end{aligned} \tag{43}$$

Appendix B : Levenberg-Marquardt method

The Levenberg-Marquardt method is a method for solving the nonlinear least squares problem:

$$\min_{\boldsymbol{\gamma}} S(\boldsymbol{\gamma}) \quad \text{where} \quad S(\boldsymbol{\gamma}) = \sum_{i=1}^{n} [y_i - f_i(\boldsymbol{\gamma})]^2,$$

where $f_i(\boldsymbol{\gamma})$'s are nonlinear functions of the parameter $\boldsymbol{\gamma} \in \mathbb{R}^p$. The key idea is to linearly approximate $f_i(\boldsymbol{\gamma} + \boldsymbol{\delta}) \approx f_i(\boldsymbol{\gamma}) + J_i^T \boldsymbol{\delta}$, for a small $\boldsymbol{\delta} \in \mathbb{R}^p$, where J_i is the Jacobian of f_i at $\boldsymbol{\gamma}$. Denote

$$\mathbf{y} = (y_1, \ldots, y_n)^T, \quad \mathbf{f}(\boldsymbol{\gamma}) = (f_1(\boldsymbol{\gamma}), \ldots, f_n(\boldsymbol{\gamma}))^T,$$

and $\mathbf{J}$ to be the $n \times p$ matrix with rows $J_1^T, \ldots, J_n^T$. The resulting linearized least squares problem involves, for given $\boldsymbol{\gamma}$ solving for $\boldsymbol{\delta}$ the equation

$$(\mathbf{J}^T\mathbf{J} + \lambda \, \mathrm{diag}(\mathbf{J}^T\mathbf{J}))\boldsymbol{\delta} = \mathbf{J}^T(\mathbf{y} - \mathbf{f}(\boldsymbol{\gamma})), \tag{44}$$

for a regularization parameter $\lambda > 0$. Note that, this solution bears similarity with the ridge regression estimate. However, the formulation in (44) is according to the observation by Marquardt that if each component of the gradient is scaled according to the curvature then there is a larger movement in the directions where the gradient is smaller. In practice, the regularization parameter λ is chosen adaptively to facilitate convergence.

Appendix C : Newton-Raphson procedure

We briefly describe the key steps of the Newton-Raphson procedure for optimizing the objective function (6). As in the implementation of the Levenberg-Marquardt algorithm, we break the iterative procedure in three steps. The update of $\boldsymbol{a}$ is still performed by the Levenberg-Marquardt algorithm (9), while keeping $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$ fixed at the current values. However, we employ Newton-Raphson to update $\boldsymbol{\theta}$ and $\boldsymbol{\beta}$. Fixing $\boldsymbol{a}$, $\boldsymbol{\beta}$ at the current estimates $\boldsymbol{a}^*$ and $\boldsymbol{\beta}^*$, respectively, we update θ_i's from the current estimates θ_i^* by

$$\theta_i^{new} = \theta_i^* - \left[\sum_{l=1}^{N_i} \sum_{j=1}^{m_{il}} \frac{\partial^2 \ell_{ilj}}{\partial \theta_i^2} \right]^{-1} \sum_{l=1}^{N_i} \sum_{j=1}^{m_{il}} \frac{\partial \ell_{ilj}}{\partial \theta_i}, \tag{45}$$

where the quantities on the right hand side are all evaluated at $(\boldsymbol{a}^*, \boldsymbol{\theta}^*, \boldsymbol{\beta}^*)$, and

$$\frac{\partial \ell_{ilj}}{\partial \theta_i} = -2\widetilde{\varepsilon}_{ilj}\frac{\partial}{\partial \theta_i}\widetilde{X}_{il}(t_{ilj}) + 2\frac{\lambda_2\theta_i}{\sum_{l=1}^{N_i} m_{il}}$$

$$\frac{\partial^2 \ell_{ilj}}{\partial \theta_i^2} = -2\widetilde{\varepsilon}_{ilj}\frac{\partial^2}{\partial^2 \theta_i}\widetilde{X}_{il}(t_{ijl}) + 2(\frac{\partial}{\partial \theta_i}\widetilde{X}_{il}(t_{ijl}))^2 + 2\frac{\lambda_2}{\sum_{l=1}^{N_i} m_{il}}.$$

Similarly, the Newton-Raphson update for $\boldsymbol{\beta}$ is given by

$$\boldsymbol{\beta}^{new} = \boldsymbol{\beta}^* - \left[\sum_{i=1}^{n}\sum_{l=1}^{N_i}\sum_{j=1}^{m_{il}} \frac{\partial^2 \ell_{ilj}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} + 2\mathbf{B}\right]^{-1} \left(\sum_{i=1}^{n}\sum_{l=1}^{N_i}\sum_{j=1}^{m_{il}} \frac{\partial \ell_{ilj}}{\partial \boldsymbol{\beta}} + 2\mathbf{B}\boldsymbol{\beta}^*\right), \tag{46}$$

where the quantities on the right hand side are again evaluated at $(\boldsymbol{a}^*, \boldsymbol{\theta}^*, \boldsymbol{\beta}^*)$, and

$$\begin{aligned}
\frac{\partial \ell_{ilj}}{\partial \boldsymbol{\beta}} &= -2\widetilde{\varepsilon}_{ilj}\frac{\partial}{\partial \boldsymbol{\beta}}\widetilde{X}_{il}(t_{ilj}) \\
\frac{\partial^2 \ell_{ilj}}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} &= 2\frac{\partial}{\partial \boldsymbol{\beta}}\widetilde{X}_{il}(t_{ilj})\left(\frac{\partial}{\partial \boldsymbol{\beta}}\widetilde{X}_{il}(t_{ilj})\right)^T - 2\widetilde{\varepsilon}_{ilj}\frac{\partial^2}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T}\widetilde{X}_{il}(t_{ilj}).
\end{aligned}$$

Appendix D : Cubic B-spline fits to the plant data

We first consider the control group. In Table 4, we report the results using B-spline basis with knots at $1 + 11.5(1:M)/M$ for $M = 2, 3, \cdots, 12$; and using $\sigma_\varepsilon^{ini} = 0.05$, and $\sigma_\theta^{ini} = 1$ as initial estimates. In the B-spline fitting, we set the penalty matrix $\boldsymbol{B}$ to be the zero matrix, that is $\lambda_R = 0$. In the Newton-Raphson step, both λ_1 and λ_2 are estimated adaptively from the data. However, the Levenberg-Marquardt step is non-adaptive, that is it uses the initial values of λ_1 and λ_2 throughout. From Table 4, for M= 2 to 8 there is no convergence. For $M = 9$ to 12, the approximate CV scores are quite similar and the minimum is achieved at $M = 9$.

We then consider the fits for the treatment group. The results using B-splines with knots at $1 + 9.5(1:M)/M$ for $M = 2, 3, \cdots, 12$; and using $\sigma_\varepsilon^{ini} = 0.05$, and $\sigma_\theta^{ini} = 1$ are reported in Table 5. We again set $\lambda_R = 0$ (that is no penalty). As for $M = 2$ to 6, there is no convergence. For M=7 to 10, the CV scores are similar and the minimum is achieved again at M=9. For $M = 11$ and 12, the method breaks down due to numerical instability.

Appendix E : Proof details

In this section we provide the proofs of the key asymptotic results.

Proof of Theorem 1

For convenience, we introduce the following notations:

$$\ell_{i\cdot}(\theta_i, \boldsymbol{\beta}) := \sum_{l=1}^{N_i}\sum_{j=1}^{m_i} \ell_{ilj}(a_{il}, \theta_i, \boldsymbol{\beta}) \qquad \text{and} \qquad \ell_{\cdot\cdot}(\boldsymbol{\theta}, \boldsymbol{\beta}) := \sum_{i=1}^{n} \ell_{i\cdot}(\theta_i, \boldsymbol{\beta}).$$

Table 4: Approximate leave-one-curve-out CV scores for *control group*. Cubic B-spline basis with knot sequence $1 + 11.5(1 : M)/M$; and $\sigma_\varepsilon^{ini} = 0.05$, $\sigma_\theta^{ini} = 1$. (* = no convergence)

M	# L-M	# N-R	CV score
2*	216	1000	489.01162
3*	445	1000	416.69848
4*	458	1000	91.21308
5*	546	1000	74.12581
6*	337	1000	58.25487
7*	279	1000	53.69243
8*	190	1000	53.37721
9	233	195	**53.16987**
10	147	120	53.26008
11	94	79	53.26125
12	78	54	53.41077

Table 5: Approximate leave-one-curve-out CV scores for *treatment group*. Cubic B-spline basis with knot sequence $1 + 9.5(1 : M)/M$; and $\sigma_\varepsilon^{ini} = 0.05$, $\sigma_\theta^{ini} = 1$. (* = no convergence)

M	# L-M	# N-R	CV score
2*	228	1000	348.65867
3*	426	1000	422.03137
4*	233	1000	96.66250
5*	257	1000	71.77904
6*	539	1000	65.85252
7	336	277	64.25370
8	197	143	63.91828
9	125	83	**63.83346**
10	94	38	63.90003
11*	–	–	–
12*	–	–	–

Here, $\boldsymbol{\theta} := (\theta_2, \ldots, \theta_n)^T$ since $\theta_1 \equiv 0$. For $\alpha > 0$, define

$$\Omega(\alpha) := \{(\boldsymbol{\theta}, \boldsymbol{\beta}) : \boldsymbol{\theta} = \boldsymbol{\theta}^* + \alpha\boldsymbol{\eta}, \boldsymbol{\beta} = \boldsymbol{\beta}^* + \alpha\boldsymbol{\delta},\ \boldsymbol{\eta} \in \mathbb{R}^{n-1}, \boldsymbol{\delta} \in \mathbb{R}^M,\ \text{s.t.}\ \parallel \boldsymbol{\eta} \parallel^2 + \parallel \boldsymbol{\delta} \parallel^2 = 1\}. \quad (47)$$

We use X^g_{ilj} to denote $X_{il}(T_{i,j}; a_{il}, g)$ where $X_{il}(\cdot)$ is the solution of the equation $x'(t) = e^{\theta^*_i} g(x(t))$ with $x(0) = a_{il}$. We use $X_{il}(\cdot; \boldsymbol{\theta}, \boldsymbol{\beta})$ to denote the solution of (17) when $X_{il}(0) = a_{il}$, and $X_{ilj}(\boldsymbol{\theta}, \boldsymbol{\beta}) := X_{il}(T_{i,j}; \boldsymbol{\theta}, \boldsymbol{\beta})$. We define $X^{\theta_i}_{il}(\cdot; \boldsymbol{\theta}, \boldsymbol{\beta})$, $X^{\beta_r}_{il}(\cdot; \boldsymbol{\theta}, \boldsymbol{\beta})$, $X^{\theta_i, \theta_i}_{il}(\cdot; \boldsymbol{\theta}, \boldsymbol{\beta})$, $X^{\theta_i, \beta_r}_{il}(\cdot; \boldsymbol{\theta}, \boldsymbol{\beta})$ and $X^{\beta_r, \beta_{r'}}_{il}(\cdot; \boldsymbol{\theta}, \boldsymbol{\beta})$ as the partial derivatives and mixed partial derivatives of $X_{il}(\cdot; \boldsymbol{\theta}, \boldsymbol{\beta})$ with respect to θ_i, β_r, (θ_i, θ_i), (θ_i, β_r) and $(\beta_r, \beta_{r'})$, respectively. Notations such as $X^{\theta_i}_{ilj}(\boldsymbol{\theta}, \boldsymbol{\beta})$ are used to mean $X^{\theta_i}_{il}(T_{i,j}; \boldsymbol{\theta}, \boldsymbol{\beta})$. We use $g_{\boldsymbol{\beta}}$ to denote the function $\sum_{k=1}^M \beta_k \phi_k$ (for convenience henceforth dropping the subscript M from $\phi_{k,M}$) and denote its first and second derivatives by $g'_{\boldsymbol{\beta}}$ and $g''_{\boldsymbol{\beta}}$, respectively. Finally, we use $\| \cdot \|_\infty$ to mean $\| \cdot \|_{L^\infty(D)}$, and denote the operator norm of a matrix and l_2 norm of a vector by $\| \cdot \|$. We use $\boldsymbol{T}$ to denote $\{T_{i,j} : j = 1, \ldots, m_i; i = 1, \ldots, n\}$ and $\boldsymbol{\varepsilon}$ to denote $\{\varepsilon_{ilj} : j = 1, \ldots, m_i; l = 1, \ldots, N_i; i = 1, \ldots, n\}$.

Let $\boldsymbol{\eta} \in \mathbb{R}^{n-1}$ and $\boldsymbol{\delta} \in \mathbb{R}^M$ be arbitrary vectors satisfying $\parallel \boldsymbol{\eta} \parallel^2 + \parallel \boldsymbol{\delta} \parallel^2 = 1$. Define $J_{n-1} := I_{n-1} - \frac{1}{n-1}\mathbf{1}_{n-1}\mathbf{1}^T_{n-1}$ and observe that $\sum_{i=2}^n (\theta_i - \overline{\theta})^2 = \boldsymbol{\theta}^T J_{n-1} \boldsymbol{\theta}$. Define

$$W_{\boldsymbol{\beta}} := \sum_{i=1}^n \sum_{l=1}^{N_i} \sum_{j=1}^{m_i} (Y_{ilj} - X_{il}(T_{i,j}; a_{il}, \theta^*_i, \boldsymbol{\beta}^*)) \frac{\partial X_{il}}{\partial \boldsymbol{\beta}}(T_{i,j}; a_{il}, \theta^*_i, \boldsymbol{\beta}^*)$$

and $W_{\boldsymbol{\theta}}$ to be an $(n-1) \times 1$ vector with $(i-1)$-th coordinate

$$\sum_{l=1}^{N_i} \sum_{j=1}^{m_i} (Y_{ilj} - X_{il}(T_{i,j}; a_{il}, \theta^*_i, \boldsymbol{\beta}^*)) \frac{\partial X_{il}}{\partial \theta_i}(T_{i,j}; a_{il}, \theta^*_i, \boldsymbol{\beta}^*)$$

for $i = 2, \ldots, n$. Also let $W := (W^T_{\boldsymbol{\beta}}, W^T_{\boldsymbol{\theta}})^T$. Then by a second order Taylor expansion, we have,

$$\begin{aligned}
& \ell_{\cdot\cdot}(\boldsymbol{\theta}^* + \alpha_N \boldsymbol{\eta}, \boldsymbol{\beta}^* + \alpha_N \boldsymbol{\delta}) - \ell_{\cdot\cdot}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*) \\
= \ & \lambda_2 \alpha_N (2(\boldsymbol{\theta}^*)^T J_{n-1} \boldsymbol{\eta} + \alpha_N \boldsymbol{\eta}^T J_{n-1} \boldsymbol{\eta}) + 2\alpha_N [\boldsymbol{\delta}^T, \boldsymbol{\eta}^T] \begin{bmatrix} W_{\boldsymbol{\beta}} \\ W_{\boldsymbol{\theta}} \end{bmatrix} \\
& + \alpha^2_N [\boldsymbol{\delta}^T, \boldsymbol{\eta}^T] \begin{bmatrix} \mathcal{G}_{\beta\beta}(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\beta}}) & \mathcal{G}_{\beta\theta}(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\beta}}) \\ \mathcal{G}_{\theta\beta}(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\beta}}) & \mathcal{G}_{\theta\theta}(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\beta}}) \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta} \\ \boldsymbol{\eta} \end{bmatrix} \\
& - \alpha^2_N [\boldsymbol{\delta}^T, \boldsymbol{\eta}^T] \begin{bmatrix} \mathcal{H}_{\beta\beta}(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\beta}}) & \mathcal{H}_{\beta\theta}(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\beta}}) \\ \mathcal{H}_{\theta\beta}(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\beta}}) & \mathcal{H}_{\theta\theta}(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\beta}}) \end{bmatrix} \begin{bmatrix} \boldsymbol{\delta} \\ \boldsymbol{\eta} \end{bmatrix}, \qquad (48)
\end{aligned}$$

where $(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\beta}})$ satisfies $\parallel \overline{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \parallel \leq \alpha_N$ and $\parallel \overline{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \parallel \leq \alpha_N$. Note that $(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\beta}})$ depends on $(\boldsymbol{a}, \boldsymbol{T})$ and $(\boldsymbol{\eta}, \boldsymbol{\delta})$, but not on $\boldsymbol{\varepsilon}$. In the above, $\mathcal{G}_{\beta\beta}(\boldsymbol{\theta}, \boldsymbol{\beta})$ is the $M \times M$ matrix

$$\sum_{i=1}^n \sum_{l=1}^{N_i} \sum_{j=1}^{m_i} \left(\frac{\partial X_{il}}{\partial \boldsymbol{\beta}}(T_{i,j}; a_{il}, \theta_i, \boldsymbol{\beta}) \right) \left(\frac{\partial X_{il}}{\partial \boldsymbol{\beta}}(T_{i,j}; a_{il}, \theta_i, \boldsymbol{\beta}) \right)^T;$$

$\mathcal{G}_{\theta\beta}(\boldsymbol{\theta}, \boldsymbol{\beta})$ is the $(n-1) \times M$ matrix with $(i-1)$-th row

$$\sum_{l=1}^{N_i} \sum_{j=1}^{m_i} \frac{\partial X_{il}}{\partial \theta_i}(T_{i,j}; a_{il}, \theta_i, \boldsymbol{\beta}) \left(\frac{\partial X_{il}}{\partial \boldsymbol{\beta}}(T_{i,j}; a_{il}, \theta_i, \boldsymbol{\beta}) \right)^T, \quad i = 2, \ldots, n;$$

$\mathcal{G}_{\theta\theta}(\boldsymbol{\theta},\boldsymbol{\beta})$ is the $(n-1)\times(n-1)$ diagonal matrix with the $(i-1)$-th diagonal entry

$$\sum_{l=1}^{N_i}\sum_{j=1}^{m_i}\left(\frac{\partial X_{il}}{\partial\theta_i}(T_{i,j};a_{il},\theta_i,\boldsymbol{\beta})\right)^2, \qquad i=2,\ldots,n;$$

$\mathcal{G}_{\beta\theta}(\boldsymbol{\theta},\boldsymbol{\beta})=\mathcal{G}_{\theta\beta}(\boldsymbol{\theta},\boldsymbol{\beta})^T$; $\mathcal{H}_{\beta\beta}(\boldsymbol{\theta},\boldsymbol{\beta})$ is the $M\times M$ matrix

$$\sum_{i=1}^{n}\sum_{l=1}^{N_i}\sum_{j=1}^{m_i}(Y_{ilj}-X_{il}(T_{i,j};a_{il},\theta_i,\boldsymbol{\beta}))\frac{\partial^2 X_{il}}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T}(T_{i,j};a_{il},\theta_i,\boldsymbol{\beta});$$

$\mathcal{H}_{\theta\beta}(\boldsymbol{\theta},\boldsymbol{\beta})$ is the $(n-1)\times M$ matrix with $(i-1)$-th row

$$\sum_{l=1}^{N_i}\sum_{j=1}^{m_i}(Y_{ilj}-X_{il}(T_{i,j};a_{il},\theta_i,\boldsymbol{\beta}))\frac{\partial^2 X_{il}}{\partial\theta_i\partial\boldsymbol{\beta}^T}(T_{i,j};a_{il},\theta_i,\boldsymbol{\beta}), \qquad i=2,\ldots,n;$$

$\mathcal{H}_{\theta\theta}(\boldsymbol{\theta},\boldsymbol{\beta})$ is the $(n-1)\times(n-1)$ matrix with $(i-1)$-th diagonal entry

$$\sum_{l=1}^{N_i}\sum_{j=1}^{m_i}(Y_{ilj}-X_{il}(T_{i,j};a_{il},\theta_i,\boldsymbol{\beta}))\frac{\partial^2 X_{il}}{\partial\theta_i^2}(T_{i,j};a_{il},\theta_i,\boldsymbol{\beta}), \qquad i=2,\ldots,n;$$

and $\mathcal{H}_{\beta\theta}(\boldsymbol{\theta},\boldsymbol{\beta})=\mathcal{H}_{\theta\beta}(\boldsymbol{\theta},\boldsymbol{\beta})^T$. Let $\mathcal{G}_{*,\theta\theta}$, $\mathcal{G}_{*,\beta\theta}$, $\mathcal{G}_{*,\theta\beta}$ and $\mathcal{G}_{*,\beta\beta}$ denote the expectations of $\mathcal{G}_{\theta\theta}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*)$, $\mathcal{G}_{\beta\theta}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*)$, $\mathcal{G}_{\theta\beta}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*)$ and $\mathcal{G}_{\beta\beta}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*)$ with respect to $(\boldsymbol{a},\boldsymbol{T})$. For future reference, we define the $(M+n-1)\times(M+n-1)$ symmetric matrix $\mathcal{G}(\boldsymbol{\theta},\boldsymbol{\beta})$ as

$$\mathcal{G}(\boldsymbol{\theta},\boldsymbol{\beta})=\begin{bmatrix}\mathcal{G}_{\beta\beta}(\boldsymbol{\theta},\boldsymbol{\beta}) & \mathcal{G}_{\beta\theta}(\boldsymbol{\theta},\boldsymbol{\beta})\\ \mathcal{G}_{\theta\beta}(\boldsymbol{\theta},\boldsymbol{\beta}) & \mathcal{G}_{\theta\theta}(\boldsymbol{\theta},\boldsymbol{\beta})\end{bmatrix}.$$

We define $\mathcal{H}(\boldsymbol{\theta},\boldsymbol{\beta})$ and $\mathcal{G}_*$ analogously.

The following decomposition of the residuals is used throughout:

$$Y_{ilj}-X_{ilj}(\boldsymbol{\theta},\boldsymbol{\beta})=\varepsilon_{ilj}+(X^g_{ilj}-X_{ilj}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*))+(X_{ilj}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*)-X_{ilj}(\boldsymbol{\theta},\boldsymbol{\beta})). \tag{49}$$

Without loss of generality in the following we assume that $\alpha_N M^{3/2}\to 0$, so that in particular the bounds (61) - (69) are valid. The proof of Theorem 1 then follows from the following sequence of lemmas.

Lemma A.1 : *Let $\boldsymbol{\gamma}=(\boldsymbol{\delta}^T,\boldsymbol{\eta}^T)^T$, and W be as defined earlier. Then, with probability tending to 1, uniformly in $\boldsymbol{\gamma}$ such that $\|\boldsymbol{\gamma}\|=1$, we have*

$$|\boldsymbol{\gamma}^T W| \;=\; \left[O(\sigma_\varepsilon M^{1/2}\sqrt{\log(\overline{N}\overline{m})})+O(M^{-p}(\overline{N}\overline{m})^{1/2})\right]\sqrt{\boldsymbol{\gamma}^T\mathcal{G}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*)\boldsymbol{\gamma}}. \tag{50}$$

Lemma A.2 : *With $\boldsymbol{\gamma}$ as in Lemma A.1, uniformly over $\boldsymbol{\gamma}$, we have*

$$\begin{aligned}&\boldsymbol{\gamma}^T\begin{bmatrix}\mathcal{G}_{\beta\beta}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}})-\mathcal{G}_{\beta\beta}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*) & \mathcal{G}_{\beta\theta}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}})-\mathcal{G}_{\beta\theta}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*)\\ \mathcal{G}_{\theta\beta}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}})-\mathcal{G}_{\theta\beta}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*) & \mathcal{G}_{\theta\theta}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}})-\mathcal{G}_{\theta\theta}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*)\end{bmatrix}\boldsymbol{\gamma}\\ =\;& O(\alpha_N M^{3/2}(\overline{N}\overline{m})^{1/2})\sqrt{\boldsymbol{\gamma}^T\mathcal{G}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*)\boldsymbol{\gamma}}+O(\alpha_N^2 M^3\overline{N}\overline{m}).\end{aligned} \tag{51}$$

Lemma A.3 : *There exists a constant $c_7 > 0$ such that, with $\boldsymbol{\gamma}$ as in Lemma A.1, uniformly over $\boldsymbol{\gamma}$, we have*

$$\boldsymbol{\gamma}^T \mathcal{G}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*)\boldsymbol{\gamma} \geq \boldsymbol{\gamma}^T \mathcal{G}_* \boldsymbol{\gamma}(1 - o_P(1)) \geq c_7 \kappa_M^{-1} \overline{N}\overline{m}(1 - o_P(1)). \tag{52}$$

Lemma A.4 : *With $\boldsymbol{\gamma}$ as in Lemma A.1, with probability tending to 1, uniformly over $\boldsymbol{\gamma}$,*

$$\begin{aligned} & \boldsymbol{\gamma}^T \begin{bmatrix} \mathcal{H}_{\beta\beta}(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\beta}}) - \mathcal{H}_{\beta\beta}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*) & \mathcal{H}_{\beta\theta}(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\beta}}) - \mathcal{H}_{\beta\theta}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*) \\ \mathcal{H}_{\theta\beta}(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\beta}}) - \mathcal{H}_{\theta\beta}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*) & \mathcal{H}_{\theta\theta}(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\beta}}) - \mathcal{H}_{\theta\theta}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*) \end{bmatrix} \boldsymbol{\gamma} \\ = \; & O(\alpha_N M^{3/2} (\overline{N}\overline{m})^{1/2}) \sqrt{\boldsymbol{\gamma}^T \mathcal{G}_* \boldsymbol{\gamma}} + O(\alpha_N M^{1/2} \overline{N}\overline{m}) + O(\alpha_N^2 M^3 \overline{N}\overline{m}) \\ & + O(\alpha_N M^{5/2-p} \overline{N}\overline{m}) + O(\sigma_\varepsilon \alpha_N M^3 (\overline{N}\overline{m})^{1/2} \sqrt{\log(\overline{N}\overline{m})}). \end{aligned} \tag{53}$$

Finally, using (49), (60), and (65)-(69), we have

$$\begin{aligned} & \max\{\| \mathcal{H}_{\beta\beta}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*) \|, \| \mathcal{H}_{\beta\theta}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*) \|, \| \mathcal{H}_{\theta\theta}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*) \|\} \\ = \; & O_P(\sigma_\varepsilon M (\overline{N}\overline{m})^{1/2}) + O(M^{-(p-1)} \overline{N}\overline{m}). \end{aligned} \tag{54}$$

Combining (51) - (54), from (48), with probability tending to 1, uniformly in $\boldsymbol{\gamma}$,

$$\begin{aligned} & \ell_{\cdot\cdot}(\boldsymbol{\theta}^* + \alpha_N \boldsymbol{\eta}, \boldsymbol{\beta}^* + \alpha_N \boldsymbol{\delta}) - \ell_{\cdot\cdot}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*) \\ \geq \; & \alpha_N^2 \boldsymbol{\gamma}^T \mathcal{G}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*) \boldsymbol{\gamma} \\ & - \alpha_N \left(O(\sigma_\varepsilon M^{1/2} \sqrt{\log(\overline{N}\overline{m})}) + O(M^{-p} (\overline{N}\overline{m})^{1/2}) + O(\alpha_N^2 M^{3/2} (\overline{N}\overline{m})^{1/2}) \right) \sqrt{\boldsymbol{\gamma}^T \mathcal{G}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*) \boldsymbol{\gamma}} \\ & - \alpha_N \lambda_2 \| \boldsymbol{\theta}^* \| O(1) - \alpha_N^2 O(\alpha_N M^{3/2} (\overline{N}\overline{m})^{1/2}) \sqrt{\boldsymbol{\gamma}^T \mathcal{G}_* \boldsymbol{\gamma}} \\ & - \alpha_N^2 O((\alpha_N M^{1/2} + \alpha_N^2 M^3 + \alpha_N M^{5/2-p} + M^{-(p-1)}) \overline{N}\overline{m}) \\ & - \alpha_N^2 O((\sigma_\varepsilon \alpha_N M^3 + \sigma_\varepsilon M) (\overline{N}\overline{m})^{1/2} \sqrt{\log(\overline{N}\overline{m})}) \\ \geq \; & c_4 \kappa_M^{-1} \alpha_N^2 \overline{N}\overline{m}(1 - o_P(1)) \end{aligned} \tag{55}$$

where $c_4 > 0$ is some constant. The last step uses Lemma A.3 and the following fact:

[**Q**] For any positive definite matrix A, with $\| A^{-1} \| \leq \kappa$, if $2c\sqrt{\kappa} < 1$, then for all $\mathbf{x}$ such that $\| \mathbf{x} \| = 1$

$$\mathbf{x}^T A \mathbf{x} - c\sqrt{\mathbf{x}^T A \mathbf{x}} \geq \frac{1}{2} \mathbf{x}^T A \mathbf{x}$$

Thus, with probability tending to 1, there is a local minimum $(\widehat{\boldsymbol{\theta}}, \widehat{\boldsymbol{\beta}})$ of the objective function (5) with $\| \widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^* \|^2 + \| \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}^* \|^2 \leq \alpha_N^2$. This completes the proof of Theorem 1.

Proof of Theorem 2

We make use of the following inequality due to Halperin and Pitt (Mitrinovic, Pecaric and Fink, 1991, page 8): *If f is locally absolutely continuous and f'' is in $L_2([0, A])$, then for any $\epsilon > 0$ the following inequality holds*

$$\int_0^A f'^2 \leq K(\epsilon) \int_0^A f^2 + \epsilon \int_0^A f''^2$$

where $K(\epsilon) = 1/\epsilon + 12/A^2$.

Define $X_i(t,x)$ as the sample path $X_{il}(t; a_{il}, \theta_i^*, \boldsymbol{\beta}^*)$ when $a_{il} = x$. Since $\theta_1 = 0$ and $X_{il}^{\boldsymbol{\beta}}(\cdot; \boldsymbol{\theta}, \boldsymbol{\beta})$ is given by (38) (Appendix A), in order to prove Theorem 2, it is enough to find a lower bound on

$$\min_{\|\mathbf{b}\|=1} \int \int_0^1 \left[\int_0^t g_{\mathbf{b}}(X_1(u,x))/g_{\boldsymbol{\beta}^*}(X_1(u,x))du \right]^2 f_T(t)dtdF_a(x)$$

where $g_{\mathbf{b}}(u) = \mathbf{b}^T \boldsymbol{\phi}(u)$. By **A5**, without loss of generality we can take the density $f_T(\cdot)$ to be uniform on $[0,1]$. Let

$$R(t,x) := \int_0^t g_{\mathbf{b}}(X_1(u,x))/g_{\boldsymbol{\beta}^*}(X_1(u,x))du.$$

Then,

$$\begin{aligned} r(t,x) &:= \frac{\partial}{\partial t} R(t,x) = \frac{g_{\mathbf{b}}(X_1(t,x))}{g_{\boldsymbol{\beta}^*}(X_1(t,x))} \\ r'(t,x) &:= \frac{\partial}{\partial t} r(t,x) = \left[\frac{g'_{\mathbf{b}}(X_1(t,x))}{g_{\boldsymbol{\beta}^*}(X_1(t,x))} - \frac{g_{\mathbf{b}}(X_1(t,x)) g'_{\boldsymbol{\beta}^*}(X_1(t,x))}{g^2_{\boldsymbol{\beta}^*}(X_1(t,x))} \right] X'_1(t,x) \\ &= \left[\frac{g'_{\mathbf{b}}(X_1(t,x))}{g_{\boldsymbol{\beta}^*}(X_1(t,x))} - \frac{g_{\mathbf{b}}(X_1(t,x)) g'_{\boldsymbol{\beta}^*}(X_1(t,x))}{g^2_{\boldsymbol{\beta}^*}(X_1(t,x))} \right] g_{\boldsymbol{\beta}^*}(X_1(t,x)) \end{aligned}$$

From this, and the fact that the coordinates of $\boldsymbol{\phi}'(u)$ are of the order $O(M^{3/2})$, coordinates of $\boldsymbol{\phi}(u)$ are of the order $O(M^{1/2})$, and all these functions are supported on intervals of length $O(M^{-1})$, we obtain that, uniformly in x,

$$\int_0^1 (r'(t,x))^2 dt = O(M^2). \tag{56}$$

Application of Halperin-Pitt inequality with $f(x) = \int_0^1 R(t,x)^2 dt$ yields

$$\int \int_0^1 (r(t,x))^2 dtdF_a(x) \leq (1/\epsilon + 12) \int \int_0^1 (R(t,x))^2 dtdF_a(x) + \epsilon \int \int_0^1 (r'(t,x))^2 dtdF_a(x). \tag{57}$$

Take $\epsilon = k_0 M^{-2}$ for some $k_0 > 0$, then by (56),

$$\int \int_0^1 (R(t,x))^2 dtdF_a(x) \quad \geq \quad k_1 M^{-2} \int \int_0^1 (r(t,x))^2 dtdF_a(x) - k_2 M^{-2},$$

for constants $k_1, k_2 > 0$ dependent on k_0. Rewrite $\int \int_0^1 (r(t,x))^2 dtdF_a(x)$ as

$$\int \int_x^{X_1(1,x)} \frac{g^2_{\mathbf{b}}(v)}{g^3_{\boldsymbol{\beta}^*}(v)} dv dF_a(x) = \int g^2_{\mathbf{b}}(v) h(v) dv \tag{58}$$

where $h(v) = g^{-3}_{\boldsymbol{\beta}^*}(v) \int \mathbf{1}_{\{x \leq v \leq X_1(1,x)\}} dF_a(x)$. If the knots are equally spaced on $[x_0 + \delta, x_1 - \delta]$ for some constant $\delta > 0$ is bounded below, then $\inf_{v \in D_0} h(v)$ is bounded below (even as $M \to \infty$) where D_0 is the union of the supports of $\{\phi_{k,M}\}_{k=1}^M$, which contained in $[x_0 + \delta/2, x_1 - \delta/2]$ for M sufficiently large). In this case, $\int \int_x^{X_1(1,x)} \frac{g^2_{\mathbf{b}}(v)}{g^3_{\boldsymbol{\beta}^*}(v)} dv \geq k_3$ for some constant $k_3 > 0$. Thus, by appropriate choice of ϵ, we have $\int \int_0^1 (R(t,x))^2 dtdF_a(x) \geq k_4 M^{-2}$ for some $k_4 > 0$, which yields $\kappa_M = O(M^2)$.

Proof of Proposition 1

The proof is based on the following lemmas.

Lemma A.5: *Let $\mathcal{P}_d$ be the class of all polynomials $p(x) = \sum_{j=0}^{d} \beta_j x^j$ of degree d on $[0,1]$ such that $|p|_\infty = 1$. Then there exists a constant $c > 0$ such that*

$$|p|_\infty \geq c \max_{0 \leq j \leq d} |\beta_j|.$$

Lemma A.6: *Let μ be a measure on the interval $[0,1]$ with the property that for any $L > 0$, there exists a constant $C(L) > 0$ such that for any interval $A \subset [0,1]$, $\mu(B)/\mu(A) \geq C(L)$ for all intervals $B \subset A$ with $length(B)/length(A) \geq L$. Then for any polynomial p of degree d on $[0,1]$, there exists a constant $c > 0$ such that*

$$\int_A p^2 d\mu \geq c \sup_{u \in A} |p(u)|^2 \mu(A).$$

For the next lemma, assume that the knots are $t_1 = \cdots = t_{d+1} = 0, t_{M+1} = \cdots = t_{M+d+1} = 1$ and $0 < t_{d+2} < \cdots < t_M < 1$. Note that we have placed extra knots at 0 and 1 in order to obtain a B-spline basis. Let $\boldsymbol{\psi} := \{\psi_j : j = 1, ..., M\}$ be the (unnormalized) B-spline basis with the knots $\{t_j : j = d+2, ..., M\}$. Let $\boldsymbol{\beta} \in \mathbb{R}^M$, and consider the spline $s(x) := \sum_{j=1}^{M} \beta_j \psi_j(x)$. Then on the interval $A_i := [t_i, t_{i+1}]$, $s(x) = \sum_{i-d \leq j \leq i} \beta_j \psi_j(x)$ with $\sum_{i-d \leq j \leq i} \psi_j(x) = 1$.

Lemma A.7: *Assume that μ is a measure on $[0,1]$ satisfying the properties of Lemma A.6 above. Consider the vector $\boldsymbol{\psi}$ of B-splines on $[0,1]$ of degree d with well-conditioned knots at $t_{d+2}, ..., t_M$, i.e., the sequence $\{M(t_{i+1} - t_i) : i = d+1, ..., M\}$ remains bounded between two positive constants for any M. Then there exist constants $c_{12}, c_{13} > 0$ (which do not depend on $t_{d+2}, ..., t_M$) such that all the eigenvalues of the matrix $\int \boldsymbol{\psi}\boldsymbol{\psi}^T d\mu$ are between $c_{12} \min_{d+1 \leq i \leq M} \mu(A_i)$ and $c_{13} \max_{d+1 \leq i \leq M} \mu(A_i)$.*

Lemma A.8: *Let h be a bounded nonnegative function on $[0,1]$ which is bounded away from zero except perhaps near 0 and 1. Assume that $\lim_{x \to 0} x^{-\gamma} h(x)$ and $\lim_{x \to 1} (1-x)^{-\gamma} h(x)$ are positive constants for some $0 < \gamma \leq 1$. Let $\boldsymbol{\psi}$ be a (unnormalized) B-spline basis (as in Lemma A.7). Then all the eigenvalues of $\int \boldsymbol{\psi}\boldsymbol{\psi}^T h dx$ are bounded between $c_{10} M^{-1-\gamma}$ and $c_{11} M^{-1}$ for some positive constants $c_{10}, c_{11} > 0$.*

Observe that under the stated condition on the density of F_a in the proposition, the function $h(v)$ appearing in (58) has the same behavior as stated in Lemma A.8 (after a change of location and scale). Proposition 1 now follows from using Halperin-Pitt inequality as in (57), but now taking $\epsilon \sim M^{-2-\gamma}$.

Rate bounds

In this subsection, we summarize approximations of various quantities that are useful in proving Lemmas A.1-A.4. First, by **A3** we have the following:

$$\| g_{\boldsymbol{\beta}}^{(j)} - g_{\boldsymbol{\beta}^*}^{(j)} \|_\infty = O(\alpha_N M^{j+1/2}) \text{ if } \| \boldsymbol{\beta} - \boldsymbol{\beta}^* \| \leq \alpha_N, \quad j = 0, 1, 2. \tag{59}$$

Next, from **A3** and **A4**, for M large enough, solutions $\{X_{il}(t; \boldsymbol{\theta}, \boldsymbol{\beta}) : t \in [0,1]\}$ exist for all $(\boldsymbol{\theta}, \boldsymbol{\beta})$ such that $\max\{\| \boldsymbol{\theta} - \boldsymbol{\theta}^* \|, \| \boldsymbol{\beta} - \boldsymbol{\beta}^* \|\} \leq \alpha_N$. This also implies that the solutions $X_{il}^{\theta_i}(\cdot; \boldsymbol{\theta}, \boldsymbol{\beta})$, $X_{il}^{\beta_r}(\cdot; \boldsymbol{\theta}, \boldsymbol{\beta})$, $X_{il}^{\theta_i, \theta_i}(\cdot; \boldsymbol{\theta}, \boldsymbol{\beta})$, $X_{il}^{\theta_i, \beta_r}(\cdot; \boldsymbol{\theta}, \boldsymbol{\beta})$ and $X_{il}^{\beta_r, \beta_{r'}}(\cdot; \boldsymbol{\theta}, \boldsymbol{\beta})$ exist on $[0,1]$ for all $(\boldsymbol{\theta}, \boldsymbol{\beta})$ such that

$\max\{\| \boldsymbol{\theta} - \boldsymbol{\theta}^* \|, \| \boldsymbol{\beta} - \boldsymbol{\beta}^* \|\} \leq \alpha_N$, since the latter are linear differential equations where the coefficient functions depend on $X_{il}(t;\boldsymbol{\theta},\boldsymbol{\beta})$ (see Appendix A). Moreover, by *Gronwall's lemma* (Lemma F.1), (59) and the fact that $\| g^{(j)}_{\boldsymbol{\beta}^*} \|_\infty = O(1)$ for $j = 0,1,2$ (again by **A3**), all these solutions are bounded for all θ_i and a_{il}, by compactness of $\text{supp}(F_a)$.

Hence, if $\alpha_N M^{3/2} = o(1)$, then using Corollary F.2 (in Appendix F), the fact that $\| g^{(j)}_{\boldsymbol{\beta}^*} \|_\infty = O(1)$ for $j = 0,1,2$, and the expressions for the ODEs for the partial and mixed partial derivatives (see Appendix A), after some algebra we obtain the following (almost surely):

$$\| X_{il}(\cdot;\boldsymbol{\theta}^*,\boldsymbol{\beta}^*) - X^g_{il}(\cdot) \|_\infty = O(M^{-p}). \tag{60}$$

The same technique can be used to prove the following (almost surely):

$$\begin{aligned}
\| X_{il}(\cdot;\boldsymbol{\theta},\boldsymbol{\beta}) - X_{il}(\cdot;\boldsymbol{\theta}^*,\boldsymbol{\beta}^*) \|_\infty &= O(\alpha_N M^{1/2}) && (61)\\
\| X^{\theta_i}_{il}(\cdot;\boldsymbol{\theta},\boldsymbol{\beta}) - X^{\theta_i}_{il}(\cdot;\boldsymbol{\theta}^*,\boldsymbol{\beta}^*) \|_\infty &= O(\alpha_N M^{3/2}) && (62)\\
\max_{1\leq r\leq M} \| X^{\beta_r}_{il}(\cdot;\boldsymbol{\theta}^*,\boldsymbol{\beta}^*) \|_\infty &= O(M^{-1/2}) && (63)\\
\max_{1\leq r\leq M} \| X^{\beta_r}_{il}(\cdot;\boldsymbol{\theta},\boldsymbol{\beta}) - X^{\beta_r}_{il}(\cdot;\boldsymbol{\theta}^*,\boldsymbol{\beta}^*) \|_\infty &= O(\alpha_N M) && (64)\\
\| X^{\theta_i,\theta_i}_{il}(\cdot;\boldsymbol{\theta},\boldsymbol{\beta}) - X^{\theta_i,\theta_i}_{il}(\cdot;\boldsymbol{\theta}^*,\boldsymbol{\beta}^*) \|_\infty &= O(\alpha_N M^{5/2}) && (65)\\
\max_{1\leq r\leq M} \| X^{\theta_i,\beta_r}_{il}(\cdot;\boldsymbol{\theta}^*,\boldsymbol{\beta}^*) \|_\infty &= O(M^{1/2}) && (66)\\
\max_{1\leq r\leq M} \| X^{\theta_i,\beta_r}_{il}(\cdot;\boldsymbol{\theta},\boldsymbol{\beta}) - X^{\theta_i,\beta_r}_{il}(\cdot;\boldsymbol{\theta}^*,\boldsymbol{\beta}^*) \|_\infty &= O(\alpha_N M^2) && (67)\\
\max_{1\leq r,r'\leq M} \| X^{\beta_r,\beta_{r'}}_{il}(\cdot;\boldsymbol{\theta}^*,\boldsymbol{\beta}^*) \|_\infty &= O(1) && (68)\\
\max_{1\leq r,r'\leq M} \| X^{\beta_r,\beta_{r'}}_{il}(\cdot;\boldsymbol{\theta},\boldsymbol{\beta}) - X^{\beta_r,\beta_{r'}}_{il}(\cdot;\boldsymbol{\theta}^*,\boldsymbol{\beta}^*) \|_\infty &= O(\alpha_N M^{3/2}) && (69)
\end{aligned}$$

whenever $\max\{\| \boldsymbol{\theta} - \boldsymbol{\theta}^* \|, \| \boldsymbol{\beta} - \boldsymbol{\beta}^* \|\} \leq \alpha_N$.

To illustrate the key arguments, we prove (63) and (64). By (38), and the fact that $\| \phi_r \|_\infty = O(M^{1/2})$ and is supported on an interval of length $O(M^{-1})$, (63) follows; in fact it holds for all $(\boldsymbol{\theta},\boldsymbol{\beta}) \in \Omega(\alpha_N)$ with $\Omega(\alpha)$ as defined in (47). Next, note that the function ϕ_r is Lipschitz with Lipschitz constant $O(M^{3/2})$ and is supported on an interval of length $O(M^{-1})$. Since (31) (in Appendix A) is a linear differential equation, using Corollary F.2 with

$$\begin{aligned}
& \delta f(t,x)\\
= \;& x\left[e^{\theta_i} g'_{\boldsymbol{\beta}}(X_{il}(t;\boldsymbol{\theta},\boldsymbol{\beta})) - e^{\theta^*_i} g'_{\boldsymbol{\beta}^*}(X_{il}(t;\boldsymbol{\theta}^*,\boldsymbol{\beta}^*))\right] + \; e^{\theta_i}\phi_r(X_{il}(t;\boldsymbol{\theta},\boldsymbol{\beta})) - e^{\theta^*_i}\phi_r(X_{il}(t;\boldsymbol{\theta}^*,\boldsymbol{\beta}^*))\\
= \;& (e^{\theta_i} - e^{\theta^*_i})\left[x g'_{\boldsymbol{\beta}}(X_{il}(t;\boldsymbol{\theta},\boldsymbol{\beta})) + \phi_r(X_{il}(t;\boldsymbol{\theta},\boldsymbol{\beta}))\right] + x e^{\theta^*_i}(g'_{\boldsymbol{\beta}}(X_{il}(t;\boldsymbol{\theta},\boldsymbol{\beta})) - g'_{\boldsymbol{\beta}}(X_{il}(t;\boldsymbol{\theta}^*,\boldsymbol{\beta}^*)))\\
& + x e^{\theta^*_i}(g'_{\boldsymbol{\beta}}(X_{il}(t;\boldsymbol{\theta}^*,\boldsymbol{\beta}^*))) - g'_{\boldsymbol{\beta}^*}(X_{il}(t;\boldsymbol{\theta}^*,\boldsymbol{\beta}^*))) + e^{\theta^*_i}(\phi_r(X_{il}(t;\boldsymbol{\theta},\boldsymbol{\beta})) - \phi_r(X_{il}(t;\boldsymbol{\theta}^*,\boldsymbol{\beta}^*))),
\end{aligned}$$

we obtain (64) by using (61) and the following facts: on $[0,1]$, $|X^{\beta_r}_{il}(t)| = O(M^{-1/2})$ for all $(\boldsymbol{\theta},\boldsymbol{\beta}) \in \Omega(\alpha_N)$; $\| g''_{\boldsymbol{\beta}} \|_\infty = O(\alpha_N M^{5/2})$; $\| g'_{\boldsymbol{\beta}} - g'_{\boldsymbol{\beta}^*} \|_\infty = O(\alpha_N M^{3/2})$; and $\alpha_N M^{3/2} = o(1)$.

Proof of lemmas

Proof of Lemma A.1 : Using (49), write

$$D_1(\boldsymbol{\gamma}) \;:=\; \boldsymbol{\gamma}^T W \;=\; \sum_{i=1}^{n}\sum_{l=1}^{N_i}\sum_{j=1}^{m_i} (\varepsilon_{ilj} + \Delta_{ilj})\boldsymbol{\gamma}^T \mathbf{v}_{ilj},$$

where $\Delta_{ilj} := X^g_{ilj} - X_{ilj}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*)$, and $\mathbf{v}_{ilj}$ is the $(M+n-1)\times 1$ vector with the first M coordinates given by $\mathbf{v}^{\beta}_{ilj} := X^{\boldsymbol{\beta}}_{il}(T_{i,j}; a_{il}, \theta^*_i, \boldsymbol{\beta}^*)$, and the last $(n-1)$ coordinates given by $\mathbf{v}^{\theta}_{ilj} := X^{\theta_i}_{il}(T_{i,j}; a_{il}, \theta^*_i, \boldsymbol{\beta}^*)\mathbf{e}_{i-1}$, where $\mathbf{e}_i$ is the i-th canonical basis vector in $\mathbb{R}^{n-1}$, and $\mathbf{e}_0 := \mathbf{0}_{n-1}$. Notice that $\mathcal{G}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*) = \sum_{i=1}^n \sum_{l=1}^{N_i} \sum_{j=1}^{m_i} \mathbf{v}_{ilj}\mathbf{v}_{ilj}^T$. Thus, by Cauchy-Schwarz inequality, and the fact that $\max_{i,l,j} |\Delta_{ilj}| = O(M^{-p})$ (by (60)) we have, uniformly in $\boldsymbol{\gamma}$,

$$|\sum_{i=1}^n \sum_{l=1}^{N_i} \sum_{j=1}^{m_i} \Delta_{ilj}\boldsymbol{\gamma}^T\mathbf{v}_{ilj}| \;=\; O(M^{-p}(\overline{N}\overline{m})^{1/2})\sqrt{\boldsymbol{\gamma}^T\mathcal{G}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*)\boldsymbol{\gamma}}.$$

Since ε_{ilj} are i.i.d. $N(0, \sigma^2_\varepsilon)$, we also have

$$\sum_{i=1}^n \sum_{l=1}^{N_i} \sum_{j=1}^{m_i} \varepsilon_{ilj}\boldsymbol{\gamma}^T\mathbf{v}_{ilj} \sim N(0, \sigma^2_\varepsilon\boldsymbol{\gamma}^T\mathcal{G}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*)\boldsymbol{\gamma})$$

conditional on $(\boldsymbol{a}, \mathbf{T})$. Since the (conditional) Gaussian process

$$f(\boldsymbol{\gamma}) := \frac{\sum_{i=1}^n \sum_{l=1}^{N_i} \sum_{j=1}^{m_i} \varepsilon_{ilj}\boldsymbol{\gamma}^T\mathbf{v}_{ilj}}{\sigma_\varepsilon\sqrt{\boldsymbol{\gamma}^T\mathcal{G}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*)\boldsymbol{\gamma}}}$$

is a smooth function over $\mathbb{S}^{M+n-1}$ (the unit sphere centered at 0 in $\mathbb{R}^{M+n-1}$), and since by assumption $M = O((\overline{N}\overline{m})^d)$ for some $d > 0$, using a covering of the sphere $\mathbb{S}^{M+n-1}$ by balls of radius $\epsilon_M \sim (\overline{N}\overline{m})^{-D}$ for an appropriately chosen $D > 0$, and using the fact that $P(N(0,1) > t) \leq t^{-1}(2\pi)^{-1/2}\exp(-t^2/2)$, for $t > 0$, we conclude that uniformly in $\boldsymbol{\gamma}$,

$$\sum_{i=1}^n \sum_{l=1}^{N_i} \sum_{j=1}^{m_i} \varepsilon_{ilj}\boldsymbol{\gamma}^T\mathbf{v}_{ilj} = O(\sigma_\varepsilon M^{1/2}\sqrt{\log(\overline{N}\overline{m})})\sqrt{\boldsymbol{\gamma}^T\mathcal{G}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*)\boldsymbol{\gamma}}$$

except on a set with probability converging to zero. This completes the proof of the lemma.

Proof of Lemma A.2 : Define $\mathbf{u}_{ilj}$ the same way as $\mathbf{v}_{ilj}$ is defined in the proof of Lemma A.1, with $(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*)$ replaced by $(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\beta}})$. Express $D_2(\boldsymbol{\gamma}) := \boldsymbol{\gamma}^T(\mathcal{G}(\overline{\boldsymbol{\theta}}, \overline{\boldsymbol{\beta}}) - \mathcal{G}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*))\boldsymbol{\gamma}$ as

$$\begin{aligned} & \boldsymbol{\gamma}^T\left[\sum_{i=1}^n \sum_{l=1}^{N_i} \sum_{j=1}^{m_i} \mathbf{u}_{ilj}\mathbf{u}_{ilj}^T - \sum_{i=1}^n \sum_{l=1}^{N_i} \sum_{j=1}^{m_i} \mathbf{v}_{ilj}\mathbf{v}_{ilj}^T\right]\boldsymbol{\gamma} \\ = \;& \boldsymbol{\gamma}^T\left[\sum_{i=1}^n \sum_{l=1}^{N_i} \sum_{j=1}^{m_i} ((\mathbf{u}_{ilj} - \mathbf{v}_{ilj})\mathbf{v}_{ilj}^T + (\mathbf{u}_{ilj} - \mathbf{v}_{ilj})(\mathbf{u}_{ilj} - \mathbf{v}_{ilj})^T + \mathbf{v}_{ilj}(\mathbf{u}_{ilj} - \mathbf{v}_{ilj})^T)\right]\boldsymbol{\gamma}. \end{aligned}$$

Then, by Cauchy-Schwarz inequality and (62) and (64), and the arguments used in the proof of Lemma A.1,

$$|D_2(\boldsymbol{\gamma})| \;=\; O(\alpha_N M^{3/2}(\overline{N}\overline{m})^{1/2})\sqrt{\boldsymbol{\gamma}^T\mathcal{G}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*)\boldsymbol{\gamma}} + O(\alpha_N^2 M^3\overline{N}\overline{m}).$$

Proof of Lemma A.3 : Define $D_3(\boldsymbol{\gamma}) := \boldsymbol{\gamma}^T(\mathcal{G}(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*) - \mathcal{G}_*)\boldsymbol{\gamma}$. Then

$$D_3(\boldsymbol{\gamma}) = \sum_{i=1}^n \sum_{l=1}^{N_i} \sum_{j=1}^{m_i} u_{ilj}(\boldsymbol{\gamma}) + \sum_{i=1}^n \sum_{l=1}^{N_i} \sum_{j=1}^{m_i} w_{ilj}(\boldsymbol{\gamma}),$$

where $u_{ilj}(\boldsymbol{\gamma}) = \boldsymbol{\gamma}^T(\nabla_i X_{ilj}\nabla_i X_{ilj}^T - \mathbb{E}[(\nabla_i X_{ilj}\nabla_i X_{ilj}^T)|a_{il}])\boldsymbol{\gamma}$ and $w_{ilj}(\boldsymbol{\gamma}) = \boldsymbol{\gamma}^T(\mathbb{E}[(\nabla_i X_{ilj}\nabla_i X_{ilj}^T)|a_{il}] - \mathbb{E}[\nabla_i X_{ilj}\nabla_i X_{ilj}^T)])\boldsymbol{\gamma}$, where, for notational simplicity,

$$\nabla_i X_{ilj} = \begin{bmatrix} X_{il}^{\boldsymbol{\beta}}(T_{i,j}) \\ X_{il}^{\theta_i}(T_{i,j})\mathbf{e}_{i-1} \end{bmatrix}, \qquad i = 1, \ldots, n.$$

Note that, the random variables $u_{ilj}(\boldsymbol{\gamma})$ have zero conditional mean (given a_{il}), are uniformly bounded and the variables $Z_{ij}(\boldsymbol{\gamma}) := \sum_{l=1}^{N_i} u_{ilj}(\boldsymbol{\gamma})$ are independent. Similarly, the random variables $\{w_{ilj}\}_{i,j}$ have zero mean are uniformly bounded and the variables $\sum_{j=1}^{m_i} w_{ilj}(\boldsymbol{\gamma})$ are independent. Indeed, for each fixed (i,l), the variables $\{w_{ilj}\}_{j=1}^{m_i}$ are identical since $T_{i,j}$ are i.i.d. Moreover, the collections $\{u_{ilj}(\boldsymbol{\gamma})\}$ and $\{w_{ilj}(\boldsymbol{\gamma})\}$ are differentiable functions of γ. Define $\mathcal{G}_*(\boldsymbol{a}) := \mathbb{E}(\mathcal{G}_*(\boldsymbol{\theta}^*, \boldsymbol{\beta}^*)|\boldsymbol{a})$. Then, since $Z_{ij}(\boldsymbol{\gamma})$ are uniformly bounded by $K_1\overline{N}$ for some constant $K_1 > 0$, and are independent given $\boldsymbol{a}$, we have

$$\begin{aligned} \mathrm{Var}(\sum_{i=1}^{n}\sum_{j=1}^{m_i} Z_{ij}(\boldsymbol{\gamma})|\boldsymbol{a}) &= \sum_{i=1}^{n}\sum_{j=1}^{m_i}\mathbb{E}[(Z_{ij}(\boldsymbol{\gamma}))^2|\boldsymbol{a}] \\ &\le \sum_{i=1}^{n}\sum_{j=1}^{m_i} N_i \sum_{l=1}^{N_i}\mathbb{E}[u_{ilj}^2(\boldsymbol{\gamma})|a_{il}] \\ &\le K_2\overline{N}\sum_{i=1}^{n}\sum_{l=1}^{N_i}\sum_{j=1}^{m_i}\mathbb{E}[|u_{ilj}(\boldsymbol{\gamma})||a_{il}] \;\le\; 2K_2\overline{N}\boldsymbol{\gamma}^T\mathcal{G}_*(\boldsymbol{a})\boldsymbol{\gamma}. \end{aligned}$$

In the above, second inequality uses $(\sum_{i=1}^N x_i)^2 \le N\sum_{i=1}^N x_i^2$, and the last follows from fact that $u_{ilj}(\boldsymbol{\gamma})$ is a difference of two nonnegative quantities, the second one being the conditional expectation of the first one given $\boldsymbol{a}$. Thus, applying Bernstein's inequality, for every $v > 0$, for every $\boldsymbol{\gamma} \in \mathbb{S}^{M+n-1}$,

$$\mathbb{P}(|\sum_{i=1}^{n}\sum_{l=1}^{N_i}\sum_{j=1}^{m_i} u_{ilj}(\boldsymbol{\gamma})| > v|\boldsymbol{a}) \le 2\exp\left(-\frac{v^2/2}{2K_2\overline{N}\boldsymbol{\gamma}^T\mathcal{G}_*(\boldsymbol{a})\boldsymbol{\gamma} + K_1\overline{N}v/3}\right).$$

Thus, using an entropy argument as in the proof of Lemma A.1, we conclude that given $\delta > 0$ there exist positive constants $C_1(\delta)$ and $C_2(\delta)$ such that on the set $\{\boldsymbol{a}|\boldsymbol{\gamma}^T\mathcal{G}_*(\boldsymbol{a})\boldsymbol{\gamma} \ge C_2(\delta)\overline{N}M\log(\overline{N}\overline{m})\}$,

$$\mathbb{P}\left(\sup_{\boldsymbol{\gamma}\in\mathbb{S}^{M+n-1}} \frac{|\sum_{i=1}^{n}\sum_{l=1}^{N_i}\sum_{j=1}^{m_i} u_{ilj}(\boldsymbol{\gamma})|}{\sqrt{\boldsymbol{\gamma}^T\mathcal{G}_*(\boldsymbol{a})\boldsymbol{\gamma}}} > C_1(\delta)(\overline{N}M\log(\overline{N}\overline{m}))^{1/2} \mid \boldsymbol{a}\right) \le (\overline{N}\overline{m})^{-\delta}. \tag{70}$$

On the other hand, using an inversion formula for block matrices,

$$\begin{aligned} \|\,\mathcal{G}_*^{-1}\,\| &= \|\begin{bmatrix}\mathcal{G}_{*,\beta\beta} & \mathcal{G}_{*,\beta\theta} \\ \mathcal{G}_{*,\theta\beta} & \mathcal{G}_{*,\theta\theta}\end{bmatrix}^{-1}\| \\ &= \|\begin{bmatrix} C_*^{-1} & -C_*^{-1}\mathcal{G}_{*,\beta\theta}(\mathcal{G}_{*,\theta\theta})^{-1} \\ -(\mathcal{G}_{*,\theta\theta})^{-1}\mathcal{G}_{*,\theta\beta}C_*^{-1} & (\mathcal{G}_{*,\theta\theta})^{-1} + (\mathcal{G}_{*,\theta\theta})^{-1}\mathcal{G}_{*,\theta\beta}C_*^{-1}\mathcal{G}_{*,\beta\theta}(\mathcal{G}_{*,\theta\theta})^{-1}\end{bmatrix}\| \\ &= O(\kappa_M(\overline{N}\overline{m})^{-1}), \end{aligned} \tag{71}$$

where $C_* := \mathcal{G}_{*,\beta\beta} - \mathcal{G}_{*,\beta\theta}(\mathcal{G}_{*,\theta\theta})^{-1}\mathcal{G}_{*,\theta\beta}$. The last equality in (71) is because **A6** together with (63) implies in particular that $\| \mathcal{G}_{*,\beta\theta}(\mathcal{G}_{*,\theta\theta})^{-1} \| = O(1)$. Now, from the facts that

$$\boldsymbol{\gamma}^T\mathcal{G}_*\boldsymbol{\gamma} \geq K_3\frac{\overline{N}\overline{m}}{\kappa_M} \ \text{(by (71))} \qquad \text{and} \qquad \min\{\overline{N},\overline{m}\} \gg \kappa_M M\log(\overline{N}\overline{m}),$$

for some constant $K_3 > 0$, so that $\boldsymbol{\gamma}^T\mathcal{G}_*\boldsymbol{\gamma} \gg \overline{m}M\log(\overline{N}\overline{m})$, and using arguments similar to those leading to (70) we have, for some $C_3(\delta) > 0$,

$$\mathbb{P}\left(\sup_{\boldsymbol{\gamma}\in\mathbb{S}^{M+n-1}} \frac{|\sum_{i=1}^n\sum_{l=1}^{N_i}\sum_{j=1}^{m_i} w_{ilj}(\boldsymbol{\gamma})|}{\sqrt{\boldsymbol{\gamma}^T\mathcal{G}_*\boldsymbol{\gamma}}} > C_3(\delta)(\overline{m}M\log(\overline{N}\overline{m}))^{1/2}\right) \leq (\overline{N}\overline{m})^{-\delta}. \tag{72}$$

Now, observing that $\sum_{i=1}^n\sum_{l=1}^{N_i}\sum_{j=1}^{m_i} w_{ilj}(\boldsymbol{\gamma}) = \boldsymbol{\gamma}^T\mathcal{G}_*(\boldsymbol{a})\boldsymbol{\gamma} - \boldsymbol{\gamma}^T\mathcal{G}_*\boldsymbol{\gamma}$, using fact [**Q**], and combining with (70) and (72), we obtain that, there is a $C_4(\delta) > 0$ such that

$$\mathbb{P}\left(\sup_{\boldsymbol{\gamma}\in\mathbb{S}^{M+n-1}} \frac{|D_3(\boldsymbol{\gamma})|}{\sqrt{\boldsymbol{\gamma}^T\mathcal{G}_*\boldsymbol{\gamma}}} \geq C_4(\delta)(\overline{m}^{1/2} + \overline{N}^{1/2})(M\log(\overline{N}\overline{m}))^{1/2}\right) = O((\overline{N}\overline{m})^{-\delta}). \tag{73}$$

(Note that, if the time points $\{t_{ilj}\}$ were independently and identically distributed for different curves (i,l), then quantity $(\overline{m}^{1/2}+\overline{N}^{1/2})(M\log(\overline{N}\overline{m}))^{1/2}$ in (73) can be replaced by $(M\log(\overline{N}\overline{m}))^{1/2}$). From (73) it follows that

$$\boldsymbol{\gamma}^T\mathcal{G}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*)\boldsymbol{\gamma} \ \geq \ \boldsymbol{\gamma}^T\mathcal{G}_*\boldsymbol{\gamma}(1-o_P(1)) \ \geq \ c_6\kappa_M^{-1}\overline{N}\overline{m}(1-o_P(1))$$

for some constant $c_6 > 0$ and for sufficiently large $\overline{N}$.

Proof of Lemma A.4 : Using (42) and (43) in Appendix A, for any t, we can express the $(M+n-1)\times(M+n-1)$ matrix with blocks $X_{il}^{\boldsymbol{\beta},\boldsymbol{\beta}^T}(t) := ((X_{il}^{\beta_r,\beta_{r'}}(t)))_{r,r'=1}^M$, $\mathbf{e}_{i-1}X_{il}^{\theta_i,\boldsymbol{\beta}^T}(t)$, $X_{il}^{\boldsymbol{\beta},\theta_i}(t)\mathbf{e}_{i-1}^T$ and $X_{il}^{\theta_i,\theta_i}(t)$, as $U_{il}(t) + U_{il}(t)^T + V_{il}(t)$ where

$$U_{il}(t) \ = \ e^{\theta_i}g_{\boldsymbol{\beta}}(X_{il}(t))\int_0^t \frac{1}{g_{\boldsymbol{\beta}}(X_{il}(s))}\begin{bmatrix} X_{il}^{\boldsymbol{\beta}}(s) \\ X_{il}^{\theta_i}(s)\mathbf{e}_{i-1}\end{bmatrix}\begin{bmatrix}\boldsymbol{\phi}'(X_{il}(s)) \\ \mathbf{0}_{n-1}\end{bmatrix}^T ds \tag{74}$$

and $\| V_{il}(t) \| = O(1)$ uniformly in t, i and l. Note that, in the above description, all the sample paths and their derivatives are evaluated at $(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*)$.

Observe that, since $Y_{ilj} - X_{il}(T_{i,j};\boldsymbol{\theta}^*,\boldsymbol{\beta}^*) = \varepsilon_{ilj} + \Delta_{ilj}$, where Δ_{ilj} is as in the proof of Lemma A.1, we have

$$\begin{aligned}
&\mathcal{H}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}}) - \mathcal{H}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*) \\
= \ &\sum_{i=1}^n\sum_{l=1}^{N_i}\sum_{j=1}^{m_i}(X_{ilj}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*) - X_{ilj}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}}))\begin{bmatrix} X_{ilj}^{\boldsymbol{\beta},\boldsymbol{\beta}^T}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}}) & X_{ilj}^{\boldsymbol{\beta},\theta_i}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}})\mathbf{e}_{i-1}^T \\ \mathbf{e}_{i-1}X_{ilj}^{\theta_i,\boldsymbol{\beta}^T}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}}) & X_{ilj}^{\theta_i,\theta_i}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}})\mathbf{e}_{i-1}\mathbf{e}_{i-1}^T\end{bmatrix} \\
&+\sum_{i=1}^n\sum_{l=1}^{N_i}\sum_{j=1}^{m_i}(\varepsilon_{ilj}+\Delta_{ilj}) \ \cdot \\
&\begin{bmatrix} X_{ilj}^{\boldsymbol{\beta},\boldsymbol{\beta}^T}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}}) - X_{ilj}^{\boldsymbol{\beta},\boldsymbol{\beta}^T}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*) & (X_{ilj}^{\boldsymbol{\beta},\theta_i}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}}) - X_{ilj}^{\boldsymbol{\beta},\theta_i}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*))\mathbf{e}_{i-1}^T \\ \mathbf{e}_{i-1}(X_{ilj}^{\theta_i,\boldsymbol{\beta}^T}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}}) - X_{ilj}^{\theta_i,\boldsymbol{\beta}^T}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*)) & (X_{ilj}^{\theta_i,\theta_i}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}}) - X_{ilj}^{\theta_i,\theta_i}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*))\mathbf{e}_{i-1}\mathbf{e}_{i-1}^T\end{bmatrix}.
\end{aligned} \tag{75}$$

First break the last summation in the last term of (75) into two parts – one corresponding to Δ_{ilj}'s and the other corresponding to ε_{ilj}'s. Then, using (60), (65), (67) and (69), we conclude that the sum involving Δ_{ilj} is $O(\alpha_N M^{5/2-p}\overline{N}\overline{m})$. The summation involving ε_{ilj}'s can be expressed as a linear function of $\boldsymbol{\varepsilon}$ with coefficients that are functions of $\boldsymbol{a}, \boldsymbol{T}$ and $\boldsymbol{\gamma}$, and depend smoothly on $\boldsymbol{\gamma}$. From this, conditionally on $(\boldsymbol{a}, \boldsymbol{T})$, this term is coordinatewise normally distributed with standard deviation $O(\sigma_\varepsilon \alpha_N M^{5/2}(\overline{N}\overline{m})^{1/2})$ for each fixed $\boldsymbol{\gamma}$. We can conclude from this by an entropy argument (similar to the one used in the proof of Lemma A.1) that the supremum of this term over all $\boldsymbol{\gamma} \in \mathbb{S}^{M+n-1}$ is $O(\sigma_\varepsilon \alpha_N M^3 (\overline{N}\overline{m})^{1/2}\sqrt{\log(\overline{N}\overline{m})})$ with probability tending to 1.

Next, using (74) we express the first term of (75) as

$$\begin{aligned}
&\sum_{i=1}^{n}\sum_{l=1}^{N_i}\sum_{j=1}^{m_i}(X_{ilj}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*) - X_{ilj}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}}))\left[U_{il}(t_{ilj}) + U_{il}(t_{ilj})^T + V_{il}(t_{ilj})\right] \\
&+\sum_{i=1}^{n}\sum_{l=1}^{N_i}\sum_{j=1}^{m_i}(X_{ilj}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*) - X_{ilj}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}}))\cdot \\
&\begin{bmatrix} X_{ilj}^{\boldsymbol{\beta},\boldsymbol{\beta}^T}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}}) - X_{ilj}^{\boldsymbol{\beta},\boldsymbol{\beta}^T}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*) & (X_{ilj}^{\boldsymbol{\beta},\theta_i}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}}) - X_{ilj}^{\boldsymbol{\beta},\theta_i}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*))\mathbf{e}_{i-1}^T \\ \mathbf{e}_{i-1}(X_{ilj}^{\theta_i,\boldsymbol{\beta}^T}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}}) - X_{ilj}^{\theta_i,\boldsymbol{\beta}^T}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*)) & (X_{ilj}^{\theta_i,\theta_i}(\overline{\boldsymbol{\theta}},\overline{\boldsymbol{\beta}}) - X_{ilj}^{\theta_i,\theta_i}(\boldsymbol{\theta}^*,\boldsymbol{\beta}^*))\mathbf{e}_{i-1}\mathbf{e}_{i-1}^T \end{bmatrix}. \qquad (76)
\end{aligned}$$

The second sum is $O(\alpha_N^2 M^3 \overline{N}\overline{m})$ by (61), (65), (67) and (69). Again, by (61), the contribution in the first sum for the term involving $V_{il}(t_{ilj})$'s is $O(\alpha_N M^{1/2}\overline{N}\overline{m})$. Finally, by Cauchy-Schwarz inequality, and (74), and the facts that $g_{\boldsymbol{\beta}}$ is bounded both above and below (for $x \geq x_0$), we have for all $0 \leq t \leq 1$,

$$|\boldsymbol{\gamma}^T U_{il}(t)\boldsymbol{\gamma}|^2 \;\leq\; c_g \int_0^t \left(\boldsymbol{\gamma}^T \begin{bmatrix} X_{il}^{\boldsymbol{\beta}}(s) \\ X_{il}^{\theta_i}(s)\mathbf{e}_{i-1}\end{bmatrix}\right)^2 ds \int_0^1 \left(\boldsymbol{\gamma}^T \begin{bmatrix} \boldsymbol{\phi}'(X_{il}(s)) \\ \mathbf{0}_{n-1}\end{bmatrix}\right)^2 ds,$$

for some constant $c_g > 0$ depending on $g_{\boldsymbol{\beta}^*}$, $\boldsymbol{\theta}$ and x_0. Notice that $\int \boldsymbol{\phi}'(x)(\boldsymbol{\phi}'(x))^T dx \leq c_7 M^2 I_M$ for some $c_7 > 0$. Furthermore,

$$\sup_{\boldsymbol{\gamma}\in\mathbb{S}^{M+n-1}} \sup_{t\in[0,1]} \int_0^t \left(\boldsymbol{\gamma}^T \begin{bmatrix} X_{il}^{\boldsymbol{\beta}}(s) \\ X_{il}^{\theta_i}(s)\mathbf{e}_{i-1}\end{bmatrix}\right)^2 ds = O(1).$$

These, together with an application of Cauchy-Schwarz inequality, and using manipulations as in the proof of Lemmas A.2 and A.3, shows that the sum involving the terms $U_{il}(t_{ilj})$'s in the expression (76) is $O(\alpha_N M^{3/2}(\overline{N}\overline{m})^{1/2})\sqrt{\boldsymbol{\gamma}^T \mathcal{G}_* \boldsymbol{\gamma}}$. Lemma A.4 now follows.

Proof of Lemma A.5 : Suppose that the result is false. Then for any sequence of positive constant δ_n decreasing to zero, we can find a polynomial p_n with coefficients $\beta_{n,j}$, $j = 0, ..., d$, such that $|p_n|_\infty \leq \delta_n \max_{0\leq j\leq d}|\beta_{n,j}|$. Let q_n be the polynomial whose coefficients $\{\gamma_{n,j}\}$ are obtained by dividing β_j's by $\max_{0\leq j\leq d}|\beta_{n,j}|$. Note that $\max|\gamma_{n,j}| = 1$ for any n. By the usual compactness argument we can find a subsequence of $\{\gamma_{n,j}\}$, which we continue to denote by $\{\gamma_{n,j}\}$, such that $\gamma_{n,j} \to \gamma_j$, $j = 0, ..., d$, where $\max_j |\gamma_j| = 1$. However the supremum norm of the limiting polynomial q of degree d with coefficients γ_j is zero which implies that $\gamma_j = 0$ for all j. This leads to a contradiction.

Proof of Lemma A.6 : Note that for any interval $A = [a, b]$, $a < b$, we can write $\int_A p^2 d\mu = \int_0^1 q^2 d\mu_A$, where $q(z) = p(a + (b - a)z)$ and $d\mu_A(z) = d\mu(a + (b - a)z)$. Since $|q|_\infty = \sup_{u\in A}|p(u)|$,

we may take $|q|_\infty = 1$. Let x^* be a point in $[0,1]$ at which $q(x^*)$ equals ± 1. Then for any $0 \le x \le 1$, $q^2(x) = 1 + (x - x^*)^2 q(x^{**}) q''(x^{**})$ for some x^{**} in $[0,1]$. Using Lemma A.5, we see that there is a constant c_{14} such that $|p''|_\infty \le c_{14}$ for all polynomials with $|p|_\infty = 1$. So $|q^2(x) - 1| \le c_{14}(x - x^*)^2$. So we can find an interval $I \subset [0,1]$ of length at least $L = (2c_{14})^{-1/2}$ containing x^* such that $q^2(x) \ge 1/2$. Let $B = a + (b-a)I$. Then $length(B)/length(A) = length(I) \ge L$. Consequently,

$$\begin{aligned} \int_A p^2 d\mu &= \int_0^1 q^2 d\mu_A \ \ge \ \frac{1}{2} \int_I d\mu_A \\ &= \frac{1}{2}\mu(B) \ \ge \ \frac{1}{2} C(L) \mu(A). \end{aligned}$$

The result now follows.

Proof of Lemma A.7 : This result is clearly true for $d = 0$. We will prove it for the case $d \ge 1$. Let $\boldsymbol{\beta} \in \mathbb{R}^k$ and let $s(x) = \boldsymbol{\beta}^T \boldsymbol{\psi}(x)$. Since s is a convex combination of $\beta_{i-d}, ..., \beta_i$ on the interval A_i, we have

$$\int s^2 d\mu = \sum_{d+1 \le i \le M} \int_{t_i}^{t_{i+1}} s^2 d\mu \le \sum_{d+1 \le i \le M} \sum_{i-d \le j \le i} \beta_j^2 \mu(A_i).$$

This establishes the upper bound for the largest eigenvalue of $\int \boldsymbol{\psi}\boldsymbol{\psi}^T$. We will now establish the result on the lower bound of the smallest eigenvalue.

Using property (viii) in chapter XI in de Boor (1978), we know that $\sup_{t_{i+1} \le x \le t_{i+d+1}} |s(x)| \ge c_{15}|\beta_i|$ for all i for some constant $c_{15} > 0$. Denote $m_0 = \min_{d+1 \le i \le M} \mu(A_i)$. Hence for any $d \le i \le M$, by Lemma A.6 we have

$$\begin{aligned} \int_{t_{i+1}}^{t_{i+d+1}} s^2 d\mu &= \sum_{i+1 \le j \le i+d+1} \int_{t_j}^{t_{j+1}} s^2 d\mu \\ &\ge c \sum_{i+1 \le j \le i+d+1} \sup_{t_j \le x \le t_{j+1}} |s(x)|^2 \mu(A_j) \\ &\ge c \sup_{t_{i+1} \le x \le t_{i+d+1}} |s(x)|^2 m_0 \ \ge \ c_{16} \beta_i^2 m_0. \end{aligned}$$

Incidentally, the same type of inequality holds for any $i = 1, .., d-1$. Consequently we have $\int s^2 d\mu \ge c_{17} \sum \beta_i^2 m_0$ for some constant $c_{17} > 0$. This completes the proof.

Proof of Lemma A.8: This follows from Lemma A.7 once we take $d\mu(x) = h(x)dx$.

Appendix F : Perturbation of Differential Equations

For nonparametric estimation of the gradient function g, we need to control the effect of lack of fit to g (meaning that g may not be exactly represented in the given basis $\{\phi_k(\cdot)\}$) on the sample paths $\{X_{il}(t) : t \in [0,1]\}$. It is convenient to do this study under a general setting of first order differential equations where the state variable $x(\cdot)$ is d-dimensional for $d \ge 1$. Our aim is to control the perturbation of the sample paths and its derivatives with respect to the parameters governing the differential equation when the *true* gradient function g is perturbed by an arbitrary function $\delta g(\cdot)$.

We present two different results about the perturbation of the solution paths of the initial value problem:

$$x' = f(t,x), \quad x(t_0) = x_0, \tag{77}$$

where $x \in \mathbb{R}^d$, when the function f is perturbed by a smooth function.

Theorem F.1 (Deuflhard and Bornemann, 2002, p.80) : *On the augmented phase space Ω let the mappings f and δf be continuous and continuously differentiable with respect to the state variable. Assume that for $(t_0, x_0) \in \Omega$, the initial value problem (77), and the perturbed problem*

$$x' = f(t,x) + \delta f(t,x), \quad x(t_0) = x_0, \tag{78}$$

have the solutions x and $\overline{x} = x + \delta x$, respectively. Then for t_1 sufficiently close to t_0, there exists a continuous matrix-valued mapping $M : \Delta \to \mathbb{R}^{d\times d}$ on $\Delta = \{(t,s) \in \mathbb{R}^2 : t \in [t_0, t_1], s \in [t_0, t]\}$ such that the perturbation δx is represented by

$$\delta x(t) = \int_{t_0}^{t} M(t,s)\delta f(s, \overline{x}(s))ds, \quad \textit{for all } t \in [t_0, t_1]. \tag{79}$$

Note that, the point t_1 can be chosen so that, the one-parameter family of initial value problems

$$x' = f(t,x) + \lambda \cdot \delta f(t,x), \quad x(t_0) = x_0, \tag{80}$$

has a corresponding solution $\phi(\cdot;\lambda) \in C^1([t_0, t_1], \mathbb{R}^d)$ for each parameter value $\lambda \in [0,1]$. In particular, $\phi(\cdot;0) = x(\cdot)$ and $\phi(\cdot;1) = \overline{x}(\cdot)$.

Propagation matrix and its relationship to perturbation

Let Φ^{t,t_0} denote the map such that $x(t) = \Phi^{t,t_0}x_0$ is the unique solution of the initial value problem (77). The following result (Theorem 3.1 in Deuflhard and Bornemann, 2002, p.77) describes the dependence of the map on the gradient function f.

Theorem F.2 : *On the extended state space Ω let f be continuous and p-times continuously differentiable, $p \geq 1$, with respect to the state variable. Moreover, suppose that for $(t_0, x_0) \in \Omega$ the unique solution of the initial value problem (77) exists up to some time $t > t_0$. Then there is a neighborhood of the the state x_0 where for all $s \in [t_0, t_1]$, the evolution*

$$x \to \Phi^{s,t_0}x$$

is p-times continuously differentiable with respect to the state variable. In other words, the evolution inherits from the right side the smoothness properties with respect to the state variable.

Then, the linearized perturbation of the state, due to a perturbation δx_s of the state at time s, namely,

$$\delta x(t) \approx \Phi^{t,s}(x(s) + \delta x_s) - \Phi^{t,s}x(s)$$

is given by $\delta x(t) = W(t,s)\delta x_s$ where

$$W(t,s) = D_\xi \Phi^{t,s}\xi \,\big|_{\xi = \Phi^{s,t_0}x_0} \in \mathbb{R}^{d\times d} \tag{81}$$

is the *Jacobi matrix*. Note that, $W(t,s)$ satisfies the differential equation:

$$\frac{d}{dt}W(t,s) = f_x(t, \Phi^{t,t_0}x_0)W(t,s), \tag{82}$$

with initial condition $W(s,s) = I$. $W(t,s)$ is called the *propagation matrix* belonging to x.

In general, we can express the matrix $M(t,s)$ appearing in Theorem F.1 as

$$M(t,s) = \int_0^1 W(t,s;\lambda)d\lambda, \tag{83}$$

where $W(t,s;\lambda)$ is the propagation matrix belonging to $\phi(\cdot;\lambda)$, and hence solves the *homogeneous* differential equation

$$\frac{d}{dt}W(t,s;\lambda) = f_x(t,\phi(t;\lambda))W(t,s;\lambda), \qquad W(s,s) = I.$$

From this, the following corollary follows easily.

Corollary F.1 : *If the limit $\delta f \to 0$ is uniform in a neighborhood of the graph of the solution x, then the linearization*

$$\delta x(t) \approx \int_{t_0}^t W(t,s)\delta f(s,x(s))ds, \quad \textit{for all } t \in [t_0,t_1]$$

holds.

Gronwall's Lemma and its implications

Lemma F.1 (Gronwall's Lemma): *Let $\psi, \chi \in C([t_0,t_1],\mathbb{R})$ be nonnegative functions and $\rho \geq 0$. Then the integral inequality*

$$\psi(t) \leq \rho + \int_{t_0}^t \chi(s)\psi(s)ds, \qquad \textit{for all } t \in [t_0,t_1]$$

implies

$$\psi(t) \leq \rho \exp\left(\int_{t_0}^t \chi(s)\psi(s)ds\right) \qquad \textit{for all } t \in [t_0,t_1].$$

In particular, $\psi \equiv 0$ holds for $\rho = 0$.

An immediate application of Lemma 1 is that, it gives a bound for $\| W(t,s;\lambda) \|$. Indeed, if

$$\| f_x(t;\phi(t;\lambda)) \| \leq \chi(t), \quad \text{for all } \lambda \in [0,1], \tag{84}$$

then taking $\psi(t) = \| W(t,s;\lambda) \|$ (note that $\psi(\cdot)$ depends on s) and $\rho = \| W(s,s;\lambda) \| = \| I \| = 1$, we obtain

$$\| W(t,s;\lambda) \| \leq \exp\left(\int_s^t \chi(u)du\right), \quad \text{for all } t_0 \leq s < t \leq t_1, \text{ for all } \lambda \in [0,1]. \tag{85}$$

Condition (84) holds in particular if $\| f_x(t,\cdot) \|_\infty \leq \chi(t)$, and then, from Theorem F.1, we obtain the important result,

Corollary F.2 : *If f is such that $\| f_x(t,\cdot) \|_\infty \leq \chi(t)$ for a function $\chi(\cdot)$ bounded on $[t_0,t_1]$, and $\| \delta f(t,\cdot) \|_\infty \leq \tau(t)$ for some nonnegative function $\tau(\cdot)$ on $[t_0,t_1]$, then*

$$\| \delta x(t) \| \leq \int_{t_0}^t \exp\left(\int_s^t \chi(u)du\right)\tau(s)ds, \qquad \textit{for all } t \in [t_0,t_1]. \tag{86}$$

Note that, even though $M(t,s)$ in (79) in general depends on x_0, the bound in (86) does not. This has the implication that if one can prove the existence of solutions $\{\phi(\cdot;\lambda) : \lambda \in [0,1]\}$ on an interval $[t_0,t_1]$ for an arbitrary collection of initial conditions x_0, and the conditions of Corollary F.2 hold, then the same perturbation bound (86) applies uniformly to each one of them.